8. Supplementary Material

Proof of Proposition 1

Proposition. *$\sum_{i\in\mathbb{F}}\langle\theta_i,\mu_i\rangle = \sum_{i\in\mathbb{F}}\langle\theta_i^\phi,\mu_i\rangle$, whenever $\mu_1,\dots,\mu_k$ obey the coupling constraints.*

Proof. $\sum_{i\in\mathbb{F}}\langle\theta^\phi,\mu_i\rangle = \sum_{i\in\mathbb{F}}\langle\theta,\mu_i\rangle + \underbrace{\sum_{ij\in\mathbb{E}}\langle\phi_{(i,j)},A_{(i,j)}\mu_i\rangle + \langle\phi_{(j,i)},A_{(j,i)}\mu_j\rangle}_{(*)} = \sum_{i\in\mathbb{F}}\langle\theta,\mu_i\rangle$, where $(*)=0$ due to $\phi_{(i,j)} = -\phi_{j,i)}$ and $A_{(i,j)}\mu_i = A_{(j,i)}\mu_j$. □

Proof of Proposition 2

Proposition. *Let $\text{conv}(X_i) = \{\mu_i : A_i\mu_i \le b_i\}$ with $A_i \in \mathbb{R}^{n\times m}$. Let the messages in problem (15) have size $n_1,\dots,n_{|J|}$. Then (15) is a linear program with $O(n+n_1+\ldots+n_{|J|})$ variables and $O(m+n_1+\ldots+n_{|J|})$ constraints.*

Proof. From LP-duality we know that $\mu_i^* \in \text{argmin}_{\mu_i:A\mu_i\le b_i}\langle c,\mu_i\rangle$ iff $\exists y \ge 0 : A_i^\top y = c_i$ and $\langle b_i - A_i\mu_i^*, y\rangle = 0$. Hence, (15) can be rewritten as

$$\begin{array}{ll}\max\limits_{y\ge0,\Delta_{(i,j_1)},\dots,\Delta_{(i,j_l)}} & \langle\delta,\theta^{\phi+\Delta}\rangle \\ \text{s.t.} & \langle b_i - A_i\mu_i^*, y\rangle = 0 \\ & A_i^\top y = \theta^{\phi+\Delta} \\ & \Delta_{(i,j)}(s)\begin{cases}\le 0, & \nu_i(s)=0\\ \ge 0, & \nu_i(s)=1\end{cases} \\ & \quad\text{where } \nu_i := A_{(i,j)}\mu_i^*\end{array} \tag{19}$$

$\theta^{\phi+\Delta}$ is a linear expression and μ^* is constant during the computation, hence (19) is a LP. □

Proof of Lemma 1 and Lemma 2

Lemma. *Let $ij \in \mathbb{E}$ be a pair of factors related by the coupling constraints and $\phi_{(i,j)}$ be a corresponding dual vector. Let $x_i^* \in \text{argmin}_{x_i\in X_i}\langle\theta_i^\phi,x_i\rangle$ and $\Delta_{(i,j)}$ satisfy*

$$\Delta_{(i,j)}(s)\begin{cases}\ge 0, & \nu(s)=1\\ \le 0, & \nu(s)=0\end{cases},\ \text{where } \nu := A_{(i,j)}x_i^*\,. \tag{20}$$

Then $x_i^* \in \text{argmin}_{x_i\in X_i}\langle\theta_i^{\phi+\Delta},x_i\rangle$ implies $D(\phi) \le D(\phi+\Delta)$.

Proof. Let $x_j^* \in \text{argmin}_{x_j\in X_j}\langle\theta_j^\phi,x_j\rangle$ be a solution of (13) at which the dual lower bound (11) is attained before the update and $x_j^{**} \in \text{argmin}_{x_j\in X_j}\langle\theta^\phi - A_{(j,i)}^\top\Delta_{(i,j)}^*,x_j\rangle$ be an integral solution at which the dual lower bound is attained after ϕ has been updated. Variable x_i^* as chosen in (13) is optimal for θ^ϕ and for $\theta^{\phi+\Delta}$ by construction. We need to prove

$$\begin{aligned}&\langle\theta_i^\phi,x_i^*\rangle + \sum_{j\in J}\langle\theta_j^\phi,x_j^*\rangle \\ &\le \langle\theta_i^\phi + \sum_{j\in J}A_{(i,j)}^\top\Delta_{(i,j)}^*, x_i^*\rangle + \sum_{j\in J}\langle\theta_j^\phi - A_{(j,i)}^\top\Delta_{(i,j)}^*, x_j^{**}\rangle\,.\end{aligned} \tag{21}$$

We shuffle all terms with variables $\Delta_{(i,j)}^*$, $j\in J$ to the right side and all other terms to the left side.

$$\begin{aligned}&\langle\theta_i^\phi,x_i^*-x_i^*\rangle + \sum_{j\in J}\langle\theta_j^\phi,x_j^*-x_j^{**}\rangle \\ &\le \langle\sum_{j\in}A_{(i,j)}^\top\Delta_{(i,j)}^*,x_i^*\rangle - \sum_{j\in J}\langle A_{(j,i)}^\top\Delta_{(i,j)}^*,x_j^{**}\rangle\end{aligned} \tag{22}$$

All terms on the left side are smaller than zero due to the choice of x_j^* being minimizers w.r.t. θ_j^ϕ. Hence, it will be enough to prove the above inequality when assuming the left side to be zero. We rewrite the scalar products by transposing $A_{(i,j)}^\top$ and $A_{(j,i)}^\top$.

$$0 \le \sum\nolimits_{j\in J}\left\{\langle\Delta_{(i,j)}^*, A_{(i,j)}x_i^* - A_{(j,i)}x_j^{**}\rangle\right\} \tag{23}$$

Due to $A_{(j,i)}x_j^{**} \in \{0,1\}^{\dim(\phi_{(i,j)})}$ and $A_{(i,j)}x_i^* \in \{0,1\}^{\dim(\phi_{(i,j)})}$ by Definition 1 and $\Delta_{(i,j)}^* \lesseqgtr 0$ whenever $A_{(i,j)}x_i^* \lesseqgtr 0$, the result follows. □

Lemma. *Let $\Delta \in AD(\theta_i^\phi,x_i^*,J)$ then $D(\phi) \le D(\phi+\Delta)$.*

Proof. Analoguous to the proof of Lemma 1. □

Proof of Theorem 1

Theorem. *Algorithm 2 monotonically increasis the dual lower bound (11).*

Proof. We prove that (i) the receiving messages and (ii) the sending messages step improve (11).
(i) Directly apply Lemma 1. (ii) The difficulty here is that we compute descent directions from the current dual variables ϕ in parallel and then apply all of them simultaneously. By Lemma 2, the send message step is non-decreasing when called for each set $J_1,\dots,J_l$ in Algorithm 2. The dual lower bound $L(\phi)$ is concave, hence we apply Jensen's inequality and note that $\omega_1+\ldots+\omega_l \le 1$ to obtain the result. □

Proof of Theorem 2

Theorem. *If θ^ϕ is marginally consistent, the dual lower bound $D(\phi)$ cannot be improved by Algorithm 2.*

First, we need two technical lemmata.

Lemma 3. *Let* $X \subset \{0,1\}^n$, $A \in \{0,1\}^{K\times n}$ *and* $Ax \in \{0,1\}^K$ $\forall x \in X$. *Let* $x^* \in X$ *be given and define* $\nu^* := Ax^*$. *Let* $\Delta \in \mathbb{R}^K$ *be given such that* $\Delta(s) \begin{cases} \geq 0, & \nu^*(s)=1 \\ \leq 0, & \nu^*(s)=0 \end{cases}$. *Then (i)* $x^* \in \operatorname{argmin}_{x\in X}\langle -\Delta, Ax\rangle$ *and (ii) for* $x^{**} \in \operatorname{argmin}_{x\in X}\langle -\Delta, Ax\rangle$, $\nu^{**} = Ax^{**}$ *it holds that* $\Delta(s) = 0$ *whenever* $\nu^*(s) \neq \nu^{**}(s)$.

Proof. Let $x \in X$ and define $\nu = Ax$. Then

$$\begin{aligned}\langle -\Delta, Ax\rangle \\ &= \underbrace{\sum_{s:\nu^*(s)=1=\nu(s)} -\Delta(s)}_{(*)} + \underbrace{\sum_{s:\nu(s)=1>0=\nu^*(s)} -\Delta(s)}_{(**)} \\ &\geq \underbrace{\sum_{s:\nu^*(s)=1} -\Delta(s)}_{(***)} \\ &= \langle -\Delta, Ax^*\rangle \end{aligned} \tag{24}$$

because $(*) \geq (***)$ due to $\Delta(s) \geq 0$ for $\nu^*(s)=1$ and $(**) \geq 0$ due to $\Delta(s) \leq 0$ for $\nu^*(s)=0$. This proves (i) and (ii) is proven by observing that $(**) = 0$ and $(*) = (***)$ must also hold. □

Lemma 4. *Let* $x_i^*, x_i^{**} \in \operatorname{argmin}_{x_i\in X_i}\langle\theta^\phi, x_i\rangle$ *be two solutions to the* i*-th factor for the current reparametrization* θ^ϕ. *If* Δ *is admissible w.r.t.* x_i^* *then* Δ *is also admissible w.r.t.* x_i^{**}.

Proof. As both x_i^* and x_i^{**} are optimal to θ^ϕ and x_i^* is also optimal to $\theta^{\phi+\Delta}$, we have $\langle\Delta_{(i,j)}, A_{(j,i)}x_i^*\rangle \leq \langle\Delta_{(i,j)}, A_{(j,i)}x_i^{**}\rangle$. By Lemma 3, (i) also $\langle -\Delta_{(i,j)}, A_{(j,i)}x_i^*\rangle \leq \langle -\Delta_{(i,j)}, A_{(j,i)}x_i^{**}\rangle$ holds, hence equality must hold. This shows $x_i^{**} \in \operatorname{argmin}_{x_i\in X_i}\langle\theta^{\phi+\Delta}, x_i\rangle$. Second, Lemma 3, (ii) implies that $\Delta(s) = 0$ whenever $\nu^*(s) \neq \nu^{**}(s)$. This proves that $\Delta_{(i,j)}(s) \begin{cases} \geq 0, & \nu^{**}(s)=1 \\ \leq 0, & \nu^{**}(s)=0 \end{cases}$, $\nu^{**} := A_{(i,j)}x_i^{**}$. □

Proof of Theorem 2. It is sufficient to show that for marginally consistent θ^ϕ for $\mathbb{S}$, the update Δ computed by Algorithm 1 on an arbitrary factor $i \in \mathbb{F}$ and some set $J \subset \mathcal{N}_{\mathbb{G}}(i)$ has the following properties: (i) $L(\phi) = L(\phi+\Delta)$, (ii) $\theta^{\phi+\Delta}$ is marginally consistent for $\mathbb{S}$. For an easier proof, we only consider the case $J = \{j\}$. The general case can be proven analoguously.

(i) Let $x_i^* \in \mathbb{S}_i$, $x_j^* \in \mathbb{S}_j$ with $A_{(i,j)}x_i^* = A_{(j,i)}x_j^*$. We have to show that

$$\min_{x_i\in X_i}\langle\theta_i^\phi, x_i\rangle + \min_{x_j\in X_j}\langle\theta_j^\phi, x_j\rangle = \min_{x_i\in X_i}\langle\theta_i^{\phi+\Delta}, x_i\rangle + \min_{x_j\in X_j}\langle\theta_j^{\phi+\Delta}, x_j\rangle \tag{25}$$

Due to x_i^* optimal to $\theta_i^{\phi+\Delta}$, since by Lemma 4 the update Δ is admissible for x_i^*, it remains to show that $x_j^* \in \operatorname{argmin}_{x_j\in X_j}\langle\theta^{\phi+\Delta}, x_j\rangle$. As $x_j^* \in \operatorname{argmin}_{x_j\in X_j}\langle\theta^{\phi}, x_j\rangle$, it is sufficient to prove that $x_j^* \in \operatorname{argmin}_{x_j\in X_j}\langle -\Delta_{(i,j)}, A_{(j,i)}x_j\rangle$. This follows from Lemma 3 (i). We conclude by noting $\langle\theta_i^\phi, x_i^*\rangle + \langle\theta_j^\phi x_j\rangle = \langle\theta_i^{\phi+\Delta}, x_i^*\rangle + \langle\theta_j^{\phi+\Delta} x_j\rangle$.

(ii) The computations in (i) show that $\mathbb{S}_i \subseteq \operatorname{argmin}_{x_i\in X_i}\langle\theta_i^{\phi+\Delta}, x_i\rangle$ and $\mathbb{S}_j \subseteq \operatorname{argmin}_{x_j\in X_j}\langle\theta_j^{\phi+\Delta}, x_j\rangle$. The reparametrizations of all other factors stay the same: $\theta_k^{\phi+\Delta} = \theta_k^\phi$ for $k \in \mathbb{F}\backslash\{i,j\}$. Hence, $\theta^{\phi+\Delta}$ is marginally consistent for $\mathbb{S}$ after the update. □

9. Special Cases: Graphical Model Solvers

We will show how Algorithm 2 subsumes known message-passing algorithms MSD [74], TRWS [48], SRMP [49] and MPLP [28] for MAP-inference with common graphical models, considered in Example 1.

Solver Primitives (13) and (15). As it can be seen, all factors in (5) are of the form $X_i = \{(1,0,\ldots,0),(0,1,0,\ldots,0),\ldots,(0,\ldots,0,1)\}$ and $\operatorname{conv}(X_i) = \{\mu \geq 0 : \langle\mathbb{1},\mu\rangle = 1\}$ is a $\dim_(X_i)$-dimensional simplex.

In all message passing algorithms [48, 49, 74, 28], there are two types of invokations of Algorithm 1 together with solutions of the accompanying optimization problem (13) and (15):

Alg. 1 input	Factor Optimization (13)	Reparametrization adjustment (15)
$i = u \in \mathsf{V}$ $J = \{uv\}$ $uv \in \mathsf{E}$	$\min_{x_u\in X_u}\{\theta_u^\phi(x_u)\}$	$\Delta^*_{(u,uv)}(x_u) = \min_{x'_u\in X_u}\theta_u^\phi(x'_u) - \theta_u^\phi(x_u)$
$i = uv \in \mathsf{E}$ $J = \{u\}$ $u \in \mathsf{V}$	$\min\{\theta_u^\phi(x_u,x_v)\}$ $(x_u,x_v) \in X_u \times X_v$	$\Delta^*_{(uv,u)}(x_u) = \min_{x'_{uv}\in X_{uv}}\theta_{uv}^\phi(x'_{uv}) - \min_{x_v\in X_v}\{\theta_{uv}(x_u,x_v)\}$

MAP-inference Solvers. In Table 1 we state solvers MSD [74], TRWS [48], SRMP [49] and MPLP [28] as special cases of our framework. Factors are visited in the order they are read in.

Algorithm	Current factor	$J_{receive}$	$J_1 \dot\cup \dots \dot\cup J_l$	ω
MSD [74]	$u \in \mathsf{V}$	$\mathcal{N}_{\mathbb{G}}(u)$	$\{uv\} \subset \mathcal{N}_{\mathbb{G}}(u)$	$\omega_1, \ldots = 1/\lvert\mathcal{N}_{\mathbb{G}}(u)\rvert$
	$uv \in \mathsf{E}$	$\varnothing$	—	—
MPLP [28]	$u \in \mathsf{V}$	$\varnothing$	—	—
	$uv \in \mathsf{E}$	$\{u, v\}$	$\{u\}, \{v\}$	$\omega_1 = 1/2 = \omega_2$
TRWS [48] SRMP [49]			forward pass:	
	$u \in \mathsf{V}$	$\{uv : v \in \mathcal{N}_{\mathsf{G}}(u), v < u\}$	$\{uv\} : v \in \mathcal{N}_{\mathsf{G}}(u), v > u$	$\omega_1, \ldots = 1/\max(\{v \in \mathcal{N}_{\mathsf{G}}(u) : v > u\}, \{v \in \mathcal{N}_{\mathsf{G}}(u) : v < u\})$
			backward pass:	
	$u \in \mathsf{V}$	$\{uv : v \in \mathcal{N}_{\mathsf{G}}(u), v > u\}$	$\{uv\} : v \in \mathcal{N}_{\mathsf{G}}(u), v < u$	$\omega_1, \ldots = 1/\max(\{v \in \mathcal{N}_{\mathsf{G}}(u) : v > u\}, \{v \in \mathcal{N}_{\mathsf{G}}(u) : v < u\})$
	$uv \in \mathsf{E}$	$\varnothing$	—	—

Table 1. [74, 48, 49, 28] as special cases of Algorithm 2.

Remark 1. *We have only treated the case of unary $\theta_u, u \in \mathsf{V}$ and pairwise potentials $\theta_{uv}, uv \in \mathsf{E}$ here. MPLP [28] and SRMP [49] can be applied to higher order potentials as well, which we do not treat here.SRMP [49] is a generalisation of TRWS [48] to the higher-order case.*

Remark 2. *There are convergent message-passing algorithms such that factors comprise trees [72, 65]. Their analysis is more difficult, hence we omit it here.*

Note that our framework generalizes upon [48, 49, 28, 74, 65, 72] in several ways: (i) Our factors need not be simplices or trees. (ii) Our messages need not be marginalization between unary/pairwise/triplet/. . . factors. (iii) We can compute message updates on more than one coupling constraint simultaneously, i.e. we may choose $J_1 \dot\cup \dots \dot\cup J_l$ in Algorithm 2 to be different than singleton sets. (i) and (ii) affect LP-modeling, (iii) affects computational efficiency: By considering multiple messages at once in Procedure 1, we may be able to make larger updates Δ^*, resulting in faster convergence.