A Bregman Projection Derivation

The objective (Reg) can be equivalently interpreted as a Bregman projection. This interpretation has been explored by Ravikumar et al. (2010) as a basis for proximal updates and also Benamou et al. (2015) for the optimal transport problem. Here, we review the transformation because it is central to the algorithm of Ravikumar et al. (2010), upon which our main theoretical results are based.

By definition of the Bregman projection with respect to the negative entropy, $\Phi = -H$, we have

$$\begin{aligned}\mathcal{D}_\Phi(\boldsymbol{\mu}, \mathbb{1}) &=_+ \langle \boldsymbol{\mu}, \log \boldsymbol{\mu} - \mathbb{1}\rangle - \langle \log \mathbb{1}, \boldsymbol{\mu} - \mathbb{1}\rangle \\ &= -H(\boldsymbol{\mu})\end{aligned}$$

where $\mathbb{1}$ is a vector of ones of the same size as the marginal vector and $=_+$ denotes the two sides are equal up to a constant. Substituting this into (Reg) and multiplying through by η yields the objective:

$$\min \quad \eta\langle C, \boldsymbol{\mu}\rangle + \mathcal{D}_\Phi\left(\boldsymbol{\mu}, \mathbb{1}\right) \qquad \text{s.t.} \quad \boldsymbol{\mu} \in \mathbb{L}_m.$$

Note the similarity to a projected mirror descent update over $\mathbb{L}_m$ starting from $\mathbb{1}$ (Bubeck, 2015; Nemirovsky and Yudin, 1983). Using this insight and performing a single gradient update in the dual, we can transform the problem into a single Bregman projection of the vector. The unprojected marginal vector $\boldsymbol{\mu}'$ satisfies

$$\nabla\Phi(\boldsymbol{\mu}') = \nabla\Phi(\mathbb{1}) - \eta C,$$

where $\nabla\Phi(\boldsymbol{\mu}) = -\nabla H(\boldsymbol{\mu}) = \log \boldsymbol{\mu}$ is the dual map and $(\nabla\Phi)^{-1}(\boldsymbol{\mu}) = \nabla\Phi^*(\boldsymbol{\mu}) = \exp(\boldsymbol{\mu})$ is the inverse dual map. We have $\boldsymbol{\mu}' = \exp(-\eta C)$ and the solution to the mirror descent update is $\mathcal{P}_{\mathbb{L}_m}(\exp(-\eta C))$. Therefore it is sufficient to solve the following Bregman projection problem:

$$\min \quad \mathcal{D}_\Phi\left(\boldsymbol{\mu}, \exp(-\eta C)\right) \qquad \text{s.t.} \qquad \boldsymbol{\mu} \in \mathbb{L}_m$$

The projection, however, cannot be computed in closed form due to the complex geometry of $\mathbb{L}_m$. Sinkhorn-like algorithms such as those used in Cuturi (2013) are unavailable because the transportation polytopes $\mathcal{U}_d(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$ are dependent on variables $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ which are also involved in the projection operation.

B Derivation of EMP Update Rules

We present the derivations of the update rules similar to Ravikumar et al. (2010) for a given edge $ij \in \mathcal{E}$ based on the Bregman projections onto the individual constraint sets $\mathcal{X}_{ij\to i}$, $\mathcal{X}_{ij,i}$, $\mathcal{X}_{ij\to j}$, $\mathcal{X}_{ij,j}$. We refer the reader to Ravikumar et al. (2010) for the original algorithm and derivation. We derive only the first two projections; the last two can be found by exchanging the indices.

(a) For the projection $\boldsymbol{\mu}' = \mathcal{P}_{\mathcal{X}_{ij\to i}}(\boldsymbol{\mu})$, where

$$\mathcal{X}_{ij\to i} = \{\boldsymbol{\mu} \ : \ \boldsymbol{\mu}_{ij}\mathbb{1} = \boldsymbol{\mu}_i\},$$

there are no constraints on any edges or vertices other than ij and i. Therefore, $\forall k \neq i$, $\boldsymbol{\mu}'_k = \boldsymbol{\mu}_k$. Similarly, $\forall k\ell \neq ij$, $\boldsymbol{\mu}'_{k\ell} = \boldsymbol{\mu}_{k\ell}$.

The Lagrangian of the projection is given in terms of primal variables $\boldsymbol{\mu}$ and dual variables α:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\mu}', \alpha) &= \sum_{x_i, x_j} \boldsymbol{\mu}'_{ij}(x_i, x_j)\left(\log \frac{\boldsymbol{\mu}'_{ij}(x_i, x_j)}{\boldsymbol{\mu}_{ij}(x_i, x_j)} - 1\right) + \sum_{x_i} \boldsymbol{\mu}'_i(x_i)\left(\log \frac{\boldsymbol{\mu}'_i(x_i)}{\boldsymbol{\mu}_i(x_i)} - 1\right) + \alpha^\top\left(\boldsymbol{\mu}'_{ij}\mathbb{1} - \boldsymbol{\mu}'_i\right) \\ &= \sum_{x_i, x_j} \boldsymbol{\mu}'_{ij}(x_i, x_j)\left(\log \frac{\boldsymbol{\mu}'_{ij}(x_i, x_j)}{\boldsymbol{\mu}_{ij}(x_i, x_j)} - 1 + \alpha(x_i)\right) + \sum_{x_i} \boldsymbol{\mu}'_i(x_i)\left(\log \frac{\boldsymbol{\mu}'_i(x_i)}{\boldsymbol{\mu}_i(x_i)} - 1 - \alpha(x_i)\right).\end{aligned}$$

By the first-order optimality condition, the primal solution in terms of the dual variables is

$$\begin{aligned}\boldsymbol{\mu}'_{ij}(x_i, x_j) &= \boldsymbol{\mu}_{ij}(x_i, x_j)e^{-\alpha(x_i)} \\ \boldsymbol{\mu}'_i(x_i) &= \boldsymbol{\mu}_i(x_i)e^{\alpha(x_i)}.\end{aligned}$$

Substituting this solution back in to the Lagrangian, we have

$$\mathcal{L}(\alpha) = -\sum_{x_i,x_j} \boldsymbol{\mu}_{ij}(x_i,x_j)e^{-\alpha(x_i)} - \sum_{x_i} \boldsymbol{\mu}_i(x_i)e^{\alpha(x_i)}.$$

Again, by the first-order optimality condition, the dual solution is

$$\alpha^*(x_i) = \frac{1}{2}\log\frac{\sum_{x_j}\boldsymbol{\mu}_{ij}(x_i,x_j)}{\boldsymbol{\mu}_i(x_i)}.$$

Substituting this value for α^* into the primal solution yields the desired result.

(b) Again, for the projection onto

$$\mathcal{X}_{ij,i} = \{\boldsymbol{\mu} \ : \ \boldsymbol{\mu}_i^\top \mathbb{1} = 1, \ \mathbb{1}^\top \boldsymbol{\mu}_{ij}\mathbb{1} = 1\},$$

only $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_{ij}$ are affected. $\mathcal{X}_{ij,i}$ enforces that the variables $\boldsymbol{\mu}_{ij}$ and $\boldsymbol{\mu}_i$ each sum to one. It is well known and easy to show that the Bregman projection with respect to the negative entropy is simply the $\boldsymbol{\mu}_{ij}$ and $\boldsymbol{\mu}_i$ normalized by their sums. This normalization can also be written as a multiplicative update of the same form by observing that

$$\boldsymbol{\mu}'_{ij}(x_i,x_j) = \boldsymbol{\mu}'_{ij}(x_i,x_j)e^{-\xi^*_{ij}}$$
$$\boldsymbol{\mu}'_i(x_i) = \boldsymbol{\mu}'_i(x_i)e^{-\xi^*_i},$$

where $\xi^*_{ij} = \log\sum_{x_i,x_j}\boldsymbol{\mu}_{ij}(x_i,x_j)$ and $\xi^*_i = \log\sum_{x_i}\boldsymbol{\mu}_i(x_i)$. Again, these can be derived via the Lagrangian.

C Extensions of EMP

C.1 Dual EMP

We may also equivalently interpret the multiplicative updates in Algorithm 1 and Algorithm 2 as additive updates of the dual variables. The dual interpretation is consistent with past work in dual MAP algorithms (Sontag et al., 2011) and may be more practical to avoid numerical issues in implementation. Instead of tracking the primal variables $\boldsymbol{\mu}$, we track a sum of the dual variables with ζ for each vertex and edge. Enforcing consistency between a given joint distribution and its marginals in (a) yields updated dual variable sums

$$\zeta'_{ij}(x_i,x_j) \leftarrow \zeta_{ij}(x_i,x_j) - \alpha^*(x_i) \qquad \zeta'_i(x_i) \leftarrow \zeta_i(x_i) + \alpha^*(x_i),$$

where again $\alpha^*(x_i) = \frac{1}{2}\log\frac{\sum_{x_j}\boldsymbol{\mu}_{ij}(x_i,x_j)}{\boldsymbol{\mu}_i(x_i)}$. The same is done for the vertex j in (c) with indices exchanged. The normalization step in (b) yields

$$\zeta'_{ij}(x_i,x_j) \leftarrow \zeta_{ij}(x_i,x_j) - \xi^*_{ij} \qquad \zeta'_i(x_i,x_j) \leftarrow \zeta_i(x_i) - \xi^*_i,$$

where $\xi^*_{ij} = \log\sum_{x_i,x_j}\boldsymbol{\mu}_{ij}(x_i,x_j)$ and $\xi^*_i = \log\sum_{x_i}\boldsymbol{\mu}_i(x_i)$. Again, the same is done for (d). The primal marginal vector is recovered with

$$\boldsymbol{\mu} = \exp(-\eta C + \zeta).$$

We will later make explicit the dual formulation as it will aid in the theoretical analysis.

C.2 Clique Constraints

The version of EMP presented in the paper is for the $\mathbb{L}_2$ local polytope, which enforces only pairwise consistency among the variables with edges, but this can be fairly easily extended. In this section, we discuss higher order pseudo-marginals and their constraints. Consider the polytope that enforces consistency on all subsets of $\mathcal{V}$ of size k and below, denoted by $\mathcal{C}$. We use the notation of Meshi et al. (2012). The constraint set is written as

$$\mathbb{L}_{\mathcal{C}} \stackrel{\text{def.}}{=} \left\{\boldsymbol{\mu} \geq 0 \ : \ \begin{array}{ll} \boldsymbol{\mu}_i \in \Sigma_m & \forall i \in \mathcal{V} \\ \boldsymbol{\mu}_i(x_i) = \sum_{x_{c\setminus i}}\boldsymbol{\mu}_c & \forall x_i \in \chi, i \in c, c \in \mathcal{C}, \end{array}\right\}. \tag{4}$$

where $x_{c\setminus i}$ denotes a marginalization over all variables except i. For convenience, we may also now account for higher-order interactions in the model itself:

$$\max \quad \sum_{c\in\mathcal{C}} \sum_{x_c\in\chi^k} \theta_c(x_c)\boldsymbol{\mu}_c(x_c) + \sum_{i\in\mathcal{V}} \sum_{x_i\in\chi} \theta_i(x_i)\boldsymbol{\mu}_i(x_i) \quad \text{s.t.} \quad \boldsymbol{\mu} \in \mathbb{L}_{\mathcal{C}}$$

The projection operation in (Proj) is the same for $C = -\theta$. Analogous update rules to Proposition 1 can derived with exactly the same procedure. For a given subset $c \in \mathcal{C}$ and vertex i, we have that $\boldsymbol{\mu}' = \mathcal{P}_{ij\to i}(\boldsymbol{\mu})$ constitutes the update

$$\boldsymbol{\mu}'_c(x_c) = \boldsymbol{\mu}_c(x_c)\sqrt{\frac{\boldsymbol{\mu}_i(x_i)}{\sum_{x_{c\setminus i}} \boldsymbol{\mu}_c(x_c)}}$$
$$\boldsymbol{\mu}'_i(x_i) = \boldsymbol{\mu}_c(x_c)\sqrt{\frac{\sum_{x_{c\setminus i}} \boldsymbol{\mu}_c(x_c)}{\boldsymbol{\mu}_i(x_i)}}.$$

The normalization updates are identical as well. As in the presented EMP algorithm, we can design greedy and cyclic algorithms around these update equations. The theoretical analysis in Section 6 will focus on the case with edges only. We leave the general analysis of $\mathbb{L}_{\mathcal{C}}$ for future work.

D Omitted Proofs and Derivations from Section 6

D.1 Derivation of the Lyapunov function (3)

For convenience, L is restated here:

$$\begin{aligned} L(\lambda,\xi) = &- \sum_{ij\in\mathcal{E}} \sum_{x_i,x_j\in\chi} \exp\left(-\eta C_{ij}(x_i,x_j) - \lambda_{ij}(x_i) - \lambda_{ji}(x_j) - \xi_{ij}\right) \\ &- \sum_{i\in\mathcal{V}} \sum_{x\in\chi} \exp\left(-\eta C_i(x) - \xi_i + \sum_{j\in N_r(i)} \lambda_{ij}(x) + \sum_{j\in N_c(i)} \lambda_{ji}(x)\right) \\ &- \sum_{ij\in\mathcal{E}} \xi_{ij} - \sum_{i\in\mathcal{V}} \xi_i + \sum_{ij\in\mathcal{E}} \sum_{x_i,x_j\in\chi} \exp(-\eta C_{ij}(x_i,x_j)) + \sum_i \sum_{x\in\chi} \exp(-\eta C_i(x)). \end{aligned} \tag{5}$$

The Lagrangian of (Proj) with primal variables $\boldsymbol{\mu}$ and dual variables (λ,ξ) can be written as

$$\begin{aligned} \mathcal{L}(\boldsymbol{\mu},\lambda,\xi) = \mathcal{D}_\Phi(\boldsymbol{\mu},\exp(-\eta C)) &+ \sum_{ij} \left(\lambda_{ij}^\top(\boldsymbol{\mu}_{ij}\mathbb{1} - \boldsymbol{\mu}_i) + \lambda_{ji}^\top(\boldsymbol{\mu}_{ij}^\top\mathbb{1} - \boldsymbol{\mu}_i)\right) \\ &+ \sum_{ij} \xi_{ij}(\mathbb{1}^\top\boldsymbol{\mu}_{ij}\mathbb{1} - 1) + \sum_i \xi_i(\boldsymbol{\mu}_i^\top\mathbb{1} - 1), \end{aligned}$$

where

$$\begin{aligned} \mathcal{D}_\Phi(\boldsymbol{\mu},\exp(-\eta C)) = &\sum_{ij} \sum_{x_i,x_j} \boldsymbol{\mu}_{ij}(x_i,x_j)\left(\log\boldsymbol{\mu}_{ij}(x_i,x_j) + \eta C_{ij}(x_i,x_j) - 1\right) \\ &+ \sum_i \sum_x \boldsymbol{\mu}_i(x)\left(\log\boldsymbol{\mu}_i(x) + \eta C_i(x) - 1\right) \\ &+ \sum_{ij} \sum_{x_i,x_i} \exp(-\eta C_{ij}(x_i,x_j)) + \sum_i \sum_x \exp(-\eta C_i(x)). \end{aligned}$$

The partial derivatives with respect to $\boldsymbol{\mu}_{ij}(x_i,x_j)$ and $\boldsymbol{\mu}_i(x)$ are given by

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\mu}_{ij}(x_i,x_j)} = \log\boldsymbol{\mu}_{ij}(x_i,x_j) + \eta C_{ij}(x_i,x_j) + \lambda_{ij}(x_i) + \lambda_{ji}(x_j) + \xi_{ij}$$
$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\mu}_i(x)} = \log\boldsymbol{\mu}_i(x) + \eta C_i(x) + \xi_i - \sum_{j\in N_r(i)} \lambda_{ij}(x_i) + \sum_{j\in N_c(i)} \lambda_{ji}(x_j).$$

Setting the derivatives to zero gives the solution $\boldsymbol{\mu}$ in terms of the dual variables:

$$\boldsymbol{\mu}_{ij}(x_i, x_j) = \exp\left(-\eta C_{ij}(x_i, x_j) - \lambda_{ij}(x_i) - \lambda_{ji}(x_j) - \xi_{ij}\right)$$
$$\boldsymbol{\mu}_i(x) = \exp\left(-\eta C_i(x) - \xi_i + \sum_{j \in N_r(i)} \lambda_{ij}(x) + \sum_{j \in N_c(i)} \lambda_{ji}(x)\right).$$

By substituting $\boldsymbol{\mu}$ in $\mathcal{L}$, we obtain the Lyapunov function L.

D.2 Proof of Lemma 1

In this section we prove Lemma 1. We restate the result for the reader's convenience.

Lemma 3. *For a given edge $ij \in \mathcal{E}$, let $\boldsymbol{\mu}'$ and (λ', ξ') denote the updated primal and dual variables after a projection from one of (a)–(d) in Proposition 1. We have the following improvements on L. If $\boldsymbol{\mu}'$ is equal to:*

(a) $\mathcal{P}_{\mathcal{X}_{ij\to i}}(\boldsymbol{\mu})$, then $L(\lambda', \xi') - L(\lambda, \xi) = 2h^2(\boldsymbol{\mu}_{ij}\mathbb{1}, \boldsymbol{\mu}_i)$

(b) $\mathcal{P}_{\mathcal{X}_{ij,i}}(\boldsymbol{\mu})$, then $L(\lambda', \xi') - L(\lambda, \xi) \geq 0$

(c) $\mathcal{P}_{\mathcal{X}_{ij\to j}}(\boldsymbol{\mu})$, then $L(\lambda', \xi') - L(\lambda, \xi) = 2h^2(\boldsymbol{\mu}_{ij}^\top\mathbb{1}, \boldsymbol{\mu}_j)$

(d) $\mathcal{P}_{\mathcal{X}_{ij,j}}(\boldsymbol{\mu})$, then $L(\lambda', \xi') - L(\lambda, \xi) \geq 0$.

Proof. Let L and L' denote the values of the Lyapunov function before and after the projection in each case.

(a) Due to the projection $\boldsymbol{\mu}' = \mathcal{P}_{\mathcal{X}_{ij\to i}}(\boldsymbol{\mu})$, only $\boldsymbol{\mu}_{ij}$ and $\boldsymbol{\mu}_i$ change values.

$$\begin{aligned}
L' - L &= \sum_{x_i, x_j} \left(\boldsymbol{\mu}_{ij}(x_i, x_j) - \boldsymbol{\mu}'_{ij}(x_i, x_j)\right) + \sum_x \left(\boldsymbol{\mu}_i(x) - \boldsymbol{\mu}'_i(x)\right) \\
&= \sum_{x_i, x_j} \boldsymbol{\mu}_{ij}(x_i, x_j) \left(1 - \sqrt{\frac{\boldsymbol{\mu}_i(x_i)}{\sum_{x'} \boldsymbol{\mu}_{ij}(x_i, x')}}\right) + \sum_x \boldsymbol{\mu}_i(x) \left(1 - \sqrt{\frac{\sum_{x'} \boldsymbol{\mu}_{ij}(x, x')}{\boldsymbol{\mu}_i(x_i)}}\right) \\
&= \left\|\sqrt{\boldsymbol{\mu}_{ij}\mathbb{1}} - \sqrt{\boldsymbol{\mu}_i}\right\|_2^2 = 2h^2(\boldsymbol{\mu}_{ij}\mathbb{1}, \boldsymbol{\mu}_i).
\end{aligned}$$

(b) Due to the projection $\boldsymbol{\mu}' = \mathcal{P}_{\mathcal{X}_{ij\to i}}(\boldsymbol{\mu})$ change, again only $\boldsymbol{\mu}_{ij}$ and $\boldsymbol{\mu}_i$, but they are simply normalized. From the derivation of the updates, we can see that only dual variables ξ_i and ξ_{ij} are updated in order for the normalization to occur. We have, from the update rule in Proposition 1

$$\xi'_{ij} = \xi_{ij} - \log \sum_{x_i, x_j} \boldsymbol{\mu}_{ij}(x_i, x_j)$$
$$\xi'_i = \xi_i - \log \sum_x \boldsymbol{\mu}_i(x).$$

The improvement on the Lyapunov function can then be written as

$$\begin{aligned}
L' - L &= \sum_{x_i, x_j} \boldsymbol{\mu}_{ij}(x_i, x_j) \left(1 - \exp(\xi'_{ij} - \xi_{ij})\right) + \sum_x \boldsymbol{\mu}_i(x) \left(1 - \exp(\xi'_i - \xi_i)\right) \\
&\quad + \xi'_{ij} - \xi_{ij} + \xi'_i - \xi_i \\
&= \sum_{x_i, x_j} \boldsymbol{\mu}_{ij}(x_i, x_j) - \log \sum_{x_i, x_j} \boldsymbol{\mu}_{ij}(x_i, x_j) - 1 \\
&\quad + \sum_x \boldsymbol{\mu}_i(x) - \log \sum_x \boldsymbol{\mu}_i(x) - 1,
\end{aligned}$$

where the second equality uses the fact that $\boldsymbol{\mu}'_{ij}$ and $\boldsymbol{\mu}_i$ both sum to one. This last expression can be shown to be non-negative by recognizing the classical inequality $x - \log x - 1 \geq 0$ for all $x > 0$.

(c) The proof of improvement is identical to (a); however, we replace vertex i with j and all row sums $\boldsymbol{\mu}_{ij}\mathbb{1}$ with column sum $\boldsymbol{\mu}_{ij}^\top\mathbb{1}$.

(d) The proof of improvement is identical to (b), but we replace i with j for the vertex marginal normalization.

□

D.3 Fixed points of EMP

We start this section by noting that all fixed points of EMP correspond to valid (constraint satisfying) primal solutions and therefore must equal global optima of the dual function.

First note that any fixed point of EMP corresponds to a candidate solution all whose constraints are satisfied. Indeed, at optimality λ^*, ξ^* satisfy:

$$\left(\boldsymbol{\mu}_\eta^*\right)_{ij}(x_i, x_j) = \exp\left(-\eta C_{ij}(x_i, x_j) - \lambda_{ij}^*(x_i) - \lambda_{ji}^*(x_j) - \xi_{ij}^*\right)$$

$$\left(\boldsymbol{\mu}_\eta^*\right)_i(x_i) = \exp\left(-\eta C_i(x_i) - \xi_i^* + \sum_{j\in N_r(i)} \lambda_{ij}^*(x_i) + \sum_{j\in N_c(i)} \lambda_{ji}^*(x_i)\right),$$

with $\boldsymbol{\mu}_\eta^* \in \mathbb{L}_2$. Since all constraints are satisfied, for all projection types $\mathcal{P}$ in Lemma 1, $\mathcal{P}(\boldsymbol{\mu}_\eta^*) = \boldsymbol{\mu}_\eta^*$.

For the converse, we proceed by contradiction. Let $\boldsymbol{\mu}$ be a fixed point of EMP. As such, all the normalization constraints (ensuring the edge and node distributions each sum to one) must be satisfied. Assume then that a constraint of type (a) or (c) is not satisfied. Without loss of generality let $ij \to i$ be the unsatisfied constraint. As a consequence of 1, the Lyapunov objective can be strictly increased by performing the corresponding Bregman projection, and therefore EMP couldn't have possibly be at a fixed point. We summarize these observations in the following proposition:

Proposition 2. *All maxima of $L(\lambda, \xi)$ are fixed points of EMP and all fixed points of EMP are maxima of $L(\lambda, \xi)$.*

D.4 Proof of Lemma 2

In this section we prove Lemma 2, we restate it here for readability:

Lemma 4. *Let λ^*, ξ^* denote the maximizers of L. The difference in function value between the optimal value of L and the value at the first iteration is upper bounded as*

$$L(\lambda^*, \xi^*) - L(\lambda^{(1)}, \xi^{(1)}) \leq \min(\|\eta C/d + \exp(-\eta C)\|_1, S).$$

Proof. We start by showing the upper bound:

$$L(\lambda^*, \xi^*) - L(\lambda^{(1)}, \xi^{(1)}) \leq \|\eta C/d + \exp(-\eta C)\|_1. \tag{6}$$

We have that $(\lambda, \xi) = (0, 0)$ when $\boldsymbol{\mu} = e^{-\eta C}$ before any updates to the primal variables. By Lemma 1, $L(0, 0, 0) \leq L(\lambda^{(1)}, \xi^{(1)})$. Then we have

$$L(\lambda^*, \xi^*) - L(\lambda^{(1)}, \xi^{(1)}) \leq L(\lambda^*, \xi^*) - L(0, 0) \leq L(\lambda^*, \xi^*).$$

We may establish an upper bound on $L(\lambda^*, \xi^*)$ by finding a feasible point in the primal objective (Proj). It is easy to verify that $\boldsymbol{\mu}$ is in $\mathbb{L}_2$ if $\forall ij \in \mathcal{E}$ and $\forall i \in \mathcal{V}$, $\boldsymbol{\mu}_{ij}(x_i, x_j) = \frac{1}{d^2}$ and $\boldsymbol{\mu}_i(x_i) = \frac{1}{d}$. With this choice of $\boldsymbol{\mu}$, the value of (Proj) is

$$\begin{aligned}
\mathcal{D}_\Phi(\boldsymbol{\mu}, \exp(^{-\eta C})) &= \sum_{ij\in\mathcal{E}} (\eta \mathbb{E}_U[C_{ij}] - 1 - \log d^2) + \sum_{i\in\mathcal{V}} (\eta \mathbb{E}_U[C_i] - 1 - \log d) \\
&\quad + \sum_{ij\in\mathcal{E}} \sum_{x_i, x_j \in \chi} \exp(-\eta C_{ij}(x_i, x_j)) + \sum_i \sum_{x\in\chi} \exp(-\eta C_i(x)) \\
&\leq \|\frac{\eta}{d} C + \exp(-\eta C)\|_1 - (|\mathcal{V}| + |\mathcal{E}|)(\log d + 1),
\end{aligned}$$

where $\mathbb{E}_U$ denotes the uniform distribution. where the last inequality follows from the fact that $C_{ij}(0,0) = C_i(0) = 0$. Therefore,

$$L(\lambda^*, \xi^*) - L(\lambda^{(1)}, \xi^{(1)}) \leq L(\lambda^*, \xi^*) \leq \|\frac{\eta}{d}C + \exp(-\eta C)\|_1 - (|\mathcal{V}| + |\mathcal{E}|)(\log d + 1).$$

We now proceed to show the following (direct) bound on $L(\lambda^*, \xi^*) - L(\lambda^{(1)}, \xi^{(1)})$:

$$L(\lambda^*, \xi^*) - L(\lambda^{(1)}, \xi^{(1)}) \leq \sum_{ij \in \mathcal{E}} \left[\log \left(\sum_{x_i, x_j \in \chi} \exp\left(-\eta C_{ij}(x_i, x_j)\right) \right) + \sum_{x_i, x_j \in \chi} \frac{\eta}{4} C_{ij}(x_i, x_j) \right] + \sum_{i \in \mathcal{V}} \left[\log \left(\sum_{x \in \chi} \exp\left(-\eta C_i(x)\right) \right) + \sum_{x \in \chi} \frac{\eta}{2} C_i(x) \right].$$

We work under the assumption that at any time k, all the component distributions of $\boldsymbol{\mu}^{(k)}$ are normalized so its entries sum to 1. Notice that in this case

$$L(\lambda^*, \xi^*) - L(\lambda^{(1)}, \xi^{(1)}) = \sum_{ij \in \mathcal{E}} \xi_{ij}^{(1)} - \xi_{ij}^* + \sum_{i \in \mathcal{V}} \xi_i^{(1)} - \xi_i^*.$$

If we initialize our algorithm to $\lambda^{(1)} = 0$, and $\xi^{(1)}$ be the normalization factors corresponding to this choice of λ, then

$$\sum_{ij \in \mathcal{E}} \xi_{ij}^{(1)} + \sum_{i \in \mathcal{V}} \xi_i^{(1)} = \sum_{ij \in \mathcal{E}} \log \left(\sum_{x_i, x_j \in \chi} \exp\left(-\eta C(x_i, x_j)\right) \right) + \sum_{i \in \mathcal{V}} \log \left(\sum_{x \in \chi} \exp\left(-\eta C(x)\right) \right).$$

Notice that at optimality λ^*, ξ^*, for all $ij \in \mathcal{E}$ and, for all x_i, x_j,

$$\exp\left(-\eta C_{ij}(x_i, x_j) - \lambda_{ij}^*(x_i) - \lambda_{ji}^*(x_j) - \xi_{ij}^*\right) = \left(\boldsymbol{\mu}_\eta^*\right)_{ij}(x_i, x_j) \in [0, 1].$$

And for all $i \in \mathcal{V}$ and for all x,

$$\exp\left(-\eta C_i(x) - \xi_i^* + \sum_{j \in N_r(i)} \lambda_{ij}^*(x) + \sum_{j \in N_c(i)} \lambda_{ji}^*(x) \right) = \left(\boldsymbol{\mu}_\eta^*\right)_i(x) \in [0, 1].$$

Therefore, for all $ij \in \mathcal{E}$ and for all x_i, x_j:

$$-\eta C_{ij}(x_i, x_j) - \lambda_{ij}^*(x_i) - \lambda_{ij}^*(x_j) - \xi_{ij}^* \leq 0 \tag{7}$$

For all $i \in \mathcal{V}$ and for all x:

$$-\eta C_i(x) - \xi_i^* + \sum_{j \in N_r(i)} \lambda_{ij}^*(x) + \sum_{j \in N_c(i)} \lambda_{ji}^*(x) \leq 0 \tag{8}$$

Summing Equations (7) and (8) over all $ij \in \mathcal{E}$, $i \in \mathcal{V}$ and $x_i, x_j, x \in \chi$ yields:

$$-\sum_{ij \in \mathcal{E}} \xi_{ij}^* - \sum_{i \in \mathcal{V}} \xi_i^* \leq \sum_{ij \in \mathcal{E}} \sum_{x_i, x_j \in \chi} \frac{\eta}{d^2} C_{ij}(x_i, x_j) + \sum_{i \in \mathcal{V}} \sum_{x \in \chi} \frac{\eta}{d} C_i(x) \tag{9}$$

And, therefore,

$$L(\lambda^*, \xi^*) - L(\lambda^{(1)}, \xi^{(1)}) \leq \sum_{ij \in \mathcal{E}} \left[\log \left(\sum_{x_i, x_j \in \chi} \exp\left(-\eta C(x_i, x_j)\right) \right) + \sum_{x_i, x_j \in \chi} \frac{\eta}{d^2} C_{ij}(x_i, x_j) \right] + \sum_{i \in \mathcal{V}} \left[\log \left(\sum_{x \in \chi} \exp\left(-\eta C(x)\right) \right) + \sum_{x \in \chi} \frac{\eta}{d} C_i(x) \right]. \tag{10}$$

Notice that the RHS of the equation above is positive since: $\sum_{i=1}^{\ell} \exp(a_i) \geq \frac{1}{\ell} \sum_{i=1}^{\ell} \exp(a_i) \geq \exp\left(\frac{\sum_{i=1}^{\ell} a_i}{\ell}\right)$ for all $\ell \in \mathbb{N}$ and all $a_1, \cdots, a_\ell \in \mathbb{R}$. Combining Equations (6) and (10) and the observation that $L(0,0) \leq L(\lambda^{(1)}, \xi^1)$ (by virtue of Lemma 1) we obtain the final result. □

In the case when all entries of C are positive it may be the case that $S \gg \| \exp(-\eta C) \|_1$.

D.5 Complete Proof of Theorem 2

In this section, we will complete the proof of Theorem 2 by handling the case of EMP-cyclic. We require two additional technical lemmas on the l_1 distance between updated variables. We will use $r(\cdot)$ and $c(\cdot)$ to denote row and column sums respectively of joint distribution matrices.

Lemma 5. *Let $a, b \in \Sigma_d$ be two points in the simplex and let $p \in \mathbb{R}_+^d$ s.t. $\min(a_i, b_i) \leq p_i \leq \max(a_i, b_i)$ for all $1 \leq i \leq d$. Let $c \in \Sigma_d$ defined as $c = \frac{p}{\sum_i p_i}$. Then:*

$$\max(\|a - c\|_1, \|b - c\|_1) \leq \|a - b\|_1$$

Proof. We only need to prove that $\|a - c\| \leq \|a - b\|_1$. From $\min(a_i, b_i) \leq p_i \leq max(a_i, b_i)$ we obtain:

$$|a_i - p_i| + |b_i - p_i| = |a_i - b_i|.$$

Let $t = \frac{1}{\sum_i p_i}$. The following relationships hold:

$$\begin{aligned} \|a - c\|_1 &= \sum_i |a - tp_i| = \sum_i |a_i - p_i + (1-t)p_i| \\ &\leq \sum_i |a_i - p_i| + \sum_i |(1-t)p_i|. \end{aligned}$$

Note that

$$\sum_i |(1-t)p_i| = |1-t| \sum_i p_i = \frac{|1-t|}{t} = |\frac{1}{t} - 1| = |\sum_i p_i - 1|,$$

and

$$\sum_i |b_i - p_i| \geq |\sum_i b_i - p_i| = |1 - \sum_i p_i| = |\sum_i p_i - 1|.$$

Therefore,

$$\|a - c\|_1 \leq \sum_i |a_i - p_i| + \sum_i |b_i - p_i| = \sum_i |a_i - b_i| = \|a - b\|_1.$$

The result follows.

□

Let $A \in \Sigma_{d \times d}$ with elements a_{ij} be a matrix representing joint distribution probabilities. For $p = \begin{bmatrix} p_1 & \dots & p_d \end{bmatrix}^\top \in \Sigma_d$, define

$$\widetilde{A} = \frac{1}{z} \begin{bmatrix} a_{11}\sqrt{\frac{p_1}{r(A)_1}} & \cdots & a_{1d}\sqrt{\frac{p_1}{r(A)_1}} \\ \vdots & \ddots & \vdots \\ a_{d1}\sqrt{\frac{p_d}{r(A)_d}} & \cdots & a_{dd}\sqrt{\frac{p_d}{r(A)_d}} \end{bmatrix}$$

where z is a normalization term, such that the new probabilities matrix sums to one. The notation $r(A)_i$ denotes the ith element of row sum vector $r(A)$.

Lemma 6. *The following inequality holds on the difference between A and $\widetilde{A}$:*

$$\|c(\widetilde{A}) - c(A)\|_1 \leq \|r(\widetilde{A}) - r(A)\|_1$$

Proof.

$$\begin{aligned}
\|c(\widetilde{A}) - c(A)\|_1 &= \sum_{j=1}^{d} \left| \sum_{i=1}^{d} \frac{a_{ij}}{z} \left(\sqrt{\frac{p_i}{r(A)_i}} - z \right) \right| \\
&\leq \sum_{i,j} \frac{a_{ij}}{z} \left| \sqrt{\frac{p_i}{r(A)_i}} - z \right| \\
&= \sum_{i} \frac{r(A)_i}{z} \left| \sqrt{\frac{p_i}{r(A)_i}} - z \right| \\
&= \sum_{i} \left| \sqrt{\frac{r(A)_i p_i}{z}} - r(A)_i \right| \\
&= \|r(\widetilde{A}) - r(A)\|_1.
\end{aligned}$$

□

This proof of Theorem 2 relies heavily on the primal and dual variables at given times throughout the algorithm. As such, it is necessary to define precise notation for these temporal events. We note that there are two loops in the algorithm: an outer loop that controls the iterations and an inner one that loops over all edges in $\mathcal{E}$. The outer loop's current iteration is given by $k \geq 0$, as defined and updated in Algorithm 1. We denote the current step of the inner loop by t where $1 \leq t \leq 4|\mathcal{E}|$. This is due to the fact that there are four projections for each edge ($\mathcal{X}_{ij\to i}$, $\mathcal{X}_{ij,i}$, $\mathcal{X}_{ij\to j}$, and $\mathcal{X}_{ij,j}$) in one full iteration for $\mathbb{L}_2$. Thus the algorithm alternates between enforcing consistency between an edge and vertex and normalizing the local distributions.

The value of $\boldsymbol{\mu}$ at iteration k and step t within iteration k is denoted by $\boldsymbol{\mu}^{(k,t)}$. For example, at the very start of the algorithm, we are at iteration $k = 1$ and step $t = 1$ with initial value $\boldsymbol{\mu}^{(1,1)}$, which is equal to $\exp(-\eta C)$ with normalized vertex marginal and edge joint distributions. The constraint set onto which a projection is made at t in any iteration is denoted by $\mathcal{X}^{(t)}$. Note that we drop k in the constraint set notation because the order in which the projections occur is always the same.

Proof of Theorem 2. Let k^* be the first iteration such that the termination condition in Algorithm 1 with respect to ϵ is met. For k such that $1 \leq k \leq k^*$, there exists $ij \in \mathcal{E}$ such that $\|r(\boldsymbol{\mu}_{ij}^{(k,1)}) - \boldsymbol{\mu}_i^{(k,1)}\|_1 \geq \epsilon$ or $\|c(\boldsymbol{\mu}_{ij}^{(k,1)}) - \boldsymbol{\mu}_j^{(k,1)}\|_1 \geq \epsilon$.

First consider the case where $\|c(\boldsymbol{\mu}_{ij}^{(k,1)}) - \boldsymbol{\mu}_j^{(k,1)}\|_1 \geq \epsilon$. Let t be chosen such that $\mathcal{X}^{(t)} = \mathcal{X}_{ij\to j}$. Note that $\boldsymbol{\mu}_j^{(k,t)}$ can move within the ϵ-ball of $c(\boldsymbol{\mu}_{ij}^{(k,t)})$ between times 1 and t of the kth iteration due to earlier projections involving vertex j. However, $\boldsymbol{\mu}_{ij}^{(k,t')} = \boldsymbol{\mu}_{ij}^{(k,1)}$ for all $t' \leq t-2$ because it is only updated at step $t-2$ where $\mathcal{X}^{(t-2)} = \mathcal{X}_{ij\to i}$. Then, by repeatedly applying the triangle inequality, we have

$$\begin{aligned}
\epsilon &\leq \|c(\boldsymbol{\mu}_{ij}^{(k,1)}) - \boldsymbol{\mu}_j^{(k,1)}\|_1 \\
&\leq \|c(\boldsymbol{\mu}_{ij}^{(k,t-2)}) - \boldsymbol{\mu}_j^{(k,1)}\|_1 \\
&\leq \|c(\boldsymbol{\mu}_{ij}^{(k,t-2)}) - \boldsymbol{\mu}_j^{(k,t)}\|_1 + \sum_{t' \in \mathcal{T}_{j,r}^{(t)} \cup \mathcal{T}_{j,c}^{(t)}} \|\boldsymbol{\mu}_j^{(k,t')} - \boldsymbol{\mu}_j^{(k,t'+2)}\|_1 \\
&\leq \|c(\boldsymbol{\mu}_{ij}^{(k,t)}) - \boldsymbol{\mu}_j^{(k,t)}\|_1 + \|c(\boldsymbol{\mu}_{ij}^{(k,t)}) - c(\boldsymbol{\mu}_{ij}^{(k,t-2)})\|_1 \\
&\quad + \sum_{t' \in \mathcal{T}_{j,r}^{(t)} \cup \mathcal{T}_{j,c}^{(t)}} \|\boldsymbol{\mu}_j^{(k,t')} - \boldsymbol{\mu}_j^{(k,t'+2)}\|_1,
\end{aligned}$$

where $\mathcal{T}_{j,r}^{(t)}$ and $\mathcal{T}_{j,c}^{(t)}$ are sets of times before t where a projection (for row and column consistency, respectively) caused $\boldsymbol{\mu}_j$ to be updated:

$$\mathcal{T}_{j,r}^{(t)} \stackrel{\text{def.}}{=} \{t' < t \ : \ \exists \ell \in N_r(i) \ \text{ s.t. } \ \mathcal{X}^{(t')} = \mathcal{X}_{j\ell \to j}\}$$
$$\mathcal{T}_{j,c}^{(t)} \stackrel{\text{def.}}{=} \{t' < t \ : \ \exists \ell \in N_c(i) \ \text{ s.t. } \ \mathcal{X}^{(t')} = \mathcal{X}_{\ell j \to j}\}.$$

Therefore, $\boldsymbol{\mu}_j^{(k,t'+2)}$ is the result of enforcing consistency with another edge of i and then normalizing $\boldsymbol{\mu}_j$. Let $e_{t'}$ denote the edge (incident on j) onto which projections are occurring at step $t' \in \mathcal{T}_{j,r}^{(t)} \cup \mathcal{T}_{j,c}^{(t)}$. From Lemma 5, if $t' \in \mathcal{T}_{j,r}(t)$, then

$$\|\boldsymbol{\mu}_j^{(k,t')} - \boldsymbol{\mu}_j^{(k,t'+2)}\|_1 \leq \|\boldsymbol{\mu}_j^{(k,t')} - r(\boldsymbol{\mu}_{e_{t'}}^{(k,t')})\|_1.$$

If $t' \in \mathcal{T}_{j,c}^{(t)}$, then

$$\|\boldsymbol{\mu}_j^{(k,t')} - \boldsymbol{\mu}_j^{(k,t'+2)}\|_1 \leq \|\boldsymbol{\mu}_j^{(k,t')} - c(\boldsymbol{\mu}_{e_{t'}}^{(k,t')})\|_1.$$

Similarly, by combining Lemma 5 and Lemma 6, we have

$$\|c(\boldsymbol{\mu}_{ij}^{(k,t)}) - c(\boldsymbol{\mu}_{ij}^{(k,t-2)})\|_1 \leq \|r(\boldsymbol{\mu}_{ij}^{(k,t)}) - r(\boldsymbol{\mu}_{ij}^{(k,t-2)})\|_1 \leq \|\boldsymbol{\mu}_i^{(k,t-2)} - r(\boldsymbol{\mu}_{ij}^{(k,t-2)})\|_1.$$

Note that since the variables are normalized at every even step, they are individually valid probability distributions, and so the Hellinger inequality can be applied. For distributions, p and q, the inequality states

$$\frac{1}{4}\|p - q\|_1^2 \leq 2h^2(p, q).$$

Therefore,

$$\begin{aligned}\frac{\epsilon^2}{4(\deg(\mathcal{G}_k)+1)} &\leq 2h^2(c(\boldsymbol{\mu}_{ij}^{(k,t)}), \boldsymbol{\mu}_j^{(k,t)}) + 2h^2(r(\boldsymbol{\mu}_{ij}^{(k,t-2)}), \boldsymbol{\mu}_i^{(k,t-2)}) \\ &\quad + \sum_{t' \in \mathcal{T}_{j,r}^{(t)}} 2h^2(r(\boldsymbol{\mu}_{e_{t'}}^{(k,t')}), \boldsymbol{\mu}_j^{(k,t')}) + \sum_{t' \in \mathcal{T}_{j,c}^{(t)}} 2h^2(c(\boldsymbol{\mu}_{e_{t'}}^{(k,t')}), \boldsymbol{\mu}_j^{(k,t')}) \\ &\leq L^{(k+1,1)} - L^{(k,1)}.\end{aligned}$$

The last inequality follows from telescoping over all steps in iteration k due to Lemma 1. This proof was for the case when $\|c(\boldsymbol{\mu}_{ij}^{(k,1)}) - \boldsymbol{\mu}_j^{(k,1)}\|_1 \geq \epsilon$. For the case when $\|r(\boldsymbol{\mu}_{ij}^{(k,1)}) - \boldsymbol{\mu}_i^{(k,1)}\|_1 \geq \epsilon$, the procedure is identical except we may ignore the term $\|c(\boldsymbol{\mu}_{ij}^{(k,t)}) - c(\boldsymbol{\mu}_{ij}^{(k,t-2)})\|_1$ since $\boldsymbol{\mu}_{ij}$ is constant within iteration k until the projection onto $X_{ij \to i}$. Thus, the improvement lower bound still holds.

Putting these results together with Lemma 2, we see that as long as a single constraint is violated above the ϵ threshold at the start of an iteration, it is possible to show that the value of L increases by at least $\epsilon^2/4(\deg(\mathcal{G}_k)+1)$ during the iteration. This implies that EMP-cyclic terminates in at most $\lceil \frac{4\mathcal{S}_0(\deg(\mathcal{G}_k)+1)}{\epsilon^2} \rceil$ iterations.

□

D.6 Proof of Theorem 3

We start by defining a version of $\mathbb{L}_2$ with slack vectors. Let ν be a vector indexed in a similar way as $\boldsymbol{\mu}$, where $\{\nu_{ij}, \nu_{ji}\}_{ij \in \mathcal{E}}$. We define the slack ν as $\nu_{ij} = \boldsymbol{\mu}_{ij}\mathbb{1} - \boldsymbol{\mu}_i$ and $\nu_{ji} = \boldsymbol{\mu}_{ij}^\top \mathbb{1} - \boldsymbol{\mu}_j$. Then we define the slack polytope $\mathbb{L}_2^\nu$ as

$$\mathbb{L}_2^\nu \stackrel{\text{def.}}{=} \left\{ \boldsymbol{\mu} \geq 0 \ : \ \begin{array}{ll} \boldsymbol{\mu}_i \in \Sigma_d & \forall i \in \mathcal{V} \\ \boldsymbol{\mu}_{ij}\mathbb{1} = \boldsymbol{\mu}_i + \nu_{ij} & \forall ij \in \mathcal{E} \\ \boldsymbol{\mu}_{ij}^\top \mathbb{1} = \boldsymbol{\mu}_j + \nu_{ji} & \forall ij \in \mathcal{E} \\ \mathbb{1}^\top \boldsymbol{\mu}_{ij} \mathbb{1} = 1 & \forall ij \in \mathcal{E} \end{array} \right\}. \tag{11}$$

Notice that by definition the slack vectors ν satisfy that, for all $ij \in \mathcal{E}$, $\nu_{ij}^\top \mathbb{1} = \nu_{ji}^\top \mathbb{1} = 0$. The main difference between $\mathbb{L}_2$ and $\mathbb{L}_2^\nu$ lies in that the joints do not marginalize exactly to the vertex probabilities but do so up to a slack. Consider the entropy-regularized linear program corresponding to $\mathbb{L}_2^\nu$:

$$\min \quad \langle C, \boldsymbol{\mu} \rangle - \frac{1}{\eta} H(\boldsymbol{\mu}) \quad \text{s.t.} \quad \boldsymbol{\mu} \in \mathbb{L}_2^\nu, \tag{Reg-slack}$$

Introducing the exact same ensemble of dual variables λ, ξ as in the Lyapunov function derivation, its dual function equals

$$\begin{aligned} L^\nu(\lambda, \xi) = &- \sum_{ij \in \mathcal{E}} \sum_{x_i, x_j \in \chi} \exp\left(-\eta C_{ij}(x_i, x_j) - \lambda_{ij}(x_i) - \lambda_{ji}(x_j) - \xi_{ij}\right) \\ &- \sum_{ij \in \mathcal{E}} \sum_{x_i, x_j \in \chi} \Big(\lambda_{ij}(x_i)\nu_{ij}(x_i) + .\lambda_{ji}(x_j)\nu_{ji}(x_j)\Big) \\ &- \sum_{i \in \mathcal{V}} \sum_{x \in \chi} \exp\left(-\eta C_i(x) - \xi_i + \sum_{j \in N_r(i)} \lambda_{ij}(x) + \sum_{j \in N_c(i)} \lambda_{ji}(x)\right) \\ &- \sum_{ij \in \mathcal{E}} \xi_{ij} - \sum_{i \in \mathcal{V}} \xi_i + \sum_{ij \in \mathcal{E}} \sum_{x_i, x_j \in \chi} \exp(-\eta C_{ij}(x_i, x_j)) + \sum_i \sum_{x \in \chi} \exp(-\eta C_i(x)). \end{aligned} \tag{12}$$

Furthermore, if λ^*, ξ^* were a set of optimal dual variables, the optimal primal $\boldsymbol{\mu}^*$ can be computed via

$$\boldsymbol{\mu}_{ij}^*(x_i, x_j) = \exp\left(-\eta C_{ij}(x_i, x_j) - \lambda_{ij}^*(x_i) - \lambda_{ji}^*(x_j) - \xi_{ij}^*\right) \tag{13}$$

$$\boldsymbol{\mu}_i^*(x_i) = \exp\left(-\eta C_i(x_i) - \xi_i^* + \sum_{j \in N_r(i)} \lambda_{ij}^*(x_i) + \sum_{j \in N_c(i)} \lambda_{ji}^*(x_i)\right). \tag{14}$$

They satisfy the same formulae as the problem without slack variables. Since dual optimality is equivalent to primal feasibility, whenever an iterate of EMP satisfies slack of ν, its corresponding primal solution is optimal for (Reg-slack).

We start with a useful manipulation lemma:

Lemma 7. *Let ν, ν' be two slack vectors and let $\boldsymbol{\mu} \in \mathbb{L}_2^\nu$. Assume $\|\nu'\|_\infty \leq \frac{1}{2d}$.*

1. *If for all $ij \in \mathcal{E}$ and $i \in \mathcal{V}$, $\boldsymbol{\mu}_i + \nu'_{ij} \in \Sigma_d$, then there exists a vector $\boldsymbol{\mu}' \in \mathbb{L}_2^{\nu'}$ such that*

$$\|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_1 \leq 2\|\nu - \nu'\|_1. \tag{15}$$

2. *If $\nu = 0$[4], then there exists a vector $\boldsymbol{\mu}' \in \mathbb{L}_2^{\nu'}$ such that*

$$\|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_1 \leq 6d \deg(\mathcal{G})\|\nu'\|_1. \tag{16}$$

Proof. First we consider the case when for all $ij \in \mathcal{E}$, $\boldsymbol{\mu}_i + \nu'_{ij}$ is a valid distribution (in other words, all its entries are in $[0, 1]$ and its values sum to 1). In this case, we can argue for the existence of $\boldsymbol{\mu}'$ via the following:

Let $\boldsymbol{\mu}'_i = \boldsymbol{\mu}_i$ for all $i \in \mathcal{V}$. Let $ij \in \mathcal{E}$ and observe that $\boldsymbol{\mu}_{ij}\mathbb{1} = \boldsymbol{\mu}_i + \nu_{ij}$ and $\boldsymbol{\mu}_{ij}^\top \mathbb{1} = \boldsymbol{\mu}_j + \nu_{ji}$. We invoke Lemma 7 in Altschuler et al. (2017) to claim the existence of $\boldsymbol{\mu}'_{ij}$ such that $\boldsymbol{\mu}'_{ij}\mathbb{1} = \boldsymbol{\mu}_i + \nu'_{ij}$ and $(\boldsymbol{\mu}'_{ij})^\top \mathbb{1} = \boldsymbol{\mu}_j + \nu'_{ji}$ and

$$\begin{aligned} \|\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}'_{ij}\|_1 &\leq 2\left(\|\boldsymbol{\mu}_i + \nu_{ij} - \boldsymbol{\mu}_i - \nu'_{ij}\|_1 + \|\boldsymbol{\mu}_j + \nu_{ji} - \boldsymbol{\mu}_j - \nu'_{ji}\|_1\right) \\ &= 2\left(\|\nu_{ij} - \nu'_{ij}\|_1 + \|\nu_{ji} - \nu'_{ji}\|_1\right). \end{aligned}$$

Setting $\boldsymbol{\mu}'$ to be the ensemble with values $\{\boldsymbol{\mu}'_i\}_{i \in \mathcal{V}}$ and $\{\boldsymbol{\mu}'_{ij}\}_{ij \in \mathcal{E}}$ the result follows.

Now we consider the case when there exist $ij \in \mathcal{E}$ such that $\boldsymbol{\mu}_i + v'_{ij}$ does not lie in the probability simplex. In this case we will have to define $\boldsymbol{\mu}'_i$ different from $\boldsymbol{\mu}_i$. Consider some $i \in \mathcal{V}$. Let $N(i)$ be the set of neighbouring vertices to i and we abuse notation slightly and use ν_{ij} for $j \in N(i)$ to denote the slack on i as of the edge marginal shared by i and j. We define $\boldsymbol{\mu}'_i$ in the following way:

[4]We do not require that $\boldsymbol{\mu}_i + \nu'_{ij} \in \Sigma_d$

1. If $\boldsymbol{\mu}_i + \nu'_{ij} \in \Sigma_d$ for all $j \in N(i)$ then let $\boldsymbol{\mu}'_i = \boldsymbol{\mu}_i$.
2. Otherwise, let $\{x_1, \cdots, x_r\} \subseteq [d]$ be the entries of $\boldsymbol{\mu}_i$ such that for all $x_\tau \in \{x_1, \cdots, x_r\}$ there exists at least one $j \in N(i)$ for which $\left[\boldsymbol{\mu}_i + \nu'_{ij}\right](x_\tau) \notin [0,1]$. Therefore, we must define $\boldsymbol{\mu}'_i$ such that

$$\max_j \|\nu'_{ij}\|_\infty \leq \boldsymbol{\mu}_i(x) \leq 1 - \max_j \|\nu'_{ij}\|_\infty,$$

which can be done by taking the convex combination of μ_i with the uniform distribution:

$$\boldsymbol{\mu}'_i = (1-\theta)\boldsymbol{\mu}_i + \frac{\theta}{d}.$$

Setting $\theta = d \max_j \|\nu'_{ij}\|_\infty$ guarantees this outcome because we are given that $\|\nu\|_\infty \leq \frac{1}{2d}$. Furthermore, we have

$$\begin{aligned}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}'_i\|_1 &= \sum_x |\boldsymbol{\mu}_i(x) - \boldsymbol{\mu}'_i(x)| \\ &\leq 2d \max_j \|\nu'_{ij}\|_\infty.\end{aligned}$$

This, in turn, implies $\sum_{i \in V} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}'_i\|_1 \leq 2d\|\nu'\|_1$. Then, we apply the result of Altschuler et al. (2017) again to achieve existence of $\{\boldsymbol{\mu}'_{ij}\}_{ij \in \mathcal{E}}$ such that

$$\begin{aligned}\|\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}'_{ij}\|_1 &\leq 2\left(\|\boldsymbol{\mu}'_i - \boldsymbol{\mu}_i - \nu'_{ij}\|_1 + \|\boldsymbol{\mu}'_j - \boldsymbol{\mu}_j - \nu'_{ji}\|_1\right) \\ &= 2\left(\|\nu'_{ij}\|_1 + \|\nu'_{ji}\|_1 + \|\boldsymbol{\mu}_i - \boldsymbol{\mu}'_i\|_1 + \|\boldsymbol{\mu}_j - \boldsymbol{\mu}'_j\|_1\right).\end{aligned}$$

Summing over these yields $\sum_{ij \in \mathcal{E}} \|\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}'_{ij}\|_1 \leq 2\|\nu'\|_1 + 2d\deg(\mathcal{G})\|\nu'\|_1$. Therefore $\|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_1 \leq 6d\deg(\mathcal{G})\|\nu'\|_1$

□

We additionally require a similar lemma which allows us to project from one polytope to another while bounding the probabilities away from zero.

Lemma 8. *Fix τ such that $0 < \tau \leq \frac{1}{8d^2}$ and a slack vector ν such that $\|\nu\|_\infty \leq \frac{1}{4d}$. If $\boldsymbol{\mu} \in \mathbb{L}_2^\nu$, then there exists a vector $\boldsymbol{\mu}' \in \mathbb{L}_2$ such that*

$$\begin{aligned}\boldsymbol{\mu}'_i(x_i) &\geq \tau \quad \forall i \in \mathcal{V},\ x_i \in \chi \\ \boldsymbol{\mu}'_{ij}(x_i, x_j) &\geq \tau \quad \forall ij \in \mathcal{E},\ x_i, x_j \in \chi \\ \|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_1 &\leq 2\|\nu\|_1 + 2(m+n)d^2\tau.\end{aligned}$$

If $\boldsymbol{\mu} \in \mathbb{L}_2$, then there exists a vector $\boldsymbol{\mu}' \in \mathbb{L}_2^\nu$ such that

$$\begin{aligned}\boldsymbol{\mu}'_i(x_i) &\geq \tau \quad \forall i \in \mathcal{V},\ x_i \in \chi \\ \boldsymbol{\mu}'_{ij}(x_i, x_j) &\geq \tau \quad \forall ij \in \mathcal{E},\ x_i, x_j \in \chi \\ \|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_1 &\leq 6d\deg(\mathcal{G})\|\nu\|_1 + 8(|\mathcal{E}| + n)d^2\tau.\end{aligned}$$

Proof. We address each case individually.

1. We use the first result from Lemma 7, which yields $\widehat{\boldsymbol{\mu}} \in \mathbb{L}_2$ such that $\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|_1 \leq 2\|\nu\|_1$. If the probabilities are already bounded below τ, then we are done; however, we must handle the worst case. As in the proof of Lemma 7, we compute a convex combination of $\widehat{\boldsymbol{\mu}}$ with the uniform distribution to draw the distribution away from zero values. Define

$$\begin{aligned}\boldsymbol{\mu}'_i &:= (1-\theta)\widehat{\boldsymbol{\mu}}_i + \frac{\theta}{d}\mathbb{1} \\ \boldsymbol{\mu}'_{ij} &:= (1-\theta)\widehat{\boldsymbol{\mu}}_{ij} + \frac{\theta}{d^2}\mathbb{1}.\end{aligned}$$

where we set $\theta = \tau d^2$ which ensures that $\theta \in [0,1]$ and $\boldsymbol{\mu} \geq \tau$. Then, note that

$$\|\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}'_i\|_1 = \sum_x |\frac{\theta}{d} - \theta\widehat{\boldsymbol{\mu}}_i(x)| \leq 2\tau d^2$$
$$\|\widehat{\boldsymbol{\mu}}_{ij} - \boldsymbol{\mu}'_{ij}\|_1 = \sum_{x_i,x_j} |\frac{\theta}{d} - \theta\widehat{\boldsymbol{\mu}}_{ij}(x_i,x_j)| \leq 2\tau d^2.$$

By the triangle inequality, we have

$$\|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_1 \leq \|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|_1 + \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}'\|_1 \leq 2\|\nu\|_1 + 2(n+m)d^2\tau.$$

2. In the second case, we start by constructing a distribution δ, which is nearly uniform but lives in the slack polytope is and bounded away from zero by at least τ.

   For each $i \in \mathcal{V}$, we take $\delta_i \in \Sigma_d$ to be the uniform distribution where $\delta_i(x) = \frac{1}{d} \geq \tau$. Since $\|\nu\|_\infty \leq \frac{1}{4d}$, we perturb the uniform distribution with ν for each $j \in N(i)$, generating $\delta_i^j := \delta_i + \nu_{ij}$. Again, we are abusing notation slightly by using ν_{ij} to denote marginalization of edge ij to vertex i. Note that $\delta_i^j \in \Sigma_d$, so we can define the product distribution $\delta_{ij} = \delta_i^j(\delta_j^i)^\top \in \mathcal{U}_d(\delta_i^j, \delta_j^i)$, which, by construction, marginalizes such that the full vector δ given by the ensemble $\{\delta_i\}_{i\in\mathcal{V}}$ and $\{\delta_{ij}\}_{ij\in\mathcal{E}}$ is in $\mathbb{L}_2^\nu$. Furthermore, each component can be bounded below as

$$\begin{aligned}\delta_{ij}(x_i,x_j) &= \frac{1}{d^2} + \frac{\nu_{ij}(x_i)}{d} + \frac{\nu_{ji}(x_i)}{d} + \nu_{ij}(x_i)\nu_{ji}(x_j) \\ &\geq \frac{1}{d^2} - \frac{1}{2d^2} - \frac{1}{16d^2} \\ &\geq \frac{1}{4d^2}.\end{aligned}$$

   Now, as before, we know there exists $\widehat{\boldsymbol{\mu}} \in \mathbb{L}_2^2$ such that $\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|_1 \leq 6d\deg(\mathcal{G})\|\nu\|_1$ from Lemma 7. Therefore, we can take the convex combination of $\boldsymbol{\mu}' = (1-\theta)\boldsymbol{\mu} + \theta\delta$ to get $\boldsymbol{\mu}' \in \mathbb{L}_2^\nu$ such that $\boldsymbol{\mu}' \geq \frac{\theta}{4d^2}$ in all entries. Taking $\theta = 4d^2\tau \in [0,1]$ ensures that $\boldsymbol{\mu}' \geq \tau$. Furthermore, the difference can be computed as

$$\|\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}'_i\|_1 = \sum_x |\frac{\theta}{d} - \theta\widehat{\boldsymbol{\mu}}(x)| \leq 8d^2\tau$$
$$\|\widehat{\boldsymbol{\mu}}_{ij} - \boldsymbol{\mu}'_{ij}\|_1 = \sum_{x_i,x_j} |\theta\delta_{ij}(x_i,x_j) - \theta\widehat{\boldsymbol{\mu}}_{ij}(x_i,x_j)| \leq 8d^2\tau.$$

   Therefore, we have $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}'\|_1 \leq 8(|\mathcal{E}| + n)d^2\tau$, which by triangle inequality implies

$$\|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_1 \leq \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_1 + \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}'\|_1 \leq 6d\deg(\mathcal{G})\|\nu\|_1 + 8(|\mathcal{E}| + n)d^2\tau.$$

□

We have now the necessary ingredients to prove the first theorem of this section, which provides a bound on the l_1 distance between the final iterate $\boldsymbol{\mu}^k$ of Algorithms 1 and 2 and the solution $\boldsymbol{\mu}_\eta^*$ of (Reg). Crucially we analyze these iterates under the assumption all their component distributions $\boldsymbol{\mu}_i^{(k)}$ for $i \in \mathcal{V}$ and $\boldsymbol{\mu}_{ij}^{(k)}$ for $ij \in \mathcal{E}$ are normalized.

Theorem 4. *Let $\boldsymbol{\mu}^{(k)}$ is the kth iterate of EMP and let $\nu^{(k)}$ be the slack vector corresponding to $\boldsymbol{\mu}^{(k)}$ such that $\|\nu^{(k)}\|_\infty \leq \frac{1}{4d}$. In other words,*

$$\nu_{ij}^{(k)} = \boldsymbol{\mu}_{ij}^{(k)}\mathbb{1} - \boldsymbol{\mu}_i^{(k)}$$
$$\nu_{ji}^{(k)} = \left(\boldsymbol{\mu}_{ij}^{(k)}\right)^\top \mathbb{1} - \boldsymbol{\mu}_j^{(k)}.$$

Fix $\tau > 0$ such that $\tau \leq \frac{1}{8d^2}$. Let $\boldsymbol{\mu}^{(k)}(2)$ be the pseudo-marginal vector in $\mathbb{L}_2$ produced by the first case of Lemma 8 when fed with $\boldsymbol{\mu}^{(k)}$ and τ. Then,

$$\sum_{i\in\mathcal{V}} \frac{1}{2}\left\|\left(\boldsymbol{\mu}^{(k)}(2)\right)_i - \left(\boldsymbol{\mu}^*_\eta\right)_i\right\|_1^2 + \sum_{ij\in\mathcal{E}} \frac{1}{2}\left\|\left(\boldsymbol{\mu}^{(k)}(2)\right)_{ij} - \left(\boldsymbol{\mu}^*_\eta\right)_{ij}\right\|_1^2 \leq (\eta\|C\|_\infty + \log 1/\tau)\left(8d\deg(\mathcal{G})\|\nu\|_1 + 10(|\mathcal{E}|+n)d^2\tau\right).$$

Proof. By definition $\boldsymbol{\mu}^{(k)} \in \mathbb{L}_2^{\nu^{(k)}}$. In fact, $\boldsymbol{\mu}^{(k)}$ is the optimizer of the following regularized linear program:

$$\min \quad \langle C, \boldsymbol{\mu}\rangle - \frac{1}{\eta}H(\boldsymbol{\mu}) \quad \text{s.t.} \quad \boldsymbol{\mu} \in \mathbb{L}_2^{\nu^{(k)}},$$

This observation follows because $\boldsymbol{\mu}^{(k)}$ is in $\mathbb{L}_2^{\nu^{(k)}}$ and its elements can be written as in (13) and (14), thus satisfying dual feasibility.

Recall that after every iteration all the component distributions are normalized. Recall that

$$\langle \eta C, \boldsymbol{\mu}^{(k)}(2)\rangle - H(\boldsymbol{\mu}^{(k)}(2)) = \mathcal{D}_\Phi\left(\boldsymbol{\mu}^{(k)}(2), \exp(-\eta C)\right) + \langle \mathbb{1}, e^{-\eta C}\rangle$$
$$\langle \eta C, \boldsymbol{\mu}^*_\eta\rangle - H(\boldsymbol{\mu}^*_\eta) = \mathcal{D}_\Phi\left(\boldsymbol{\mu}^*_\eta, \exp(-\eta C)\right) + \langle \mathbb{1}, e^{-\eta C}\rangle,$$

where $\Phi = -H$ is the negative entropy. The point $\boldsymbol{\mu}^*_\eta$ is the optimal point of the information projection $\exp(-\eta C)$ for points in $\mathbb{L}_2$. By the properties of information projections,

$$\mathcal{D}_\Phi\left(\boldsymbol{\mu}^{(k)}(2), \exp(-\eta C)\right) \geq \mathcal{D}_\Phi\left(\boldsymbol{\mu}^{(k)}(2), \boldsymbol{\mu}^*_\eta\right) + \mathcal{D}_\Phi\left(\boldsymbol{\mu}^*_\eta, \exp(-\eta C)\right).$$

Since for $\boldsymbol{\mu}^{(k)}(2)$ and $\boldsymbol{\mu}^*_\eta$, the sum of their entries is the same, by Pinsker's inequality (applied to each of the component vertex and edge distributions) this in turn implies that

$$\begin{aligned}\mathcal{D}_\Phi\left(\boldsymbol{\mu}^{(k)}(2), \exp(-\eta C)\right) - \mathcal{D}_\Phi\left(\boldsymbol{\mu}^*_\eta, \exp(-\eta C)\right) &\geq \mathcal{D}_\Phi\left(\boldsymbol{\mu}^{(k)}(2), \boldsymbol{\mu}^*_\eta\right) \\ &\geq \sum_{i\in\mathcal{V}} \frac{1}{2}\left\|\left(\boldsymbol{\mu}^{(k)}(2)\right)_i - \left(\boldsymbol{\mu}^*_\eta\right)_i\right\|_1^2 + \qquad (17)\\ &\quad \sum_{ij\in\mathcal{E}} \frac{1}{2}\left\|\left(\boldsymbol{\mu}^{(k)}(2)\right)_{ij} - \left(\boldsymbol{\mu}^*_\eta\right)_{ij}\right\|_1^2. \qquad (18)\end{aligned}$$

Let $\boldsymbol{\mu}^*_\eta(\nu^{(k)})$ in $\mathbb{L}_2^{\nu^{(k)}}$ be the vector produced by Lemma 8 applied to $\boldsymbol{\mu}^*_\eta \in \mathbb{L}_2$. Note that we utilize the existence of $\boldsymbol{\mu}^*_\eta(\nu^{(k)})$ and $\boldsymbol{\mu}^{(k)}(2)$ for analysis but we need not actually *compute* them. Expanding I yields

$$\begin{aligned}\mathcal{D}_\Phi\left(\boldsymbol{\mu}^{(k)}(2), \exp(-\eta C)\right) - \mathcal{D}_\Phi\left(\boldsymbol{\mu}^*_\eta, \exp(-\eta C)\right) &= \langle \eta C, \boldsymbol{\mu}^{(k)}(2) - \boldsymbol{\mu}^*_\eta\rangle + H(\boldsymbol{\mu}^*_\eta) - H(\boldsymbol{\mu}^{(k)}(2)) \\ &= \underbrace{\langle \eta C, \boldsymbol{\mu}^{(k)}(2) - \boldsymbol{\mu}^{(k)}\rangle + H(\boldsymbol{\mu}^{(k)}) - H(\boldsymbol{\mu}^{(k)}(2))}_{A_1} \\ &\quad + \underbrace{\langle \eta C, \boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^*_\eta(\nu^{(k)})\rangle + H(\boldsymbol{\mu}^*_\eta(\nu^{(k)})) - H(\boldsymbol{\mu}^{(k)})}_{A_2} \\ &\quad + \underbrace{\langle \eta C, \boldsymbol{\mu}^*_\eta(\nu^{(k)}) - \boldsymbol{\mu}^*_\eta\rangle + H(\boldsymbol{\mu}^*_\eta) - H(\boldsymbol{\mu}^*_\eta(\nu^{(k)}))}_{A_3}.\end{aligned}$$

Term A_2 is negative since $\boldsymbol{\mu}^{(k)}$ is the optimal point in the slack polytope. Because $\boldsymbol{\mu}^*_\eta(\nu^{(k)})$ and $\boldsymbol{\mu}^{(k)}(2)$ were constructed such that all their probabilities are lower bounded by τ, it holds that the entropies are $\log\frac{1}{\tau}$-Lipschitz

in $\|\cdot\|_1$ Terms A_1 and A_3 can be then bounded:

$$\begin{aligned}
A_1 &\leq \eta\|C\|_\infty \|\boldsymbol{\mu}^{(k)}(2) - \boldsymbol{\mu}^{(k)}\|_1 + \log\frac{1}{\tau}\|\boldsymbol{\mu}^{(k)}(2) - \boldsymbol{\mu}^{(k)}\|_1 \\
&\leq (\eta\|C\|_\infty + \log 1/\tau)\left(2\|\nu^{(k)}\|_1 + 2(|\mathcal{E}|+n)d^2\tau\right) \\
A_3 &\leq \|C\|_\infty \|\boldsymbol{\mu}_\eta^*(\nu^{(k)}) - \boldsymbol{\mu}_\eta^*\|_1 + \log\frac{1}{\tau}\|\boldsymbol{\mu}_\eta^*(\nu^{(k)}) - \boldsymbol{\mu}_\eta^*\|_1 \\
&\leq (\eta\|C\|_\infty + \log 1/\tau)\left(6d\deg(\mathcal{G})\|\nu^{(k)}\|_1 + 8(|\mathcal{E}|+n)d^2\tau\right).
\end{aligned}$$

The result then follow as

$$A_1 + A_3 \leq (\eta\|C\|_\infty + \log 1/\tau)\left(8d\deg(\mathcal{G})\|\nu^{(k)}\|_1 + 10(|\mathcal{E}|+n)d^2\tau\right).$$

□

Theorem 4, combined with the EMP algorithm's optimality condition can provide convergence guarantees for the case when $\mathbb{L}_2$ is tight and the solution is unique. We restate the main result, Theorem 3, for readability.

Theorem 5. *Let* $\eta \geq \frac{2\log(16n^2d^2)+16|\mathcal{E}|d^2}{\min(\Delta,\frac{1}{128})}$, *and* $\epsilon^{-1} > (25d\deg(\mathcal{G})|\mathcal{E}|)^2\max(\eta\|C\|_\infty, 68)$. *If* $\mathbb{L}_2$ *is tight and* $|\mathcal{V}_2^*| = 1$, *the EMP algorithm returns a MAP assignment after* $\lceil\frac{4\mathcal{S}_0(\deg(\mathcal{G}_k)+1)}{\epsilon^2}\rceil$ *iterations for EMP-cyclic and after* $\lceil\frac{4\mathcal{S}_0}{\epsilon^2}\rceil$ *iterations for EMP-greedy.*

Proof. Let $\boldsymbol{\mu}^{(k)}$ be the last internal iterate of the EMP algorithm before rounding. Since the stopping condition has been met, the slack vector $\nu^{(k)}$ corresponding to $\boldsymbol{\mu}^{(k)}$ must satisfy $\|\left(\nu^{(k)}\right)_{ij}\|_1 \leq \epsilon$ for all $ij \in \mathcal{E}$ so that $\|\nu^{(k)}\|_1 \leq 2|\mathcal{E}|\epsilon$.

Let $\boldsymbol{\mu}^{(k)}(2)$ be defined as in Theorem 4 and choose $\tau = \frac{\epsilon}{10(|\mathcal{E}|+n)d^2}$[5]. Then, the bound from Theorem 4 becomes

$$\begin{aligned}
&(\eta\|C\|_\infty + \log 1/\tau)\left(8d\deg(\mathcal{G})\|\nu^{(k)}\|_1 + 10(|\mathcal{E}|+n)d^2\tau\right) \\
&\leq (\eta\|C\|_\infty + \log 1/\tau)\left(16d\deg(\mathcal{G})|\mathcal{E}|\epsilon + 10(|\mathcal{E}|+n)d^2\tau\right) \\
&= \left(\eta\|C\|_\infty + \log\frac{10(|\mathcal{E}|+n)d^2}{\epsilon}\right)17d\deg(\mathcal{G})|\mathcal{E}|\epsilon \\
&= \left(\eta\|C\|_\infty + \log\left(10(|\mathcal{E}|+n)d^2\right) + \log\frac{1}{\epsilon}\right)17d\deg(\mathcal{G})|\mathcal{E}|\epsilon \\
&\leq \left(\eta\|C\|_\infty + \log\left(10(|\mathcal{E}|+n)d^2\right) + 2\epsilon^{-1/2}\right)17d\deg(\mathcal{G})|\mathcal{E}|\epsilon,
\end{aligned}$$

where the last inequality used the fact that $\log(x) \leq n(x^{1/n}-1)$ for $n > 0$. Choosing $\epsilon^{-1} > 425d^2\deg(\mathcal{G})^2|\mathcal{E}|^2\max\{\eta\|C\|_\infty, 68\}$ ensures that

$$\sum_{i\in\mathcal{V}}\frac{1}{2}\left\|\left(\boldsymbol{\mu}^{(k)}(2)\right)_i - \left(\boldsymbol{\mu}_\eta^*\right)_i\right\|_1^2 + \sum_{ij\in\mathcal{E}}\frac{1}{2}\left\|\left(\boldsymbol{\mu}^{(k)}(2)\right)_{ij} - \left(\boldsymbol{\mu}_\eta^*\right)_{ij}\right\|_1^2 \leq \frac{3}{25}.$$

Consequently for all $i \in \mathcal{V}$

$$\left\|\left(\boldsymbol{\mu}^{(k)}(2)\right)_i - \left(\boldsymbol{\mu}_\eta^*\right)_i\right\|_1 \leq \frac{2}{5}.$$

and for all $ij \in \mathcal{E}$

$$\left\|\left(\boldsymbol{\mu}^{(k)}(2)\right)_{ij} - \left(\boldsymbol{\mu}_\eta^*\right)_{ij}\right\|_1 \leq \frac{2}{5}.$$

[5] As long as $\epsilon \leq \frac{1}{4d}$ at least, this guarantees $\tau \leq \frac{1}{8d^2}$, so we are free to use Theorem 4

We also have

$$\begin{aligned}\|\boldsymbol{\mu}^{(k)}(2) - \boldsymbol{\mu}^{(k)}\|_1 &\leq 2\|\nu^{(k)}\|_1 + 2(|\mathcal{E}| + n)d^2\tau \\ &\leq 4|\mathcal{E}|\epsilon + \frac{\epsilon}{5} \\ &\leq 5|\mathcal{E}|\epsilon,\end{aligned}$$

which implies $\|\boldsymbol{\mu}^{(k)}(2) - \boldsymbol{\mu}^{(k)}\|_1 \leq \frac{1}{24}$ and $\|\boldsymbol{\mu}^*_\eta - \boldsymbol{\mu}^*\|_1 \leq \frac{1}{32}$ (by the condition on η, see Theorem 1). Putting these inequalities together by triangle inequality,

$$\begin{aligned}\| \left(\boldsymbol{\mu}^{(k)}\right)_i - (\boldsymbol{\mu}^*)_i \|_1 &\leq \| \left(\boldsymbol{\mu}^{(k)}\right)_i - \left(\boldsymbol{\mu}^{(k)}(2)\right)_i \|_1 + \| \left(\boldsymbol{\mu}^{(k)}(2)\right)_i - \left(\boldsymbol{\mu}^*_\eta\right)_i \|_1 + \| \left(\boldsymbol{\mu}^*_\eta\right)_i - (\boldsymbol{\mu}^*)_i \|_1 \\ &\leq \frac{1}{24} + \frac{2}{5} + \frac{1}{32} \\ &< \frac{1}{2}.\end{aligned}$$

For all $i \in \mathcal{V}$. A similar statement holds for all $ij \in \mathcal{E}$:

$$\begin{aligned}\| \left(\boldsymbol{\mu}^{(k)}\right)_{ij} - (\boldsymbol{\mu}^*)_{ij} \|_1 &\leq \| \left(\boldsymbol{\mu}^{(k)}\right)_{ij} - \left(\boldsymbol{\mu}^{(k)}(2)\right)_{ij} \|_1 + \| \left(\boldsymbol{\mu}^{(k)}(2)\right)_{ij} - \left(\boldsymbol{\mu}^*_\eta\right)_{ij} \|_1 + \| \left(\boldsymbol{\mu}^*_\eta\right)_{ij} - (\boldsymbol{\mu}^*)_{ij} \|_1 \\ &\leq \frac{1}{24} + \frac{2}{5} + \frac{1}{32} \\ &< \frac{1}{2}.\end{aligned}$$

Therefore, assuming $\boldsymbol{\mu}^*$ (the solution of $\mathbb{L}_2$) is integral,

$$\left(\text{round}(\boldsymbol{\mu}^{(k)})\right)_i = (\boldsymbol{\mu}^*)_i \text{ for all } i \in \mathcal{V}$$

and

$$\left(\text{round}(\boldsymbol{\mu}^{(k)})\right)_{ij} = (\boldsymbol{\mu}^*)_{ij} \text{ for all } ij \in \mathcal{E}.$$

□

E Experiment Details

In this section, we provide some additional details for the experiments in Section 7. As mentioned, empirical comparisons between state-of-the-art solvers and EMP-like algorithms have been studied extensively (Kappes et al., 2013; Meshi et al., 2012; Ravikumar et al., 2010; Werner, 2007). For instance, Meshi et al. (2012) found that the regularized star-based message passing algorithms greatly outperform standard optimization techniques such as FISTA and gradient descent, which do not exploit the coordinate structure of the problem.

The primary purpose of these experiments is to understand how the theoretical results in Section 6 manifest in a practical setting. In particular, we would like to understand how the convergence rates, in terms of the ability to round to the solution, behave as a function of the parameters of the problem such as graph size, choice of regularization η, and connectivity of the graph. In all experiments, we ran an LP solver on the graph in order to obtain the ground-truth MAP assignment. We only considered problems that were tight. The solver specifically is the ECOS solver through a CVXPY wrapper.

E.1 Grid Experiments

As mentioned, our first set of experiments considered solving the MAP problem on $\sqrt{n} \times \sqrt{n}$ grids, totallying n vertices. The vertices were connected by edges to their vertical and horizontal neighbors in the grid. This setting is fairly standard in the literature (Erdogdu et al., 2017; Globerson and Jaakkola, 2008; Ravikumar et al., 2010).

We considered the MAP problem with $d = 3$ labels and choose a cost vector C in the family of multi-label Potts models, another well-studied application (Wainwright and Jordan, 2008). Potts models typically have diagonal

potentials between edges. That is, we only penalize/reward when the labels on two connected vertices agree. We randomly generated the actual values of the vector. For vertex costs, we chose $C_i(x_i) \sim \text{Unif}(-0.5, 0.5)$ and for the edge costs we chose

$$C_{ij}(x_i, x_j) = \begin{cases} \beta_{ij} & x_i = x_j \\ 0 & \text{otherwise} \end{cases} \quad \forall ij, x_i, x_j,$$

where $\beta_{ij} \sim \text{Unif}\{-0.1, 0.1\}$. In the approximation results, we ran the algorithms until they had effectively converged after 80 iterations, where each iteration consisted of a full pass over the edges. For EMP-cyclic, this means simply going through all the edges once. For EMP-greedy, one iteration means the opportunity to update each edge exactly once, (although the algorithm will greedily select them in reality). Thus both algorithms update the same number of edges, though their choices will be different. Regardless, we found that 80 iterations was reasonably sufficient to observe the approximation properties. We measured the results in terms of the average Hamming distance between the LP's solution, which is integral, and the rounded solution returned by the algorithms.

E.2 Random Graph Experiments

While the grid topology offers a consistent platform to evaluate the algorithms, we also considered randomly generated graphs, specifically Erdős-Rényi random graphs. These graphs are constructed by iterating through every pair of the n vertices. Then, an edge is drawn between vertex i and j with probability p. Specifically, we chose $p = \frac{1.1 \log n}{n}$, which is just large enough that the graph is almost surely connected. We found these to be useful hyperparameter because any lower and the graph would largely be disconnected. Any higher and typically we found the LP was not tight. We chose the same multi-label Potts model for generating the cost vector C.

With these experiments, we intended to understand how diverse graph topologies would affect convergence due to randomness. In particular, we restricted the degrees of the graph to $\deg(\mathcal{G}) = 5, 10$ to observe how the algorithms behave on denser graphs.