Message-Passing Algorithms for MAP Estimation Using DC Programming (Supplementary Material)

Akshat Kumar
Univ. of Massachusetts Amherst

Shlomo Zilberstein
Univ. of Massachusetts Amherst

Marc Toussaint
Freie Universität Berlin

A Detailed Proofs and Derivations

A.1 Proposition 3

For the QP edges, it holds that $\mu_{ij}(x_i, x_j) = \mu_i(x_i)\mu_j(x_j)$. Instead of having a linear objective function $(\boldsymbol{\mu}\cdot\boldsymbol{\theta})$, we can substitute $\mu_{ij}(x_i, x_j)$ by $\mu_i(x_i)\mu_j(x_j)$ in the objective. Thus we no longer need to store the parameter $\mu_{ij}(x_i, x_j)$ nor the mean-field constraint explicitly for QP edges. Therefore, the total number of parameters is $O(k^2|L| + nk)$ and the total number of constraints is $O(2|L|k + n)$ where $L = E\backslash Q$. As the size of Q increases by 1, the size of the set L decreases by 1. This proves the proposition.

A.2 Proposition 4

The optimization problem involving the function $g(\boldsymbol{\mu}, \boldsymbol{y}; \theta, Q)$ over Ω' is given by[1]:

$$\min_{\boldsymbol{\mu},\boldsymbol{y}} -\sum_{(i,j)\in Q}\sum_{x_i,x_j}\theta(x_i,x_j)e^{y(x_i)+y(x_j)} - \sum_{(i,j)\in L}\sum_{x_i,x_j}\theta(x_i,x_j)\mu(x_i,x_j) \tag{19}$$

$$\textit{subject to}: \sum_{x_i,x_j}\mu(x_i,x_j) = 1\ \forall(i,j)\in L;$$

$$\sum_{\hat{x}_j}\mu_{ij}(x_i,\hat{x}_j) = e^{y(x_i)}\ \forall i\in V, \forall x_i, \forall Nb_l(i) \tag{20}$$

The Lagrangian $L(\boldsymbol{\mu}, \boldsymbol{y}, \lambda)$ is given by:

$$L(\boldsymbol{\mu},\boldsymbol{y},\lambda) = -\sum_{(i,j)\in Q}\sum_{x_i,x_j}\theta(x_i,x_j)e^{y(x_i)+y(x_j)} - \sum_{(i,j)\in L}\sum_{x_i,x_j}\theta(x_i,x_j)\mu(x_i,x_j) + \sum_{ij}\lambda_{ij}\Big\{\sum_{x_i,x_j}\mu(x_i,x_j)-1\Big\} + \sum_{i\in V}\sum_{j\in Nb_l(i)}\sum_{x_i}\lambda_{ji}(x_i)\Big\{\sum_{x_j}\mu_{ij}(x_i,x_j) - e^{y(x_i)}\Big\} \tag{21}$$

Notice that in the above Lagrangian, we ignored the nonnegativity constraints associated with $\mu(x_i, x_j) \geq 0$ for simplicity as they remain the same for both versions of the optimization problem. Now the KKT conditions for the stationary point are given by:

$$\frac{\partial L}{\partial\mu(x_i,x_j)} = -\theta(x_i,x_j) + \lambda_{ij} + \lambda_{ji}(x_i) + \lambda_{ij}(x_j) = 0 \tag{22}$$

The second KKT condition is:

$$\frac{\partial L}{\partial y(x_i)} = -\sum_{j\in Nb_q(i)}\sum_{x_j}\theta(x_i,x_j)e^{y(x_i)+y(x_j)} - \sum_{j\in Nb_l(i)}\lambda_{ji}(x_i)e^{y(x_i)} = 0 \tag{23}$$

which can be further simplified to:

$$\sum_{j\in Nb_q(i)}\sum_{x_j}\theta(x_i,x_j)e^{y(x_j)} + \sum_{j\in Nb_l(i)}\lambda_{ji}(x_i) = 0 \tag{24}$$

Now consider the second optimization problem of optimizing $g(\boldsymbol{\mu}; \theta, Q)$ over Ω given as:

$$\min_{\boldsymbol{\mu},\boldsymbol{y}} -\sum_{(i,j)\in Q}\sum_{x_i,x_j}\theta(x_i,x_j)\mu(x_i)\mu(x_j) - \sum_{(i,j)\in L}\sum_{x_i,x_j}\theta(x_i,x_j)\mu(x_i,x_j) \tag{25}$$

$$\textit{subject to}: \sum_{x_i,x_j}\mu(x_i,x_j) = 1\ \forall(i,j)\in L;$$

$$\sum_{\hat{x}_j}\mu_{ij}(x_i,\hat{x}_j) = \mu(x_i)\ \forall i\in V, \forall x_i, \forall Nb_l(i) \tag{26}$$

Notice that every feasible point $(\boldsymbol{\mu'}, \boldsymbol{y'}) \in \Omega'$ of the first optimization problem can be transformed into a unique feasible point $\boldsymbol{\mu} \in \Omega$ of the second problem simply by setting $\mu(\cdot,\cdot) = \mu'(\cdot,\cdot)$ and setting $\mu(\cdot) = e^{y(\cdot)}$. Thus each stationary point of the first optimization problem is a feasible point of the second problem. We now show that this feasible point indeed satisfies the KKT conditions of the second optimization problem too, thus prov-

[1] Equation numbers continue the main paper.

ing the proposition. The Lagrangian $L(\boldsymbol{\mu},\lambda)$ is given by:

$$L(\boldsymbol{\mu},\lambda) = -\sum_{(i,j)\in Q}\sum_{x_i,x_j}\theta(x_i,x_j)\mu(x_i)\mu(x_j) - \sum_{(i,j)\in L}\sum_{x_i,x_j}\theta(x_i,x_j)\mu(x_i,x_j) + \sum_{ij}\lambda_{ij}\Big\{\sum_{x_i,x_j}\mu(x_i,x_j)-1\Big\} + \sum_{i\in V}\sum_{j\in Nb_l(i)}\sum_{x_i}\lambda_{ji}(x_i)\Big\{\sum_{x_j}\mu_{ij}(x_i,x_j)-\mu(x_i)\Big\} \tag{27}$$

Now the KKT conditions for the stationary point are:

$$\frac{\partial L}{\partial \mu(x_i,x_j)} = -\theta(x_i,x_j)+\lambda_{ij}+\lambda_{ji}(x_i)+\lambda_{ij}(x_j) = 0 \tag{28}$$

The above KKT condition matches exactly with the KKT condition of Eq. (22).

The second KKT condition is:

$$\frac{\partial L}{\partial \mu(x_i)} = -\sum_{j\in Nb_q(i)}\sum_{x_j}\theta(x_i,x_j)\mu(x_j) - \sum_{j\in Nb_l(i)}\lambda_{ji}(x_i) = 0 \tag{29}$$

The above condition is also exactly the same as of the condition in Eq. (24) by noting that $\mu(x_j) = e^{y(x_j)}$. Thus we have shown that every stationary point of the first optimization problem is a feasible point of the second optimization and also satisfies the KKT conditions.

A.3 Proposition 5

The above proposition can be proved using Zangwill's global convergence theorem [2] which has been used to prove the convergence of CCCP for the convex constraints [1, Thm. 4]. They also show another variant of CCCP with non-convex constraints also converges to a stationary point [1, Sec. 4.1]. In our case, we wish to prove the convergence of CCCP for the optimization problem $\min g(\boldsymbol{\mu},\boldsymbol{y};\theta,Q)$ of Eq. (19) over constraints Ω' of Eq. (20). In principle, we can use the analysis of [1, Sec. 4.1] that handles CCCP with non-convex constraints, however we do not use the variant of CCCP presented in that section as we have non-convex equality constraints rather than D.C. inequality constraints. For a brief overview of the Zangwill's theorem, we refer to [1, Thm. 2]. We will also use some background terms from [1] such as the notion of point-to-set map, details can be found in that paper.

Roughly speaking, the CCCP iteration of Eq. (12) defines a point-to-set map $x^{l+1} = A_{cccp}(x^l)$ where A_{cccp} is the optimization problem of Eq. (12). The main idea to prove the convergence of CCCP is two fold. First we show that a fixed point of A_{cccp} is also a stationary point of the D.C. program of Eq. (19). The fixed point $x^\star$ of A_{cccp} is given by the condition $x^\star = A_{cccp}(x^\star)$. This can be easily shown by writing the KKT conditions for A_{cccp} at $x^\star$ and showing that they also satisfy the KKT conditions for the D.C. program of Eq. (19) similar to Appendix A.2. This condition holds in our case; we skip the proof for brevity.

The second step is to show that the limit points of any sequence $\{x^l\}_{l=0}^{\infty}$ generated by A_{cccp} are the fixed points of A_{cccp}. This can be shown by using the conditions of [1, Thm. 2]. We do not show the proof in detail as it is similar to the proof of convergence of CCCP with *convex* constraints [1, Thm. 4]. We provide high level arguments as follows. The main reason is that although our original D.C. program of Eq. (19) has non-convex constraints, the CCCP iteration A_{cccp} we proposed in Eq. (12) is a convex optimization problem with linear equality constraints. Therefore the convergence of A_{cccp} is implied by [1, Thm. 4], which only requires A_{cccp} to be a convex optimization problem and be uniformly compact on the constraint set. We also highlight that the [1, Remark 7] holds in our case as the constraint set for A_{cccp} in Eq. (13) is compact.

A.4 Reinterpretation of the dual updates in terms of primal parameters

This appendix derives the updates used in the inner loop of Alg. 1. Let each step of dual coordinate ascent be indexed by superscripts τ starting from zero. Initially, we set all multipliers λs to zero. Let the outer loop iterations be indexed by subscripts n and let the current outer loop iteration be $n+1$. So we have:

$$\mu(x_i,x_j) = e^{\left\{\theta(x_i,x_j)+\nabla_{\mu(x_i,x_j)}v+\lambda_{ij}(x_j)+\lambda_{ji}(x_i)-\lambda_{ij}-1\right\}}$$

$$\mu^{\mathbf{0}}(\mathbf{x_i},\mathbf{x_j}) = \mu_{\mathbf{n}}(\mathbf{x_i},\mathbf{x_j})\exp\left\{\theta(\mathbf{x_i},\mathbf{x_j})\right\}$$

$$\mu(x_i) = \frac{\nabla_{y(x_i)}v}{1+\sum_{k\in Nb_l(i)}\lambda_{ki}(x_i)}$$

$$\mu^{\mathbf{0}}(\mathbf{x_i}) = \nabla_{\mathbf{y(x_i)}}\mathbf{v}$$

For any inner loop iteration τ, we can realize the intermediate beliefs as:

$$\mu^\tau(x_i) = \frac{\nabla_{y(x_i)}v}{1+\sum_{k\in Nb_l(i)}\lambda^\tau_{ki}(x_i)}$$

$$\mu^\tau(x_i,x_j) = \mu_n(x_i,x_j)e^{\left\{\theta(x_i,x_j)+\lambda^\tau_{ij}(x_j)+\lambda^\tau_{ji}(x_i)-\lambda^\tau_{ij}\right\}}$$

Let us first consider the dual update for $\lambda^{\tau+1}_{ij}(x_j)$.

$$\lambda^{\tau+1}_{ij}(x_j) = W\Big[\frac{\nabla_{y(x_j)}v\; e^{\sum_{k\in Nb_l(j)\setminus i}\lambda^\tau_{kj}(x_j)+1}}{\sum_{x_i}\mu_n(x_i,x_j)\exp\{\theta(x_i,x_j)+\lambda^\tau_{ji}(x_i)-\lambda^\tau_{ij}\}}\Big] - 1 - \sum_{k\in Nb_l(j)\setminus i}\lambda^\tau_{kj}(x_j)$$

We can simplify the argument Z^τ of the lambert W-function as follows:

$$Z^\tau = \frac{\nabla_{y(x_j)} v \; e^{\sum_{k \in Nb_l(j) \setminus i} \lambda^\tau_{kj}(x_j) + 1}}{\sum_{x_i} \mu_n(x_i, x_j) \exp\{\theta(x_i, x_j) + \lambda^\tau_{ji}(x_i) - \lambda^\tau_{ij}\}} \frac{e^{\lambda^\tau_{ij}}(x_j)}{e^{\lambda^\tau_{ij}}(x_j)}$$

$$Z^\tau = \frac{\nabla_{y(x_j)} v \; e^{\sum_{k \in Nb_l(j)} \lambda^\tau_{kj}(x_j) + 1}}{\sum_{x_i} \mu_n(x_i, x_j) \exp\{\theta(x_i, x_j) + \lambda^\tau_{ij}(x_j) + \lambda^\tau_{ji}(x_i) - \lambda^\tau_{ij}\}}$$

$$Z^\tau = \frac{\nabla_{y(x_j)} v \;\; \exp\{\frac{\nabla_{y(x_j)}}{\mu^\tau(x_j)}\}}{\sum_{x_i} \mu^\tau(x_i, x_j)}$$

So now we have new $\lambda^{\tau+1}_{ij}(x_j)$ as:

$$\lambda^{\tau+1}_{ij}(x_j) = W\big[Z^\tau\big] - 1 - \sum_{k \in Nb_l(j) \setminus i} \lambda^\tau_{kj}(x_j)$$

Using the above equation to calculate the new $\mu^{\tau+1}(x_j)$, we have:

$$\mu^{\tau+1}(x_j) = \frac{\nabla_{y(x_j)} v}{1 + \sum_{k \in Nb_l(j) \setminus i} \lambda^\tau_{kj}(x_j) + \lambda^{\tau+1}_{ij}(x_j)}$$

$$\mu^{\tau+\mathbf{1}}(\mathbf{x_j}) = \frac{\nabla_{\mathbf{y(x_j)}} \mathbf{v}}{\mathbf{W}[\mathbf{Z}^\tau]}$$

The only other quantity affected by $\lambda^{\tau+1}_{ij}(x_j)$ is $\mu^{\tau+1}(x_i, x_j)$. We have:

$$\mu^{\tau+1}(x_i, x_j) = \mu_n(x_i, x_j) e^{\left\{\theta(x_i, x_j) + \lambda^\tau_{ji}(x_i) - \lambda^\tau_{ij}\right\}} e^{\lambda^{\tau+1}_{ij}(x_j)}$$

$$= \mu_n(x_i, x_j) e^{\left\{\theta(x_i, x_j) + \lambda^\tau_{ji}(x_i) - \lambda^\tau_{ij}\right\}} \frac{e^{W[Z]}}{e^{1 + \sum_{k \in Nb_l(j) \setminus i} \lambda^\tau_{kj}(x_j)}}$$

Multiplying and dividing the above expression by $e^{\lambda^\tau_{ij}(x_j)}$, we get

$$= \mu_n(x_i, x_j) e^{\left\{\theta(x_i, x_j) + \lambda^\tau_{ij}(x_j) + \lambda^\tau_{ji}(x_i) - \lambda^\tau_{ij}\right\}} \frac{e^{W[Z]}}{e^{1 + \sum_{k \in Nb_l(j)} \lambda^\tau_{kj}(x_j)}}$$

$$\mu^{\tau+\mathbf{1}}(\mathbf{x_i}, \mathbf{x_j}) = \mu^\tau(\mathbf{x_i}, \mathbf{x_j}) \exp\left(\mathbf{W}[\mathbf{Z}^\tau] - \frac{\nabla_{\mathbf{y(x_j)}} \mathbf{v}}{\mu^\tau(\mathbf{x_j})}\right)$$

A.5 Proposition 7

Substituting the definition of Bregman function in Eq. (18) we get the proximal iteration as:

$$\mu^{n+1} = \arg\min_{\mu \in \Omega} \left\{ g(\mu) + \frac{1}{\omega}\Big(f(\mu) - f(\mu^n) - \nabla f(\mu^n)\mu + \nabla f(\mu^n)\mu^n \Big) \right\}$$

$$= \arg\min_{\mu \in \Omega} \left\{ g(\mu) + \frac{1}{\omega}\big(f(\mu) - \nabla f(\mu^n) \cdot \mu \big) \right\}$$

Consider the D.C. program:

$$\min_{\mu \in \Omega} \{ u(\mu) - v(\mu) \}$$

equivalent to the original problem $\min_{\mu \in \Omega} g(\mu)$ with $u(\mu) = g(\mu) + \frac{1}{\omega} f(\mu)$ and $v(\mu) = \frac{1}{\omega} f(\mu)$. The CCCP iteration of Eq. (7) is given as:

$$\arg\min_{\mu \in \Omega} \{ g(\mu) + \frac{1}{\omega} f(\mu) - \frac{1}{\omega} \nabla f(\mu^n) \cdot \mu \}$$

which is equivalent to the previous proximal scheme iteration.

B Experimental Results for Max-Product for Biq Instances

Table 1 shows the complete set of results, detailing the solution quality achieved by max-product.

Table 1: Solution quality comparisons for max-product

Instance	Optimal	MP
100-1	7970	7822
100-2	11036	11036
100-3	12723	12723
100-4	10368	10368
100-5	9083	9083
100-6	10210	10065
100-7	10125	10034
100-8	11435	11435
100-9	11455	11455
100-10	12565	12565
250-1	45607	45607
250-2	44810	44810
250-3	49037	49037
250-4	41274	41270
250-5	47961	47961
250-6	41014	41014
250-7	46757	46757
250-8	35726	34450
250-9	48916	48916
250-10	40442	40442
1b.20	133	0
2b.30	121	0
3b.40	118	0
4b.50	129	0
5b.60	150	0
6b.70	146	61
7b.80	160	0
8b.90	145	0
9b.100	137	0
10b.125	154	0

References

[1] Bharath Sriperumbudur and Gert Lanckriet. On the convergence of the concave-convex procedure. In *Advances in Neural Information Processing Systems*, pages 1759–1767, 2009.

[2] W. Zangwill. *Nonlinear Programming: A Unified Approach.* Prentice Hall, Englewood Cliffs, N.J., 1969.