Solving MRF Minimization by Mirror Descent

Duy V.N. Luong, Panos Parpas, Daniel Rueckert, and Berç Rustem

Department of Computing, Imperial College London, United Kingdom

Abstract. Markov Random Fields (MRF) minimization is a well-known problem in computer vision. We consider the augmented dual of the MRF minimization problem and develop a Mirror Descent algorithm based on weighted Entropy and Euclidean Projection. The augmented dual problem consists of maximizing a non-differentiable objective function subject to simplex and linear constraints. We analyze the convergence properties of the algorithm and sharpen its convergence rate. In addition, we also use the convergence analysis to identify an optimal stepsize strategy for weighted entropy projection and an adaptive stepsize strategy for weighted Euclidean projection. Experimental results on synthetic and vision problems demonstrate the effectiveness of our approach.

1 Introduction

MRF energy minimization is a central problem in many computer vision applications. State-of-the-art algorithms to solve the MRF problem can be classified in three methodological frameworks: graph cut [1], belief propagation [2] and LP relaxation. We concentrate on the LP relaxation model for MRF problem. The two common frameworks for the LP relaxation of MRF are based on tree-reweighted message passing [3] and dual decomposition [4]. Message-passing techniques exploit acyclic structures in the MRF models and are known to be efficient. However, the convergence properties of message-passing algorithm is not fully understood. In contrast, the dual decomposition approach is connected to the theory of convex optimization, thus the convergence analysis and suboptimality can be established. In the dual framework, the graphical model is decomposed into easy slave MRFs with favourable properties such as submodular graph, acyclic graph. These slaves can be solved efficiently via dynamic programming and their solutions are used to update the parameters of the master problem in a subgradient projection manner. Compared to other methods, the dual-based approach benefits from better convergence properties and has suboptimality guarantees. Recently, improvements to the Dual Decomposition Sub Gradient technique have been made, including Nesterov's smoothing [5], First Order Primal-Dual method [6], Improved Decomposition [7].

In this paper, we develop a projection algorithm to solve the dual problem of the LP relaxation using weighted Entropy and Euclidean distances. The method is based on Mirror Descent algorithm [8,9] and its generalization on "favourable geometry" domain [10]. We employ a dual decomposition technique as in Komodakis et al. [4] to obtain the dual framework with two types of problems.

G. Bebis et al. (Eds.): ISVC 2012, Part I, LNCS 7431, pp. 587–598, 2012.

The master problem solves a non-smooth objective function subject to linear constraints. The MRFs subproblems can be solved by dynamic programming independently. In the dual LP-based algorithm [4,7], subgradient projection is often used by the master to optimally distribute the data cost between the slaves. Main drawbacks of this approach are slow convergence rate and its sensitivity to the choice of stepsize. In order to address these drawbacks, we transform the domain of dual variables to the intersection of simplexes and linear constraints. The search is performed within the simplexes before proceeding with subgradient method. As a result, our method inherits faster convergence rate from the Mirror Descent algorithm with weighted entropy distance. For the second procedure we employ the weighted Euclidean projection with an adaptive stepsize that shows significant speed up in practice. Our method does not require more memory than any other dual-based methods. The sub problems and all variables are decoupled therefore parallelizing computation is fully supported. In the worst case, this method has an $O(\frac{1}{\epsilon^2})$ complexity whereas the method based on Nesterov's smoothing technique [5,11] provide a convergence rate of $O(\frac{1}{\epsilon})$. However, those methods run an inner loop to compute a good stepsize where each inner iterations require computations of sub MRF problems. It is important to stress that this theoretical comparison is only valid in the worst case. In practice, using good adaptive stepsize strategy for first order method significantly reduces the number of iterations.

The main contributions of this paper are:

- We reformulate the original dual problem and construct the ingredients required for the Mirror Descent algorithm, including weighted distance, weighted norm, dual norm and the local Lipschitz constants.
- We provide the solutions updates using Mirror Descent algorithm for our model.
- Through the convergence analysis, we show that sequential updates by performing Entropy projection before Euclidean projection is better than parallel updates. We also use the bounded optimality to identify the optimal stepsize for entropy projection and adaptive stepsize for euclidean projection.

2 Background

Discrete MRF minimization aims to solve a general graphical multi-labelling problem. Given a set of discrete labels L, the goal is to find a labelling configuration such that it returns the minimal energy on the MRF model specified by an undirected hypergraph $G = (V, E)$ where V and E denote the sets of nodes and edges respectively. Each node $a \in V$ must admit one label from L. By $\theta_{a,i}$, we denote the unary cost of assigning label $i \in L$ to node $a \in V$. The notation $\theta_{ab,ij}$ is used to denote the pairwise cost for edge $ab \in E$. The LP relaxation of the MRF problem is defined as follows:

$$\min_{x \in X} \sum_{a \in V} \sum_{i \in L} \theta_{a,i}.x_{a,i} + \sum_{ab \in E} \sum_{i \in L} \sum_{j \in L} \theta_{ab,ij}.x_{ab,ij} \tag{1}$$

where the constraint set X is known as the *local marginal polytope* [3]. Due to the special structure of the problem above, it turns out that the dual of (1) can be solved efficiently [4]. We write the LP problem compactly as:

$$E(\theta, x) := \min_{x \in X} \langle \theta, x \rangle \tag{2}$$

In the dual approach , the original graph G is decomposed into a collection of trees (*acyclic* graphs) T. Each tree $t \in T$ corresponds to a simpler MRF problem $E^t(\theta^t, x^t)$ that can be solved efficiently by the Max Product Belief Propagation algorithm. Without loss of generality, we assume each tree contains all nodes and every edge must appear only once in T. For example, in a 2D grid graph, one tree contains all horizontal edges and one contains all verticle edges. In this setting, no constraints apply to pairwise cost and the sum of unary costs across the trees must preserve the unary cost of the original graph, ie. $\sum_{t \in T} \theta^t_{a,i} = \theta_{a,i}$. The dual-based algorithm aims to distribute the right amount of unary costs for each tree in order to maximize the dual problem.

$$\max_{\{\theta^t\} \in \Theta} \sum_{t \in T} E^t(\theta^t, x^t) \;\; \text{where} \;\; \Theta = \left\{ \sum_{t \in T} \theta^t = \theta \right\} \tag{3}$$

It is well-known that the solution to problem (3) is the lower bound of the LP problem (2). The key property in dual-based algorithms is to maintain the feasibility set Θ.

Transformation of the Dual Domain: For computational reasons, most methods to solve (3) are based on the subgradient algorithm with Euclidean projection. One disadvantage of this approach is the slow convergent rate. The choice of stepsize significantly affects the algorithm and at every iteration, all unary costs are adjusted by the same amount. In order to address this issue, ie. adjusting the unary cost differently based on the cost itself, we transform the domain of the dual problem such that it still maintains the fesibile set Θ while accelerating the search procedure. Consider the following augmented dual problem:

$$\max_{\rho \in \Delta, \lambda \in \Lambda} F(\rho, \lambda) := \max_{\rho \in \Delta, \lambda \in \Lambda} \textstyle\sum_{t \in T} E^t(\rho^t . \theta + \lambda^t, x^t) \tag{4}$$

$$\text{where:} \quad \Delta = \left\{ \rho \,\middle|\, \textstyle\sum_{t \in T} \rho^t = \mathbf{1} \;,\; \rho \succeq \mathbf{0} \right\} \;\; ; \;\; \Lambda = \left\{ \lambda \,\middle|\, \textstyle\sum_{t \in T} \lambda^t = \mathbf{0} \right\}$$

It is easy to see that the sets Δ and Λ preserve Θ. The augmented model has the same optimal objective function value as the original dual problem. Notice that if we choose a constant $\rho \in \Delta$, then our model is equivalent to Komodakis et al. [4]. The objective function $F(\rho, \lambda)$ is linear in both variables; in addition, ρ and λ are completely decoupled.

3 Mirror Descent (MD)

Mirror Descent algorithm [8,10] is a generalization of the proximal algorithm [12] with a nonlinear distance function [9] and an optimal stepsize. In order to utilize

the Mirror Descent algorithm with the special structure of our augmented dual model, we need to define the subgradient, the weighted distances and weighted norm which favour the problem's geometry. Instead of solving the augmented dual problem (4) directly, we generate a sequence of updates:

$$\begin{bmatrix}\rho^{k+1}\\ \lambda^{k+1}\end{bmatrix} = \underset{\rho\in\Delta,\lambda\in\Lambda}{\arg\max}\begin{Bmatrix}\langle F'(\rho^k),\rho\rangle - \frac{1}{\tau}D_\Delta(\rho,\rho^k)\\ +\langle F'(\lambda^k),\lambda\rangle - \frac{1}{\eta}D_\Lambda(\lambda,\lambda^k)\end{Bmatrix} \tag{5}$$

Since the function F is linear in both variables, ρ and λ are decoupled, the subgradients with respect to each variable are also disjoint. The weighted distances D_Δ, D_Λ and stepsizes τ, η are defined independently to exploit the geometry of each set. To simplify our notation, we define an index set to cover all unary terms: $I = \{(a,i)|\forall a \in V, \forall i \in L\}$. The domains Δ and Λ are built by taking the direct product of the disjoint subsets:

$$\Delta :=_\otimes \Delta_i \ ; \ \ \Lambda :=_\otimes \Lambda_i \ \ , \ \forall i \in I$$

Let $T(i)$ be the collection of trees that cover the same unary term i, then each subset reads:

$$\Delta_i = \left\{ \sum_{t\in T(i)} \rho_i^t = 1 \ , \ \rho_i^t \geq 0 \right\} \ ; \ \Lambda_i = \left\{ \sum_{t\in T(i)} \lambda_i^t = 0. \right\}$$

Subgradient: The following lemma shows how the subgradient is estimated in our algorithm.

Lemma 1. *Let $F'(\rho,\lambda)$ be defined as follows: $F'(\rho,\lambda) = [\theta.\bar{x};\bar{x}]$, where $\bar{x} \in \arg\min_{x\in X}\langle\rho.\theta+\lambda, x\rangle$. Then $F'(\rho,\lambda) \in \partial F(\rho,\lambda)$, where $\partial F(\rho,\lambda)$ denotes the subgradient of $F(\rho,\lambda)$ at the point (ρ,λ).*

Proof. The point $\bar{x}$ is suboptimal for $\min_{x\in X}\langle\rho'.\theta+\lambda', x\rangle$, therefore:

$$\begin{aligned} F(\rho',\lambda') &\leq \langle\rho'.\theta+\lambda',\bar{x}\rangle = \langle\rho.\theta+\lambda,\bar{x}\rangle + \langle\theta.\bar{x},\rho'-\rho\rangle + \langle\bar{x},\lambda'-\lambda\rangle \\ F(\rho',\lambda') &\leq F(\rho,\lambda) + \langle\theta.\bar{x},\rho'-\rho\rangle + \langle\bar{x},\lambda'-\lambda\rangle \end{aligned}$$

as required by subgradient inequality. □

Distance Function: MD generates a projection based on nonlinear distance function. Let D_C^i denotes a Bregman distance function defined on a single closed convex set C_i:

$$D_C^i(u_i,v_i) = \psi_C^i(u_i) - \psi_C^i(v_i) - \langle\nabla\psi_C^i(v_i), u_i - v_i\rangle$$

where $u_i, v_i \in C_i$ and $\psi_C^i : C_i \to \mathbb{R}$ is a *1-strongly* convex *distance-generating-function (d.g.f)*. The weighted distance function D_C defined on the domain $C :=_\otimes C_i$ is thus given by:

$$\begin{aligned} D_C(u,v) &= \sum_{i\in I}\alpha^i D_C^i(u_i,v_i) = \sum_{i\in I}\alpha^i\left[\psi_C^i(u_i) - \psi_C^i(v_i) - \langle\nabla\psi_C^i(v_i), u_i - v_i\rangle\right] \\ &:= \psi_C(u) - \psi_C(v) - \langle\nabla\psi_C(v), u - v\rangle \end{aligned}$$

where $\alpha_C^i > 0$ is the weighted parameter. The weighted d.g.f defined on C is:

$$\psi_C(u) = \sum_{i \in I} \alpha^i \psi_C^i(u_i) \tag{6}$$

Norm: Another requirement for MD is its *compatible* norm, ie. weighted d.g.f $\psi_C : C \to \mathbb{R}$ is *1-strongly* convex w.r.t the weighted norm $\|.\|_C$ [10]:

$$\|u\|_C = \sqrt{\sum_{i \in I} \alpha^i \|u_i\|_{C_i}^2} \tag{7}$$

In the formulation above, $\|.\|_{C_i}$ is a local norm that is defined based on the geometry of a subset C_i.

Dual Norm and the Local Lipschitz Constant:
From the definition of Dual Norm [13], we can derive the dual norm of (7):

$$\|u\|_{C*} = \sqrt{\sum_{i \in I} \|u_i\|_{C_i*}^2 / \alpha^i}$$

Let $\mathcal{L}_{C_i} := \sup_{u_i \in C_i} \|F'_u\|_{C_i*}$ then the *local* Lipschitz constant is given by:

$$\mathcal{L}_C = \sup_{u \in C} \|F'_u\|_{C*} = \sqrt{\sum_{i \in I} \mathcal{L}_{C_i}^2 / \alpha^i} \tag{8}$$

Note that in our notation, we refer the Lipschitz constant as *local* since it depends on the specific choice of subgradient.

Weighted Entropy Distance. With the general definitions of distance and norm, we define the *weighted entropy* distance D_Δ over the domain Δ using *entropy* d.g.f ψ_Δ^i and l_1-norm $\|.\|_1$ on individual set Δ_i:

$$\psi_\Delta^i(\rho_i) = \sum_{t \in T(i)} \rho_i^t \ln \rho_i^t \; ; \; \|\rho\|_\Delta = \sqrt{\sum_{i \in I} \alpha_\Delta^i \|\rho_i\|_1^2} \tag{9}$$

Lemma 2. *Let $\psi_\Delta : \Delta \to \mathbb{R}$ be the weighted d.g.f (6) defined with summand ψ_Δ^i. Then ψ_Δ is 1-strongly convex w.r.t the norm $\|.\|_\Delta$*

Proof.
$$\begin{aligned} \langle \nabla\psi_\Delta(u) - \nabla\psi_\Delta(v), u - v \rangle &= \sum_{i \in I} \alpha_\Delta^i \langle \nabla\psi_\Delta^i(u_i) - \nabla\psi_\Delta^i(v_i), u_i - v_i \rangle \\ &\geq \sum_{i \in I} \alpha_\Delta^i \|u_i - v_i\|_1^2 = \|u - v\|_\Delta^2 \end{aligned} \tag{10}$$

The inequality in (10) comes from the well known *1-strongly convex* property of entropy function ψ_Δ^i over the simplex Δ_i w.r.t l_1-norm [9]. □

Weighted Entropy Distance. Weighted Euclidean distance D_Λ is defined by summand ψ^i_Λ and l_2-norm on subset Λ_i as follow:

$$\psi^i_\Lambda(\lambda_i) = \frac{1}{2}\lambda_i^\top \lambda_i \; ; \; \|\lambda\|_\Lambda = \sqrt{\sum_{i\in I} \alpha^i_\Lambda \|\lambda_i\|_2^2} \tag{11}$$

Lemma 3. *Let $\psi_\Lambda : \Lambda \to \mathbb{R}$ be the weighted d.g.f* (6) *defined with summand ψ^i_Λ. Then ψ_Λ is 1-strongly convex w.r.t the norm $\|.\|_\Lambda$.*

Proof. The proof is similar to Lemma 2. □

Solution Updates: Using the weighted distance D_Δ and D_Λ above, we obtain the solutions to proximal sequence (5):

$$\begin{aligned} \rho_i^{k+1(t)} &= \frac{\rho_i^{k(t)} \exp\left(F'(\rho^k)^t_i . \tau/\alpha^i_\Delta\right)}{\sum\limits_{t\in T(i)} \rho_i^{k(t)} \exp\left(F'(\rho^k)^t_i . \tau/\alpha^i_\Delta\right)} \\ \lambda_i^{k+1(t)} &= \frac{\eta}{\alpha^i_\Lambda}\left(F'(\lambda^k)^t_i - \frac{\sum_{t\in T} F'(\lambda^k)^t_i}{T_i}\right) \end{aligned} \tag{12}$$

where T_i denotes the number of trees that cover the unary term i. It is straightforward to see that variable updates only happen at the nodes which are assigned different labels across the trees. In addition, the stepsize update for each unary term is affected by the weighting factor associated with that term. Through the convergence analysis below, we derive an optimal stepsize for entropy projection and adaptive stepsize for euclidean projection.

4 Convergence Analysis

The MD iterations (5) solve for ρ and λ independently. Since the two variables are disjoint, MD can either update them simultenously or sequentially. By examining the convergence analysis, we justify that updating sequentially provides a faster convergence rate, ie. MD updates ρ first, until there is no improvement in the dual, then it switchs to update λ. In addition, we define the optimal stepsize for entropy projection and an adaptive stepsize for euclidean projection based on the bounded sub-optimality.

4.1 Convergence Analysis

Lemma 4. *The proximal update* (5) *provides better sub-optimality in sequential manner than in parallel manner. It has the following worst case optimality:*

$$F^* - \max_{k=1..K}\{F_k\} \leq \frac{\sqrt{2}(\mathcal{L}_\Delta\sqrt{\Omega_\Delta} + \mathcal{L}_\Lambda\sqrt{\Omega_\Lambda})}{\sqrt{K}} \tag{13}$$

Proof. (Sketch) Assume we have a sequence of updates $\{\rho^k\}_{k=1}^{k_1}, \{\lambda^k\}_{k=k_1+1}^{k_1+k_2}$, follows the proof of Proposition 1.1 (i) as in [10] with ingredients of MD algorithm

that we developed in Section 3, we can obtain the following inequality:

$$\langle F'(\rho_k), \rho^* - \rho^k \rangle \leq \frac{1}{\tau} D_\Delta(\rho^*, \rho^k) - \frac{1}{\tau} D_\Delta(\rho^*, \rho^{k+1}) + \frac{\tau \|F'(\rho^k)\|^2_{\Delta *}}{2} \quad (14)$$

$$\langle F'(\lambda_k), \lambda^* - \lambda^k \rangle \leq \frac{1}{\eta} D_\Lambda(\lambda^*, \lambda^k) - \frac{1}{\eta} D_\Lambda(\lambda^*, \lambda^{k+1}) + \frac{\eta \|F'(\lambda^k)\|^2_{\Lambda *}}{2} \quad (15)$$

Let $K = k_1 + k_2$, $\hat{F} = \max_k\{F_k\}$. Summing up (14) and (15) over K iterations:

$$\begin{aligned} K(F^* - \hat{F}) &\leq \sum_{k=1}^{K} (F^* - F_k) \leq \sum_{k=1}^{k_1} \langle F'(\rho_k), \rho^* - \rho^k \rangle + \sum_{k=k_1+1}^{K} \langle F'(\lambda_k), \lambda^* - \lambda^k \rangle \\ &\leq \frac{D_\Delta(\rho^*, \rho^1)}{\tau} + \frac{D_\Lambda(\lambda^*, \lambda^{k_1+1})}{\eta} + \frac{k_1 \tau \mathcal{L}^2_\Delta + k_2 \eta \mathcal{L}^2_\Lambda}{2} \end{aligned}$$

where $\mathcal{L}_\Delta$ and $\mathcal{L}_\Lambda$ are the local Lipschitz constants. Let $\Omega_\Delta = \max_{\rho \in \Delta} D_\Delta$ and $\Omega_\Lambda = \max_{\lambda \in \Lambda} D_\Lambda$, then:

$$F^* - \hat{F} \leq \frac{\Omega_\Delta}{K\tau} + \frac{\Omega_\Lambda}{K\eta} + \frac{1}{2}\left(\frac{k_1}{K}\tau \mathcal{L}^2_\Delta + \frac{k_2}{K}\eta \mathcal{L}^2_\Lambda\right) \quad (16)$$

Inequality (16) gives the bounded sub-optimality when updates are done in sequential manner. Let B denotes the RHS of (16). If parallel updates are used, then $k_1 = k_2 = K$, and we have:

$$B \leq \frac{\Omega_\Delta}{K\tau} + \frac{\Omega_\Lambda}{K\eta} + \frac{1}{2}\left(\tau \mathcal{L}^2_\Delta + \eta \mathcal{L}^2_\Lambda\right) \quad (17)$$

From inequality (17), we can justify that sequential updates provide better sub-optimal approximation than parrallel updates. Minimizing the RHS of (17) w.r.t τ and η gives:

$$\tau = \frac{\sqrt{2\Omega_\Delta}}{\mathcal{L}_\Delta \sqrt{K}} \;\; ; \;\; \eta = \frac{\sqrt{2\Omega_\Lambda}}{\mathcal{L}_\Lambda \sqrt{K}} \quad (18)$$

Hence, the rate of convergence is bounded by:

$$F^* - \hat{F} \leq \frac{\sqrt{2}(\mathcal{L}_\Delta \sqrt{\Omega_\Delta} + \mathcal{L}_\Lambda \sqrt{\Omega_\Lambda})}{\sqrt{K}}$$

Theorem 1. *The sequential updates generated by the MD algorithm provides the following bound on sub-optimality:*

$$F^* - \max_{k=1..K}\{F_k\} \leq \frac{\sqrt{2}\left(\sum_{i \in I} |\theta_i| \sqrt{\ln(T_i)} + |\lambda^*_i| T_i\right)}{\sqrt{K}} \quad (19)$$

Proof. We want to minimize the RHS of (13). The parameters associate with two disjoint domains can be computed independently. Consider minimizing an arbitrary term:

$$\Omega\mathcal{L}^2 = \left[\sum_{i\in I}\alpha^i\Omega^i\right]\left[\sum_{i\in I}\alpha_i^{-1}\mathcal{L}_i^2\right] \tag{20}$$

Optimising (20) w.r.t α, we obtain:

$$\alpha^i = \frac{\mathcal{L}_i}{\sqrt{\Omega^i}\left[\sum_{i\in I}\mathcal{L}_i\sqrt{\Omega^i}\right]}$$

Therefore, $\Omega = 1$ and $\mathcal{L} = \sum_{i\in I}\mathcal{L}_i\sqrt{\Omega_i}$.

The Lipschitz Constant over the Domain Δ

Each subset Δ_i is equipped with $\|.\|_1$, therefore $\mathcal{L}_i^{\Delta} = \sup_{\rho_i\in\Delta_i}\|F_i'(\rho)\|_\infty = |\theta_i|$ In addition, the maximum distance Ω_Δ^i over the simplex Δ_i is defined in Proposition 5.1 (c) [9]: $\Omega_\Delta^i = \ln(T_i)$. Hence, we have: $\mathcal{L}_\Delta = \sum_{i\in I}|\theta_i|\sqrt{\ln(T_i)}$

The Lipschitz Constant over the Domain Λ

Similarly, we can compute:

$$\mathcal{L}_i^{\Lambda} = \sup_{\lambda_i\in\Lambda_i}\|F_i'(\lambda)\|_2 = \sqrt{T_i}\ \ ;\ \ \Omega_\lambda^i = T_i(\lambda_i^*)^2\ \ ;\ \ \mathcal{L}_\lambda = \sum_{i\in I}|\lambda_i^*|T_i$$

Note that, the amount λ_i^* and its maximum distance Ω_λ^i can only estimated due to the "unbounded" nature of the set Λ_i. Finally, we obtain the bound (19):

$$F^* - \hat{F}_{k=1..K} \leq \frac{\sqrt{2}\left(\sum_{i\in I}|\theta_i|\sqrt{\ln(T_i)} + |\lambda_i^*|T_i\right)}{\sqrt{K}}$$

Remarks. From Lemma 4, we have justified that sequential updates is better than parallel updates. Now, let us consider two other type of updates: using weighted entropy projection only and using weighted Euclidean projection only. Clearly, using weighted entropy only will get trapped into a local maxima because the set Δ does not cover the original feasible set Θ. However, in applications where the simplexes fully cover the original feasible set, we can obtain very fast convergence. If we use only weighted Euclidean projection to search within the same space defined by Δ, then in the worst-case, the sub-optimality is defined by:

$$\frac{\sqrt{2}\left(\sum_{i\in I}|\theta_i|T_i + |\lambda_i^*|T_i\right)}{\sqrt{K}}$$

It is easy to see that this bound is much larger than our optimal bound in the RHS of (19) as the size of the set I is very large.

4.2 Discussion

Switching Criteria: An intuitive idea is to derive switching criteria based on dual gap. When the MD sequence based on entropy projection finds a sub-optimal distribution in its domain, it will not improve the dual further. One

can define a switching point when there is no improvement in the dual objective or dual gap. However, subgradient method often fluctuates the dual objective, thus, the dual gap appears to have zic-zagging behaviour. Therefore it is not efficient to detect switching point based on the dual gap. On the other hand, an important feature of the Dual Decomposition model is that as the method converges, the number of non-agreement nodes is decreasing. This observation works better in general since it does not fluctuate as much as the dual objective. We define a threshold $\sigma \in \{1, .., 20\}$ depends on applications for switching to Euclidean projection when the number non-agreement nodes does not decrease after σ iterations.

Implementation: The proximal updates are done in sequence where we solve for ρ until the switching criteria is met, then we solve for λ. With the ingredients developed sofar, the sequence (12) reduces to:

$$\rho^t = \frac{\theta_i^t}{\sum_{t \in T} \theta_i^t} \;\; ; \;\; \omega^t = \varepsilon . \operatorname{sign}(\theta_i^t . \bar{x}_i^t) \sqrt{2 \ln(T_i)/k} \;\; ; \;\; \rho^t = \frac{\rho^t \exp(\omega^t)}{\sum_{t \in T(i)} \rho^t \exp(\omega^t)}$$

$$\theta_i^t = \rho^t \left(\sum_{t \in T} \theta_i^t \right) \tag{21}$$

$$\theta_i^t = \theta_i^t + \sqrt{2 . \frac{\Omega_i}{k}} \left(\bar{x}_i^t - \frac{\sum_{t \in T} \bar{x}_i^t}{T_i} \right) \tag{22}$$

where $\varepsilon \in (0, 2)$ is a speed up parameter. The entropy projection updates the master's parameter by (21). Equation (22) is used for Euclidean projection. Since we can compute ρ based on the current value of unary terms, the memory required is not more than any other type of dual decomposition methods.

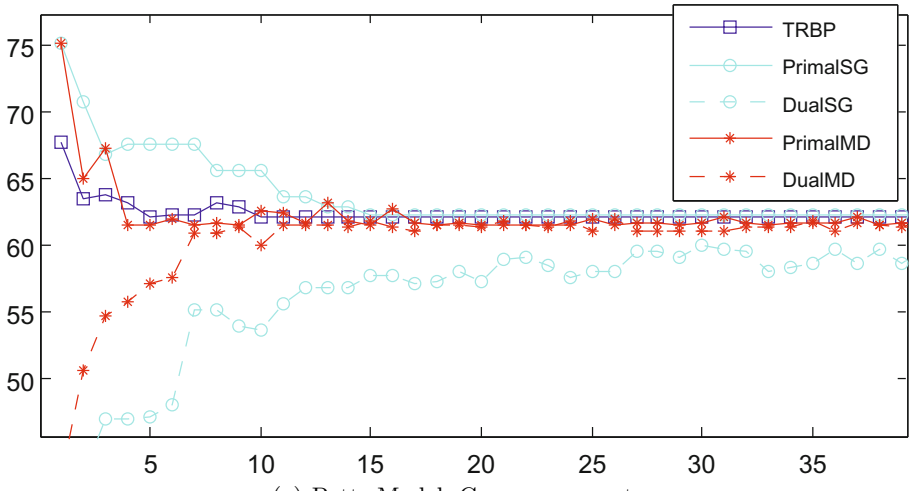
Adaptive Stepsize: The stepsize for entropy projection is optimal and can be computed analytically since we know the maximum distance of the simplex. However, we do not know the maximum distance Ω_i on the unbounded set Λ_i, therefore we estimate it by:

$$\sqrt{\frac{\Omega_i}{k}} \approx \frac{|\hat{E} - \hat{F}|}{T_i . L_k} \tag{23}$$

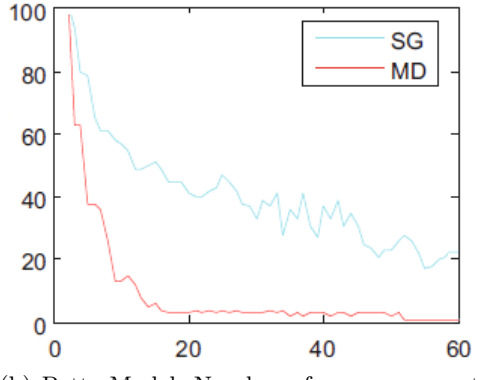
where $\hat{E}$ is the best primal solution after k iterations. At iteration k, there is L_k number of non-agreement nodes that need to be adjusted to reduce the dual gap $|\hat{E} - \hat{F}|$. The difference between primal and dual is distributed evenly for L_k nodes. In addition, for each node, this amount is dispensed evenly amongst the trees that cover the node.

5 Experiments

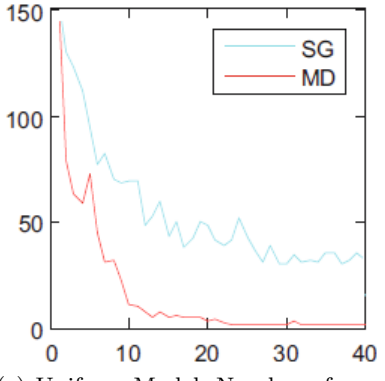
In order to demonstrate the effectiveness of our method, we present experimental results with synthetic data and segmentation problem with the UGM Matlab package [6]. In addition, we also examine our method with the Tsukuba



(a) Potts Model: Convergence rate



(b) Potts Model: Number of non-agreement Nodes



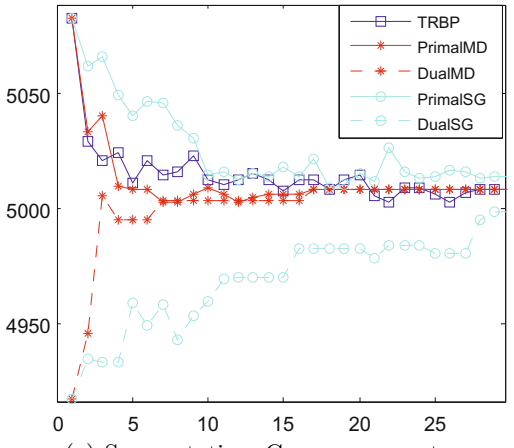
(c) Uniform Model: Number of non-agreement Nodes

Fig. 1. Synthetic data

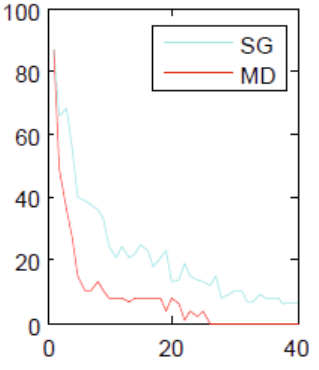
Stereo problem in MRF-Bechmark package [14]. In all experiments, we apply three methods: Tree-reweighted variants (TRBP in UGM and TRW-S in MRF-benchmarks), Mirror Descent and Sub Gradient with adaptive stepsize $\alpha = \frac{|\hat{E}-\hat{F}|}{\|F'_k\|^2}$ as suggested in [4].

Synthetic Data: For our synthetic experiments, we used a grid graph of size 20×20 and 5 labels. For the Potts model, $\theta_{a,i}$ was drawn from $\mathcal{U}(-1,+1)$, while $\theta_{ab,ij} = \omega_{ab} * \mathbb{I}(i=j)$ and $\omega_{ab} = \mathcal{N}(0,1)$. For the Uniform model, we withdraw all data from $\mathcal{U}(0,1)$, for the edge weight, we also use $\omega_{ab} = \mathcal{N}(0,1)$. For these small tests, we set the switching threshold to 5.

Figures 1(a) and 1(b) shows the convergence of primal-dual gap and number of labels to fix for the Potts model. The switch between methods occur between iterations 20 and 25. All methods converge eventually, however our method



(a) Segmentation: Convergence rate



(b) Segmentation: Number of non-agreement Nodes

Fig. 2. Segmentation data

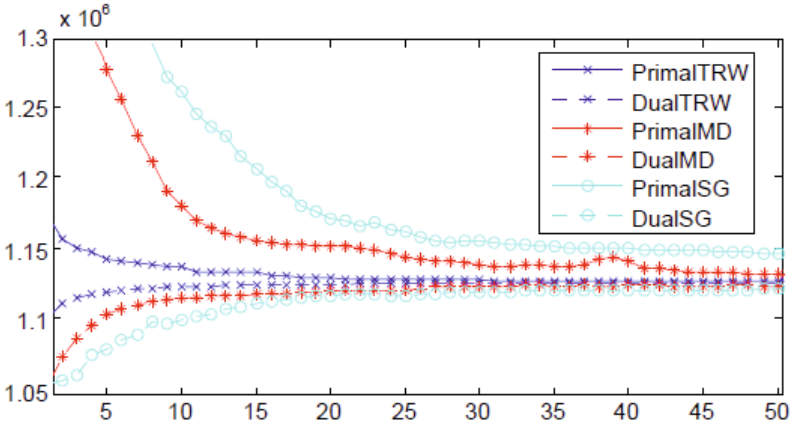


Fig. 3. Stereo: Convergence

outperforms the sub-gradient method significantly and obtains the optimal solution slightly before TRW. In the Uniform model (Figure 1(c)), the switch even not happen, MD can compute optimal labelling by entropy projection only.

Segmentation Probem: The segmentation problem is to recover a coloured X from its noisy image in the UGM package. Figures 2(a) and 2(b) show how the methods perform. Note the switching step happens between 15 and 20 iterations. After the switch to the Euclidean projection, with our adaptive stepsize, MD can recover the optimal solution at around iteration 25.

The Stereo Problem: Figure 3 shows the convergence rate for Tsukuba problem with three methods. We can see that TRW still converges fastest, with the MD method comes second. Both TRW and MD generate similar dual objective sequence after 30 iterations.

6 Conclusion

An efficient algorithm to solve the dual MRF minimization problem is presented. The method is based on Mirror Descent algorithm with weighted distance projections, weighted norms and local Lipschitz constants. After a careful analysis of the algorithm, we are able to sharpen the theoretical convergence rate as well as to improve the performance of the algorithm in practice. Mirror Descent can be applied efficiently on any bounded set, a direction of future research is to establish the relationship between the dual gap and feasible sets, and address the possibility of performing entropy projection on the unbounded set Θ.

References

1. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. 23, 1222–1239 (2001)
2. Yedidia, J.S., Freeman, W.T., Weiss, Y.: Understanding belief propagation and its generalizations. In: Exploring Artificial Intelligence in the New Millennium, pp. 239–269 (2003)
3. Wainwright, M.J., Jaakkola, T.S., Willsky, A.S.: Map estimation via agreement on (hyper)trees: Message-passing and linear-programming approaches. IEEE Trans. on Information Theory 51, 3697–3717 (2005)
4. Komodakis, N., Paragios, N., Tziritas, G.: Mrf energy minimization and beyond via dual decomposition. IEEE Trans. Pattern Anal. Mach. Intell. 33, 531–552 (2011)
5. Savchynskyy, B., Schmidt, S., Kappes, J., Schnorr, C.: A study of nesterov's scheme for lagrangian decomposition and map labeling. In: Computer Vision and Pattern Recognition, pp. 1817–1823 (2011)
6. Schmidt, M.: Ugm: Matlab code for undirected graphical models (2011)
7. Jancsary, J., Matz, G.: Convergent decomposition solvers for tree-reweighted free energies. Journal of Machine Learning Research
8. Ben-tal, A., Margalit, T., Nemirovski, A.: The ordered subsets mirror descent optimization method with applications to tomography. SIAM Journal on Optimization 12 (2001)
9. Beck, A., Teboulle, M.: Mirror descent and nonlinear projected subgradient methods for convex optimization. Operations Research Letters (2003)
10. Juditsky, A., Nemirovski, A.: First order methods for nonsmooth convex large-scale optimization, i: General purpose methods. In: Optimization for Machine Learning. MIT Press (2010)
11. Jojic, V., Gould, S., Koller, D.: Accelerated dual decomposition for map inference. In: International Conference of Machine Learning, pp. 503–510 (2010)
12. Censor, Y., Zenios, S.A.: Proximal minimization algorithm with d-functions. Journal of Optimization Theory and Applications 73, 451–464 (1992)
13. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004)
14. Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., Rother, C.: A comparative study of energy minimization methods for mrfs. IEEE Trans. Pattern Anal. Mach. Intell., 1068–1080 (2008)