## A. Numerical Experiments Details

All experiments were run on an Intel Core i7 processor with 16 GB RAM. We consider two different metrics.

1. The first metric is the original primal objective value (P), which is the actual value we wish to minimize. Since the optimization is done in the dual variables, we use Proj to project the point $\mu^{\boldsymbol{\lambda}}$ onto $\mathbb{L}_2$ and report the projection's function value, additively normalized with respect to the optimal value.

2. The second metric emphasizes the first by reporting the log-competitive ratio between the standard and accelerated variants of the the algorithms. The competitive ratio is computed as $\log\left(\frac{\langle C, \widehat{\mu}_{\mathsf{EMP}} - \mu^* \rangle}{\langle C, \widehat{\mu}_{\mathsf{Accel\text{-}EMP}} - \mu^* \rangle}\right)$, where $\widehat{\mu}_{\mathsf{EMP}}$ and $\widehat{\mu}_{\mathsf{Accel\text{-}EMP}}$ are the projections due to Proj of the outputs of EMP and Accel-EMP, respectively, and $\mu^*$ is a minimizer over $\mathbb{L}_2$. The same is computed for SMP and Accel-SMP. Thus, positive values at a given time indicate that the accelerated variant has lower error on the original objective.

**Message Passing Algorithms** We implemented the message passing algorithms with their update rules exactly as prescribed in Algorithms 1, 2, and 3. The algorithms are compared with respect to the number of updates (i.e. iterations); however, we note that the cost of each update is greater for SMP and Accel-SMP since they both require computing slacks of the entire neighborhood surrounding a give vertex.

**Block-Coordinate Methods** In addition to studying the empirical properties of the message passing algorithms, we present a supplementary empirical comparison with block-coordinate descent and its accelerated variant (Lee & Sidford, 2013). The purpose of this inclusion is to standardize how much we expect acceleration to improve the algorithms. We chose a stepsize of $1/\eta$. We note that each update in block-coordinate descent is essentially as expensive as an update of EMP.

This choice of cost vectors $C$ ensures that vertices cannot be trivially set to their minimal vertex potentials to achieve a reasonable objective value; the MAP algorithm must actually consider pairwise interactions. We evaluated each of the four algorithms on the same graph with $\eta = 1000$. Due to the inherent stochasticity of the randomized algorithms, we ran each one 10 times and took the averages and standard deviations. Since the graphs are small enough, we computed the ground-truth optimal value of (P) using a standard LP solver in CVXPY.

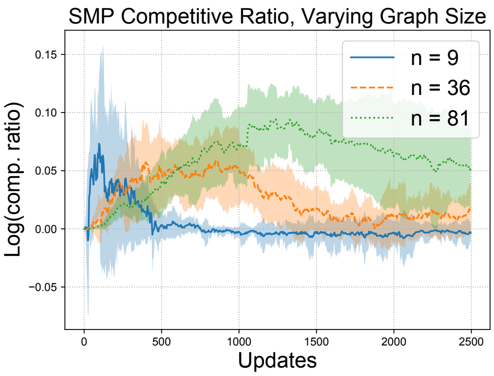

*Figure 2.* The competitive ratio of SMP with respect to Accel-SMP on (P) is compared across random graphs of varying sizes $n = 9$, 36, and 81.

In order to understand the effect of the graph size on the competitive ratio between the standard and accelerated algorithms, we generated random graphs of sizes $n = 9$, 36, and 81 with the same randomly drawn edges and cost vectors. We ran SMP and Accel-SMP for a fixed number of iterations over 10 random trials and computed the average log competitive ratio. Again, we used $\eta = 1000$. Figure 2 shows that Accel-SMP runs faster in all regimes, especially at beginning, and then the performance improvement eventually tapers after many iterations.

## B. Omitted Proofs and Derivations of Section 2

### B.1. Proof of Proposition 1

Recall that the primal objective is to solve the following:

$$\text{minimize} \quad \langle C, \mu \rangle - \frac{1}{\eta} H(\mu) \quad \text{s.t.} \quad \mu \in \mathbb{L}_2, \tag{obj}$$

where

$$H(\mu) = -\sum_{i \in V} \sum_{x_i \in \chi} \mu_i(x_i)(\log \mu_i(x_i) - 1) - \sum_{e \in E} \sum_{x_e \in \chi^2} \mu_e(x_e)(\log \mu_e(x_e) - 1).$$

Though it is not strictly necessary, we will also include a normalization constraint on the pseudo-marginal edges, which amounts to $\sum_{x_e} \mu_e(x_e) = 1$ for all $e \in E$. The Lagrangian is therefore

$$\mathcal{L}(\mu, \boldsymbol{\lambda}, \xi) = \langle C, \mu \rangle - \frac{1}{\eta} H(\mu) + \sum_{e \in E, i \in e} \boldsymbol{\lambda}_{e,i}^\top (S_{e,i} - \mu_i) + \sum_{i \in V} \xi_i (\sum_{x_i} \mu_i(x_i) - 1) + \sum_{e \in E} \xi_e (\sum_{x_e} \mu_e(x_e) - 1)$$

Taking the derivative w.r.t $\mu$ yields

$$\frac{\partial \mathcal{L}(\mu, \boldsymbol{\lambda}, \xi)}{\partial \mu_i(x_i)} = C_i(x_i) + \frac{1}{\eta} \log \mu_i(x_i) + \xi_i - \sum_{e \in N_i} \boldsymbol{\lambda}_{e,i}(x_i) \tag{6}$$

$$\frac{\partial \mathcal{L}(\mu, \boldsymbol{\lambda}, \xi)}{\partial \mu_e(x_e)} = C_e(x_e) + \frac{1}{\eta} \log \mu_e(x_e) + \xi_e + \sum_{i \in e} \boldsymbol{\lambda}_{e,i}((x_e)_i). \tag{7}$$

Here we are using $(x_e)_i$ to denote selecting the label associated with endpoint $i \in V$ from the pair of labels $x_e \in \chi^2$. The necessary conditions for optimality imply that

$$\mu_i^{\boldsymbol{\lambda},\xi}(x_i) = \exp\left( -\eta C_i(x_i) - \eta \xi_i + \eta \sum_{e \in N_i} \boldsymbol{\lambda}_{e,i}(x_i) \right) \tag{8}$$

$$\mu_e^{\boldsymbol{\lambda},\xi}(x_e) = \exp\left( -\eta C_e(x_e) - \eta \xi_e - \eta \sum_{i \in e} \boldsymbol{\lambda}_{e,i}((x_e)_i) \right), \tag{9}$$

where we use the superscripts to show that these optimal values are dependent on the dual variables, $\boldsymbol{\lambda}$ and $\xi$. The dual problem then becomes

$$\underset{\boldsymbol{\lambda}, \xi}{\text{maximize}} \quad -\frac{1}{\eta} \sum_i \sum_{x_i} \mu_i^{\lambda,\xi}(x_i) - \frac{1}{\eta} \sum_c \sum_{x_c} \mu_c^{\boldsymbol{\lambda},\xi}(x_c) - \sum_{i \in V} \xi_i - \sum_{e \in E} \xi_e$$

Note that we can solve exactly for $\xi$ as well, which simply normalizes the individual pseudo-marginals for each edge and vertex so that

$$\xi_i = \frac{1}{\eta} \log \sum_{x_i} \exp\left( -\eta C_i(x_i) - \eta \xi_i + \eta \sum_{e \in N_i} \boldsymbol{\lambda}_{e,i}(x_i) \right)$$

$$\xi_e = \frac{1}{\eta} \log \sum_{x_e} \exp\left( -\eta C_e(x_e) - \eta \xi_e - \eta \sum_{i \in e} \lambda_{e,i}((x_e)_i) \right)$$

Plugging this into $\mu^{\boldsymbol{\lambda},\xi}$ ensures that each local vertex and edge distribution is normalized to 1. Therefore the final objective becomes

$$\begin{aligned} \underset{\boldsymbol{\lambda}}{\text{minimize}} \quad & \frac{m+n}{\eta} + \frac{1}{\eta} \sum_{i \in V} \log \sum_{x_i} \exp\left( -\eta C_i(x_i) - \eta \xi_i + \eta \sum_{e \in N_i} \boldsymbol{\lambda}_{e,i}(x_i) \right) \\ & + \frac{1}{\eta} \sum_{e \in E} \log \sum_{x_e} \exp\left( -\eta C_e(x_e) - \eta \xi_e - \eta \sum_{i \in e} \lambda_{e,i}((x_e)_i) \right), \end{aligned}$$

and we can ignore the constant.

### B.2. Entropy-Regularized Message Passing Derivations

In this section, we derive the standard message passing algorithms that will be the main focus of the paper. Both come from simply computing the gradient and choosing additive updates to satisfy the optimality conditions directly.

**Proposition 2.** *The operator* $\mathsf{EMP}^\eta_{e,i} : \boldsymbol{\lambda} \mapsto \boldsymbol{\lambda}'_{e,i}(\cdot) \in \mathbb{R}^d$ *is satisfied by* $\boldsymbol{\lambda}'_{e,i}(x_i) = \boldsymbol{\lambda}_{e,i}(x_i) + \frac{1}{2\eta} \log \frac{S^{\boldsymbol{\lambda}}_{e,i}(x_i)}{\mu^{\boldsymbol{\lambda}}_i(x_i)}$.

*Proof.* From (1), the partial gradient of $L$ with respect to coordinate $(e, i, x_i)$ yields the following necessary and sufficient optimality condition:

$$S^{\boldsymbol{\lambda}}_{e,i}(x_i) = \mu^{\boldsymbol{\lambda}}_i(x_i).$$

Suppose that $\boldsymbol{\lambda}'$ satisfies this condition, and thus minimizes $L_{e,i}(\cdot; \boldsymbol{\lambda})$. We can decompose $\boldsymbol{\lambda}'$ at coordinate $(e, i, x_i)$ additively as $\boldsymbol{\lambda}'_{e,i}(x_i) = \boldsymbol{\lambda}_{e,i}(x_i) + \delta_{e,i}(x_i)$. From the definition of $\mu^{\boldsymbol{\lambda}}$, the optimality condition becomes

$$\exp(2\eta\delta_{c,i}(x_i)) = \frac{\sum_{x_j \in \chi} \mu^{\boldsymbol{\lambda}}_e(x_i, x_j)}{\mu^{\boldsymbol{\lambda}}_i(x_i)}$$

Rearranging to find $\delta_{e,i}(x_i)$ and then substituting into $\boldsymbol{\lambda}'_{e,i}(x_i)$ yields the desired result. □

Now, we can derive a lower bound on the improvement on the dual objective $L$ from applying an update of $\mathsf{EMP}$.

**Lemma 1.** *Let $\boldsymbol{\lambda}'$ be the result of applying $\mathsf{EMP}^{\eta}_{e,i}(\boldsymbol{\lambda})$ to $\boldsymbol{\lambda}$, keeping all other coordinates fixed. Then,* $L(\boldsymbol{\lambda}) - L(\boldsymbol{\lambda}') \geq \frac{1}{4\eta} \|\nu^{\boldsymbol{\lambda}}_{e,i}\|_1^2$.

*Proof.* Let $\widetilde{\mu}$ denote the unnormalized marginals. From the definition of $L$,

$$\begin{aligned} L(\boldsymbol{\lambda}) - L(\boldsymbol{\lambda}') = &\frac{1}{\eta} \log \sum_{x_i} \exp\left(-\eta C_i(x_i) + \eta \sum_{e \in N_i} \boldsymbol{\lambda}_{e,i}(x_i)\right) + \frac{1}{\eta} \log \sum_{x_e} \exp\left(-\eta C_e(x_e) - \eta \sum_{i \in e} \boldsymbol{\lambda}_{e,i}(x_i)\right) \\ &- \frac{1}{\eta} \log \sum_{x_i} \exp\left(-\eta C_i(x_i) + \eta\delta_{e,i}(x_i) + \eta \sum_{e \in N_i} \boldsymbol{\lambda}_{e,i}(x_i)\right) \\ &- \frac{1}{\eta} \log \sum_{x_e} \exp\left(-\eta C_e(x_e) - \eta\delta_{e,i}(x_i) - \eta \sum_{i \in e} \boldsymbol{\lambda}_{e,i}(x_i)\right) \end{aligned}$$

Define $\widetilde{\mu}^{\boldsymbol{\lambda}}_i(x_i) = \exp\left(-\eta C_i(x_i) + \sum_{e \in E} \boldsymbol{\lambda}_{e,i}(x_i)\right)$ and $\widetilde{\mu}_e(x_e) = \exp\left(-\eta C_e(x_e) - \sum_{i \in e} \boldsymbol{\lambda}_{e,i}(x_i)\right)$. The cost difference can then be written as

$$\begin{aligned} L(\boldsymbol{\lambda}) - L(\boldsymbol{\lambda}') &= -\frac{1}{\eta} \log\left(\sum_{x_i} \frac{\widetilde{\mu}^{\boldsymbol{\lambda}}_i(x_i) e^{\eta\delta_{e,i}(x_i)}}{\sum_{x'_i} \widetilde{\mu}^{\boldsymbol{\lambda}}_i(x'_i)}\right) - \frac{1}{\eta} \log\left(\sum_{x_e} \frac{\widetilde{\mu}^{\boldsymbol{\lambda}}_e(x_e) e^{-\eta\delta_{e,i}(x_c)}}{\sum_{x'_e} \widetilde{\mu}^{\boldsymbol{\lambda}}_e(x'_e)}\right) \\ &= -\frac{1}{\eta} \log\left(\sum_{x_i} \mu^{\boldsymbol{\lambda}}_i(x_i) \sqrt{\frac{S_{e,i}(x_i)}{\mu^{\boldsymbol{\lambda}}_i(x_i)}}\right) - \frac{1}{\eta} \log\left(\sum_{x_e} \mu^{\boldsymbol{\lambda}}_e(x_e) \sqrt{\frac{\mu^{\boldsymbol{\lambda}}_i(x_i)}{S^{\boldsymbol{\lambda}}_{e,i}(x_i)}}\right) \\ &= -\frac{2}{\eta} \log\left(\sum_{x_i} \sqrt{S^{\boldsymbol{\lambda}}_{e,i}(x_i) \mu^{\boldsymbol{\lambda}}_i(x_i)}\right) \end{aligned}$$

Note that the right-hand contains the Bhattacharyya coefficient $BC(p, q) := \sum_i \sqrt{p_i q_i}$ which has the following relationship with the Hellinger distance: $BC(p, q) = 1 - h^2(p, q)$.

The inequality then follows from exponential inequalities:

$$L(\boldsymbol{\lambda}) - L(\boldsymbol{\lambda}') = -\frac{2}{\eta} \log(1 - h^2(S^{\boldsymbol{\lambda}}_{e,i}, \mu^{\boldsymbol{\lambda}}_i)) \geq -\frac{2}{\eta} \log \exp(-h^2(S^{\boldsymbol{\lambda}}_{e,i}, \mu^{\boldsymbol{\lambda}}_i)) = \frac{2}{\eta} h^2(S^{\boldsymbol{\lambda}}_{e,i}, \mu^{\boldsymbol{\lambda}}_i)$$

Furthermore, the Hellinger inequality gives us

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

We conclude the result by applying this inequality with $p = S^{\boldsymbol{\lambda}}_{e,i}$ and $q = \mu^{\boldsymbol{\lambda}}_i$. □

**Proposition 3.** *The operator* $\mathsf{SMP}_i^\eta : \boldsymbol{\lambda} \mapsto \boldsymbol{\lambda}'_{\cdot,i}(\cdot) \in \mathbb{R}^{d|N_i|}$ *is, for all* $e \in N_i$ *and* $x_i \in \chi$*, satisfied by*

$$\begin{aligned}\boldsymbol{\lambda}'_{e,i}(x_i) = \boldsymbol{\lambda}_{e,i} &+ \frac{1}{\eta}\log S^{\boldsymbol{\lambda}}_{e,i}(x_i) \\ &- \frac{1}{\eta(|N_i|+1)}\log\left(\mu_i^{\boldsymbol{\lambda}}(x_i)\textstyle\prod_{e'\in N_i} S^{\boldsymbol{\lambda}}_{e',i}(x_i)\right),\end{aligned}$$

*Proof.* The optimality conditions require, for all $e \in N_i$,

$$S^{\boldsymbol{\lambda}'}_{e,i}(x_i) = \mu_i^{\boldsymbol{\lambda}'}(x_i),$$

which implies that

$$S^{\boldsymbol{\lambda}}_{e,i}(x_i) = \mu_i^{\boldsymbol{\lambda}}(x_i)\exp\left(\eta\delta_{e,i}(x_i) + \eta\sum_{e'\in N_i}\delta_{e',i}(x_i)\right),$$

where $\delta_{e,i}(x_i) = \boldsymbol{\lambda}'_{e,i}(x_i) - \boldsymbol{\lambda}_{e,i}(x_i)$. Then, let $e_1, e_2 \in N_i$. At optimality, it holds that

$$\frac{S^{\boldsymbol{\lambda}}_{e_1,i}(x_i)}{S^{\boldsymbol{\lambda}}_{e_2,i}(x_i)} = \frac{\exp(\eta\delta_{e_1,i}(x_i))}{\exp(\eta\delta_{e_2,i}(x_i))}$$

Substituting each $\delta_{e_2,i}(x_i)$ in terms of $\delta_{e_1,i}(x_i)$, we then have

$$S^{\boldsymbol{\lambda}}_{e,i}(x_i) = \mu_i^{\boldsymbol{\lambda}}(x_i)\exp(\eta\delta_{e,i}(x_i))\prod_{e'\in N_i}\frac{S^{\boldsymbol{\lambda}}_{e',i}(x_i)}{S^{\boldsymbol{\lambda}}_{e,i}(x_i)}\exp(\eta\delta_{e,i}(x_i)).$$

Collecting and then rearranging the above results in

$$\exp((|N_i|+1)\eta\delta_{e,i}(x_i)) = (S^{\boldsymbol{\lambda}}_{e,i}(x_i))^{|N_i|+1}\left(\mu_i^{\boldsymbol{\lambda}}(x_i)\prod_{e'\in N_i}S^{\boldsymbol{\lambda}}_{e',i}(x_i)\right)^{-1}.$$

In additive form, the update equation is

$$\delta_{e,i}(x_i) = \frac{1}{\eta}\log S^{\boldsymbol{\lambda}}_{e,i}(x_i) - \frac{1}{\eta(|N_i|+1)}\log\left(\mu_i^{\boldsymbol{\lambda}}(x_i)\prod_{e'\in N_i}S^{\boldsymbol{\lambda}}_{e',i}(x_i)\right).$$

□

## C. Omitted Proofs of Technical Lemmas of Section 4

### C.1. Proof of Random Estimate Sequences Lemma 3

**Lemma 3.** *The sequence* $\{\phi_k, \delta_k\}_{k=0}^K$ *defined in (5) is a random estimate sequence. Furthermore, it maintains the form* $\phi_k(\boldsymbol{\lambda}) = \omega_k + \frac{\gamma_k}{2}\|\boldsymbol{\lambda} - \mathbf{v}^{(k)}\|$ *for all k where*

$$\begin{aligned}\gamma_{k+1} &= (1-\theta_k)\gamma_k \\ \mathbf{v}^{(k+1)}_{e,i} &= \begin{cases}\mathbf{v}^{(k)}_{e,i} + \frac{q\theta_k}{\gamma_{k+1}}\nu^{\mathbf{y}^{(k)}}_{e,i} & \textit{if } (e,i) = (e_k, i_k) \\ \mathbf{v}^{(k)}_{e,i} & \text{otherwise}\end{cases} \\ \omega_{k+1} &= (1-\theta_k)\omega_k + \theta_k L(\mathbf{y}^{(k)}) - \frac{(\theta_k q)^2}{2\gamma_{k+1}}\|\nu^{\mathbf{y}^{(k)}}_{e_k,i_k}\|_2^2 \\ &\quad - \theta_k q\langle\nu^{\mathbf{y}^{(k)}}_{e_k,i_k}, \mathbf{v}^{(k)}_{e_k,i_k} - \mathbf{y}^{(k)}_{e_k,i_k}\rangle\end{aligned}$$

*Proof.* First we show that it is an estimate sequence by induction. Clearly this holds for the base case $\phi_0$ when $\delta_0 = 1$. Then, we assume the inductive hypothesis that $\mathbb{E}[\phi_k(\boldsymbol{\lambda})] \leq (1 - \delta_k)L(\boldsymbol{\lambda}) + \delta_k \phi_0(\boldsymbol{\lambda})$. From, here we can show

$$\begin{aligned}
\mathbb{E}[\phi_{k+1}(\boldsymbol{\lambda})] &= (1-\theta_k)\mathbb{E}[\phi_k(\boldsymbol{\lambda})] + \theta_k \mathbb{E}\left[L(\mathbf{y}^{(k)}) - \langle q\nu^{\mathbf{y}^{(k)}}_{e_k,i_k}, \boldsymbol{\lambda}_{e_k,i_k} - \mathbf{y}^{(k)}_{e_k,i_k}\rangle\right] \\
&= (1-\theta_k)\mathbb{E}[\phi_k(\boldsymbol{\lambda})] + \theta_k \mathbb{E}\left[L(\mathbf{y}^{(k)}) + \langle \nabla L(\mathbf{y}^{(k)}), \boldsymbol{\lambda} - \mathbf{y}^{(k)}\rangle\right] \\
&\leq (1-\theta_k)((1-\delta_k)L(\boldsymbol{\lambda}) + \delta_k\phi_0(\boldsymbol{\lambda})) + \theta_k L(\boldsymbol{\lambda}) \\
&= (1-\delta_{k+1})L(\boldsymbol{\lambda}) + \delta_{k+1}\phi_0(\boldsymbol{\lambda})
\end{aligned}$$

The first line uses the definition of $\phi_{k+1}$ and the second line uses the law of total expectation and the fact that $(e_k, i_k)$ is sampled uniformly. The inequality leverages the inductive hypothesis and convexity of $L$. From Nesterov (2018, §2), we know that the definition of $\delta_k$ from $\theta_k$ ensures that $\delta_k \xrightarrow{k} 0$. Therefore, $\{\phi_k, \delta_k\}_{k=0}^K$ is a random estimate sequence.

As noted, the identities are fairly standard (Nesterov, 2018; Lee & Sidford, 2013). We prove each claim in order.

- From definition of $\phi_{k+1}$, computing the second derivative of the combination shows that it is constant at $(1 - \theta_k)\gamma_k$.
- Computing the gradient with respect to block-coordinate $\boldsymbol{\lambda}_{e_k,i_k}$ of the combination shows, at optimality, we have

  $$(1-\theta_k)\gamma_k(\boldsymbol{\lambda}_{e_k,i_k} - \mathbf{v}^{(k)}_{e_k,i_k}) - q\theta_k\nu^{\mathbf{y}^{(k)}}_{e_k,i_k} = 0$$

  which implies

  $$\boldsymbol{\lambda}_{e_k,i_k} = \mathbf{v}^{(k)}_{e_k,i_k} + \frac{q\theta_k}{\gamma_{k+1}}\nu^{\mathbf{y}^{(k)}}_{e_k,i_k}$$

  For any other block-coordinate $(e, i)$, the optimality condition simply implies $\boldsymbol{\lambda}_{e,i} = \mathbf{v}^{(k)}_{e,i}$.
- The last claim can be show by inserting the minimizer, $\mathbf{v}^{(k)}_{e_k,i_k}$, into $\phi_{k+1}$. Therefore, we have

  $$\begin{aligned}
  \omega_{k+1} &:= \min_{\boldsymbol{\lambda}} \phi_{k+1} \\
  &= \phi_{k+1}(\mathbf{v}^{(k+1)}) \\
  &= (1-\theta_k)\omega_k + \frac{\gamma_{k+1}}{2}\|\mathbf{v}^{(k)} - \mathbf{v}^{(k+1)}\|_2^2 + \theta_k L(\mathbf{y}^{(k)}) - \theta_k q\langle \nu^{\mathbf{y}^{(k)}}_{e_k,i_k}, \mathbf{v}^{(k+1)}_{e_k,i_k} - \mathbf{y}^{(k)}_{e_k,i_k}\rangle \\
  &= (1-\theta_k)\omega_k + \frac{(q\theta_k)^2}{2\gamma_{k+1}}\|\nu^{\mathbf{y}^{(k)}}_{e_k,i_k}\|_2^2 + \theta_k L(\mathbf{y}^{(k)}) - \theta_k q\langle \nu^{\mathbf{y}^{(k)}}_{e_k,i_k}, \mathbf{v}^{(k)}_{e_k,i_k} + \frac{q\theta_k}{\gamma_{k+1}}\nu^{\mathbf{y}^{(k)}}_{e_k,i_k} - \mathbf{y}^{(k)}_{e_k,i_k}\rangle \\
  &= (1-\theta_k)\omega_k + \theta_k L(\mathbf{y}^{(k)}) - \frac{(q\theta_k)^2}{2\gamma_{k+1}}\|\nu^{\mathbf{y}^{(k)}}_{e_k,i_k}\|_2^2 - q\theta_k\langle \nu^{\mathbf{y}^{(k)}}_{e_k,i_k}, \mathbf{v}^{(k)}_{e_k,i_k} - \mathbf{y}^{(k)}_{e_k,i_k}\rangle
  \end{aligned}$$

□

### C.2. Proof of $\mathbb{L}_2$ Projection Lemma 5 and $\mathbb{L}_2^{\boldsymbol{\lambda}}$ Projection Lemma 6

**Lemma 5.** *For $\boldsymbol{\lambda} \in \mathbb{R}^{r_D}$ and $\mu^{\boldsymbol{\lambda}} \in \mathbb{L}_2^{\nu^{\boldsymbol{\lambda}}}$, Algorithm 4 returns a point $\widehat{\mu} = \mathsf{Proj}(\mu^{\boldsymbol{\lambda}}, 0)$ such that $\widehat{\mu}_i = \mu_i^{\boldsymbol{\lambda}}$ for all $i \in V$ and*

$$\sum_{e \in E} \|\mu_e^{\boldsymbol{\lambda}} - \widehat{\mu}_e\|_1 \leq 2 \sum_{e \in E, i \in e} \|\nu_{e,i}^{\boldsymbol{\lambda}}\|_1.$$

*Proof.* Since $\nu = 0$, we know that $\mu_i^{\boldsymbol{\lambda}} + \nu_{e,i} = \mu_i^{\boldsymbol{\lambda}} \in \Delta_n$ for all $i \in V$ and $e \in N_i$. For any $(i, j) = e \in E$, $\mathsf{Proj}$ applies Algorithm 2 of Altschuler et al. (2017) to generate $\widehat{\mu}_e$ from $\mu_e^{\boldsymbol{\lambda}}$ with the following guarantee due to Altschuler et al. (2017, Lemma 7):

$$\|\widehat{\mu}_e - \mu_e^{\boldsymbol{\lambda}}\|_1 \leq 2\|S_{e,i}^{\boldsymbol{\lambda}} - \mu_i^{\boldsymbol{\lambda}}\|_1 + 2\|S_{e,j}^{\boldsymbol{\lambda}} - \mu_j^{\boldsymbol{\lambda}}\|_1$$

and $\widehat{\mu}_e \in \mathcal{U}_d(\mu_i^{\boldsymbol{\lambda}}, \mu_j^{\boldsymbol{\lambda}})$. Applying this guarantee for all edges in $E$ gives the result. □

**Lemma 6.** *Let $\mu \in \mathbb{L}_2$ and $\boldsymbol{\lambda} \in \mathbb{R}^{rD}$. Define $\delta = \max_{e\in E, i\in e} \|\nu^{\boldsymbol{\lambda}}_{e,i}\|_1$ There exists $\widehat{\mu}$ in the slack polytope $\mathbb{L}_2^{\nu^{\boldsymbol{\lambda}}}$ such that*

$$\|\mu - \widehat{\mu}\|_1 \leq 16(m+n)d\delta + 2 \sum_{e\in E, i\in e} \|\nu^{\boldsymbol{\lambda}}_{e,i}\|_1$$

*Proof.* For convenience, we just write $\nu$ for the slack, dropping the notational dependence on $\boldsymbol{\lambda}$. We will proceed with this proof by constructing such a $\widehat{\mu}$ in two cases. We would like to show that the edge marginals $\mu_e$ can be modified to give $\widehat{\mu} \in \mathbb{L}_2^{\nu}$. To do this, we aim to use Algorithm 4 to match $\widehat{\mu}_e$ to the modified marginals $\mu_i + \nu_{e,i}$ for every $e \in E$ and $i \in e$. As long as $\mu_i + \nu_{e,i} \in \Delta_d$, setting $\mu_i' = \mu_i$ and $\mu_e' = \mu_e$ and computing $\widehat{\mu} = \mathsf{Proj}(\widetilde{\mu}, \nu)$ would return $\widehat{\mu} \in \mathbb{L}_2^{\nu}$ that satisfies the condition by Lemma 5.

However, if $\mu_i + \nu_{e,i} \notin \Delta_d$, then $\exists\, x \in \chi$ such that $\mu_i(x) + \nu_{e,i}(x) \notin [0,1]$. Consider the case where $\delta \leq \frac{1}{2d}$. We aim to create a temporary marginal vector $\mu'$ which is made by modifying $\mu_i$ appropriately until the slack can be added to $\mu_i'$ while maintaining a valid distribution. To do this, we set $\mu_i'$ as the convex combination with the uniform distribution

$$\mu_i' = (1-\theta)\mu_i + \theta\, \mathrm{Unif}(\chi)$$

for some $\theta \in [0,1]$. Choosing $\theta = d\delta$ ensures that

$$\delta \leq \mu'(x) \leq 1 - \delta \quad \forall\, x \in \chi,$$

Furthermore, $\mu_i' \in \Delta_d$ because $\Delta_d$ is convex and we have

$$\begin{aligned}
\|\mu_i' - \mu_i\|_1 &= \sum_{x\in\chi} \delta|1 - d\mu_i(x)| \\
&\leq \sum_{x} \delta + \delta d\mu_i(x) \\
&= 2d\delta
\end{aligned}$$

Then we set $\mu_e' = \mu_e$ for all $e \in E$. Using Algorithm 4, we compute $\widehat{\mu} = \mathsf{Proj}(\mu', \nu) \in \mathbb{L}_2^{\nu}$. Together with Lemma 5, we have that

$$\begin{aligned}
\|\widehat{\mu} - \mu\|_1 &= \sum_{i\in V} \|\widehat{\mu}_i - \mu_i\|_1 + \sum_{e\in E} \|\widehat{\mu}_e - \mu_e\|_1 \\
&\leq 2nd\delta + 2 \sum_{e\in E, i\in e} \|\nu_{e,i}\|_1 + \|\mu_i' - \mu_i\|_1 \\
&\leq (n+8m)d\delta + 2 \sum_{e\in E, i\in e} \|\nu_{e,i}\|_1
\end{aligned}$$

On the other hand, consider the case where $\delta > \frac{1}{2d}$. Then instead we choose the temporary marginal vector as $\mu_i' = \mu_i^{\boldsymbol{\lambda}}$ for all $i \in V$ and $\mu_e' = \mu_e$ for all $e \in E$, which ensures that $\mu_i' + \nu_{e,i} \in \Delta_d$ by definition of $\nu$. We then compute $\widehat{\mu} = \mathsf{Proj}(\mu', \nu)$, which ensures

$$\begin{aligned}
\|\widehat{\mu} - \mu\|_1 &\leq \sum_{i\in V} \|\mu_i - \mu_i^{\boldsymbol{\lambda}}\|_1 + 2 \sum_{e\in E, i\in e} \|\mu_i - \mu_i^{\boldsymbol{\lambda}}\|_1 + \|\nu_{e,i}\|_1 \\
&\leq 2n + 8m + 2 \sum_{e\in E, i\in e} \|\nu_{e,i}\|_1 \\
&\leq 4nd\delta + 16md\delta + 2 \sum_{e\in E, i\in e} \|\nu_{e,i}\|_1
\end{aligned}$$

where the second inequality uses the fact that the $l_1$ distance is bounded by 2 and the last inequality uses the assumption that $\delta > \frac{1}{2d}$. We take the worst of these two cases for the final result. □

### C.3. Proof of Proposition 4

**Proposition 4.** *Let* $\mu^* \in \mathbb{L}_2$ *be optimal,* $\boldsymbol{\lambda} \in \mathbb{R}^{r_D}$*,* $\widehat{\mu} = \mathsf{Proj}(\mu^{\boldsymbol{\lambda}}, 0) \in \mathbb{L}_2$*, and* $\delta = \max_{e\in E, i \in e} \|\nu^{\boldsymbol{\lambda}}_{e,i}\|_1$*. The following inequality holds:*

$$\begin{aligned}\langle C, \widehat{\mu} - \mu^* \rangle &\leq 16(m+n)d\|C\|_\infty \delta \\ &+ 4\|C\|_\infty \sum_{e\in E, i\in e} \|\nu^{\boldsymbol{\lambda}}_{e,i}\|_1 + \frac{n\log d + 2m \log d}{\eta}.\end{aligned}$$

*Proof.* Consider $\mu^{\boldsymbol{\lambda}}$, which may not lie in $\mathbb{L}_2$. It does, however, lie within its own slack polytope $\mathbb{L}_2^{\nu^{\boldsymbol{\lambda}}}$ from Definition 1. Therefore, it can be seen that $\mu^{\boldsymbol{\lambda}}$ is a solution to

$$\text{minimize} \quad \langle C, \mu \rangle - \frac{1}{\eta} H(\mu) \quad \text{s.t.} \quad \mu \in \mathbb{L}_2^{\boldsymbol{\lambda}} \tag{10}$$

Then, consider the point $\widehat{\mu} = \mathsf{Proj}(\mu^{\boldsymbol{\lambda}}, 0) \in \mathbb{L}_2$. Let $\mu^* \in \arg\min_{\mu \in \mathbb{L}_2} \langle C, \mu \rangle$. We have

$$\begin{aligned}\langle C, \widehat{\mu} - \mu^* \rangle &= \langle C, \widehat{\mu} - \mu^{\boldsymbol{\lambda}} + \mu^{\boldsymbol{\lambda}} - \mu^* \rangle \\ &\leq \|C\|_\infty \|\widehat{\mu} - \mu^{\boldsymbol{\lambda}}\|_1 + \langle C, \mu^{\boldsymbol{\lambda}} - \mu^* \rangle.\end{aligned} \tag{11}$$

Note that the last term in the right-hand side can be written as

$$\begin{aligned}\langle C, \mu^{\boldsymbol{\lambda}} - \mu^* \rangle &= \langle C, \mu^{\boldsymbol{\lambda}} - \widehat{\mu}^* + \widehat{\mu}^* - \mu^* \rangle \\ &\leq \|C\|_\infty \|\widehat{\mu}^* - \mu^*\|_1 + \langle C, \mu^{\boldsymbol{\lambda}} - \widehat{\mu}^* \rangle,\end{aligned} \tag{12}$$

where $\widehat{\mu}^* \in \mathbb{L}_2^{\nu^{\boldsymbol{\lambda}}}$ is the existing vector from Lemma 6 using $\mu^* \in \mathbb{L}_2$ and slack from $\boldsymbol{\lambda}$. Because $\mu^{\boldsymbol{\lambda}}$ is the solution to (10), we further have

$$\begin{aligned}\langle C, \mu^{\boldsymbol{\lambda}} - \widehat{\mu}^* \rangle &\leq \frac{1}{\eta}(H(\mu^{\boldsymbol{\lambda}}) - H(\widehat{\mu}^*)) \\ &\leq \frac{n\log d + 2m\log d}{\eta}\end{aligned} \tag{13}$$

Combining inequalities (11), (12), and (13) shows that

$$\langle C, \widehat{\mu} - \mu^* \rangle \leq \|C\|_\infty (\|\widehat{\mu} - \mu^{\boldsymbol{\lambda}}\|_1 + \|\widehat{\mu}^* - \mu^*\|_1) + \frac{n\log d + 2m\log d}{\eta}$$

Using Lemma 5 and 6, we can further bound this as

$$\langle C, \widehat{\mu} - \mu^* \rangle \leq \|C\|_\infty \left(16(m+n)d\delta + \sum_{e,i} 4\|\nu^{\boldsymbol{\lambda}}_{e,i}\|_1\right) + \frac{n\log d + 2m\log d}{\eta}.$$

□

### C.4. Proof of $G(\eta)$ Upper Bound

In the proof of Lemma 4, we used the fact that the numerator of the final convergence rate can be bounded by $G(\eta)^2$. Here, we formally state this result and prove it.

**Lemma 7.** *It holds that*

$$4L(0) - 4L(\boldsymbol{\lambda}^*) + 16m^2\eta\|\boldsymbol{\lambda}^*\|_2^2 \leq G(\eta)^2,$$

*where* $G(\eta) := 24md(m+n)(\sqrt{\eta}\|C\|_\infty + \frac{\log d}{\sqrt{\eta}})$.

The proof requires bounding both $L(0) - L(\boldsymbol{\lambda}^*)$, which we have already done in Lemma 10, and bounding the norm $\|\boldsymbol{\lambda}^*\|_2^2$. We rely on the following result from Meshi et al. (2012).

**Lemma 8.** *There exists $\boldsymbol{\lambda}^* \in \Lambda^*$ such that*

$$\begin{aligned}\|\boldsymbol{\lambda}^*\|_1 &\leq 2d(n+m)\|C\|_\infty + \frac{4d(m+n)}{\eta}\log d \\ &\leq \frac{4d(m+n)}{\eta}(\eta\|C\|_\infty + \log d)\end{aligned}$$

*Proof.* Modifying Meshi et al. (2012, Supplement Lemma 1.2) for our definition of $H$ gives us

$$\|\boldsymbol{\lambda}^*\|_1 \leq 2d(L(0) - n - m - \langle C, \mu^{\boldsymbol{\lambda}^*}\rangle + \frac{1}{\eta}H(\mu^{\boldsymbol{\lambda}^*})).$$

Using Cauchy-Schwarz and maximizing over the entropy yields the result. □

*Proof of Lemma 7.* Using these results, we can prove the claim. We bound the square root of the numerator, multiplied by $\sqrt{\eta}$:

$$\begin{aligned}\sqrt{\eta\left(4L(0) - 4L(\boldsymbol{\lambda}^*) + 16m^2\eta\|\boldsymbol{\lambda}^*\|_2^2\right)} &\leq \sqrt{8(m+n)(\eta\|C\|_\infty + \log d) + 16m^2\left(4d(m+n)\left(\eta\|C\|_\infty + \log d\right)\right)^2} \\ &\leq \sqrt{8(m+n)(\eta\|C\|_\infty + \log d)} + 16md(m+n)(\eta\|C\|_\infty + \log d) \\ &\leq 24md(m+n)(\eta\|C\|_\infty + \log d)\end{aligned}$$

The first inequality used Lemma 10 and Lemma 8. The second inequality uses the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$. The last inequality uses the fact that the first term is greater than 1 under the assumption $d \geq 2$. Dividing through by $\sqrt{\eta}$ gives the result. □

## D. Proof of Theorem 1

We begin with a complete proof of Theorem 1 for EMP and then show how to modify it slightly for SMP. The result also builds on some of the same technical lemmas used in the proof of Theorem 2.

### D.1. Edge Message Passing

The cornerstone of the proof is showing that the expected slack norms can be bounded over iterations.

**Lemma 9.** *Let $L^* = \min_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda})$ and let $\widehat{\boldsymbol{\lambda}}$ be the output of Algorithm 1 after $K$ iterations with* EMP *and a uniform distribution. For any $e \in E$ and $i \in e$, the expected norm of the constraint violation in $\mathbb{L}_2$ is bounded as*

$$\mathbb{E}\sum_{e\in E, i\in e}\|S_{e,i}^{\widehat{\boldsymbol{\lambda}}} - \mu_i^{\widehat{\boldsymbol{\lambda}}}\|_1^2 \leq \frac{8m\eta(L(0) - L^*)}{K}$$

*Proof.* From Lemma 1, we have that the expected improvement is lower bounded at each iteration

$$\mathbb{E}\left[L(\boldsymbol{\lambda}^{(k)}) - L(\boldsymbol{\lambda}^{(k+1)})\right] \geq \frac{1}{4\eta}\mathbb{E}\left[\|\nabla_{e_k, i_k}L(\boldsymbol{\lambda}^{(k)})\|_1^2\right],$$

Then, using that $\nabla_{e,i} L(\boldsymbol{\lambda}) = \mu_i^{\boldsymbol{\lambda}} - S_{e,i}^{\boldsymbol{\lambda}}$, we apply the bound $k = 1, 2, \ldots, K$:

$$\begin{aligned} L(0) - L^* &\geq \frac{1}{4\eta} \sum_{k=0}^{K-1} \mathbb{E}\left[ \| S_{e_k, i_k}^{\boldsymbol{\lambda}^{(k)}} - \mu_{i_k}^{\boldsymbol{\lambda}^{(k)}} \|_1^2 \right] \\ &= \frac{1}{8m\eta} \sum_{k=0}^{K-1} \sum_{e \in E, i \in e} \mathbb{E}\left[ \| S_{e,i}^{\boldsymbol{\lambda}^{(k)}} - \mu_i^{\boldsymbol{\lambda}^{(k)}} \|_1^2 \right] \\ &\geq \frac{K}{8m\eta} \sum_{e \in E, i \in e} \mathbb{E}\left[ \| S_{e,i}^{\widehat{\boldsymbol{\lambda}}} - \mu_i^{\widehat{\boldsymbol{\lambda}}} \|_1^2 \right] \end{aligned}$$

The equality uses the law of total expectation, conditioning on $\boldsymbol{\lambda}^{(k)}$. The second inequality uses the fact that $\widehat{\boldsymbol{\lambda}}$ is chosen to minimize the sum of squared constraint violations.

□

Next, we provide a bound on the initial function value gap.

**Lemma 10.** *For $\boldsymbol{\lambda}^* \in \Lambda^*$ it holds that $L(0) - L(\boldsymbol{\lambda}^*) \leq 2(m+n)\|C\|_\infty + \frac{2}{\eta}(m+n) \log d$.*

*Proof.* We will bound both $L(0)$ and $L(\boldsymbol{\lambda}^*)$ individually. First, from the definition

$$\begin{aligned} L(0) &= \frac{1}{\eta} \sum_i \log \sum_{x_i} \exp(-\eta C_i(x_i)) + \frac{1}{\eta} \sum_e \log \sum_{x_e} \exp(-\eta C_e(x_e)) \\ &\leq \frac{1}{\eta} \sum_i \log \sum_{x_i} \exp(\eta \|C\|_\infty) + \frac{1}{\eta} \sum_e \log \sum_{x_e} \exp(\eta \|C\|_\infty) \\ &= (n+m)\|C\|_\infty + \frac{n}{\eta} \log d + \frac{2m}{\eta} \log d. \end{aligned}$$

For $L(\boldsymbol{\lambda}^*)$, we recognize that $L$ is simply the negative of the primal problem (Reg-P), shifted by a constant amount. In particular, we have

$$-L(\boldsymbol{\lambda}^*) - \frac{1}{\eta}(n+m) = \langle C, \mu^* \rangle - \frac{1}{\eta} H(\mu^*)$$

For some $\mu^*$ that solves (Reg-P). Note that $H$ is offset with a linear term (different from the usual definition of the entropy) that exactly cancels the $-\frac{1}{\eta}(n+m)$ on the left-hand side. We then conclude $-L(\boldsymbol{\lambda}^*) \leq (m+n)\|C\|_\infty$ by Cauchy-Schwarz. Summing these two gives the desired result. □

*Proof of Theorem 1 for* EMP. Fix $\epsilon' > 0$. Lemma 9 and Lemma 10 ensure that

$$\mathbb{E} \sum_{e \in E, i \in e} \| \nu_{e,i}^{\widehat{\boldsymbol{\lambda}}} \|_1^2 \leq (\epsilon')^2$$

after

$$K = \frac{16m(m+n)(\eta\|C\|_\infty + \log d)}{(\epsilon')^2} \tag{14}$$

iterations. By Jensen's inequality and recognizing that the norms are non-negative, this also implies that

$$\mathbb{E}[\| \nu_{e,i}^{\widehat{\boldsymbol{\lambda}}} \|_1] \leq \epsilon' \quad \forall e \in E, i \in e$$

after the same number of iterations.

Now, we use the upper bound due to the approximation from Proposition 4 and take the expectation, giving

$$\begin{aligned}\mathbb{E}\left[\langle C, \widehat{\mu} - \mu^*\rangle\right] \leq \|C\|_\infty \left(8m\epsilon' + 16(m+n)d\mathbb{E}[\delta]\right) \\ + \frac{n \log d + 2m \log d}{\eta},\end{aligned}$$

where

$$\mathbb{E}\left[\delta\right]^2 \leq \mathbb{E}[\delta^2] \leq \mathbb{E} \sum_{e \in E, i \in e} \|\nu_{e,i}^{\widehat{\boldsymbol{\lambda}}}\|_1^2 \leq (\epsilon')^2.$$

Therefore, the bound becomes

$$\begin{aligned}\mathbb{E}\left[\langle C, \widehat{\mu} - \mu^*\rangle\right] \leq 24(m+n)d\|C\|_\infty \epsilon' \\ + \frac{n \log d + 2m \log d}{\eta}\end{aligned}$$

We conclude the result from substituting into (14), using the definition of $\eta$ and choosing $\epsilon' = \frac{\epsilon}{48(m+n)d\|C\|_\infty}$ □

### D.2. Star Message Passing

The significant difference between the SMP proof and the EMP proof is that there is variable improvement at each update, dependent on the degree of the node being updated. Using the distribution from (3), we ensure that the improvement becomes uniform in expectation. This analysis is similar to weighting coordinates by their coordinate-wise smoothness coefficients in coordinate gradient algorithms (Nesterov, 2012).

A slight modification of the proof of (Meshi et al., 2012) is required the get the tighter $l_1$-norm lower bound.

**Lemma 2.** *Let $\boldsymbol{\lambda}'$ be the result of applying $\mathsf{SMP}_i^\eta$ to $\boldsymbol{\lambda}$, keeping all other coordinates fixed. Then,* $L(\boldsymbol{\lambda}) - L(\boldsymbol{\lambda}') \geq \frac{1}{8|N_i|\eta} \sum_{e \in N_i} \|\nu_{e,i}^{\boldsymbol{\lambda}}\|_1^2$.

*Proof.* Meshi et al. (2012) show that

$$L(\boldsymbol{\lambda}) - L(\boldsymbol{\lambda}') = -\frac{1}{\eta} \log \left( \sum_{x_i} \left( \mu_i^{\boldsymbol{\lambda}} \prod_{e \in N_i} S_{e,i}^{\boldsymbol{\lambda}}(x_i) \right)^{\frac{1}{|N_i|+1}} \right)^{|N_i|+1},$$

and further

$$|N_i| - |N_i| \left( \sum_{x_i} \left( \mu_i^{\boldsymbol{\lambda}} \prod_{e \in N_i} S_{e,i}^{\boldsymbol{\lambda}}(x_i) \right)^{\frac{1}{|N_i|+1}} \right)^{|N_i|+1} \geq \sum_{e \in N_i} \left( 1 - \left( \sum_{x_i} \sqrt{\mu_i^{\boldsymbol{\lambda}}(x_i) S_{e,i}^{\boldsymbol{\lambda}}(x_i)} \right)^2 \right)$$

We recognize the inner term of the square as the Bhattacharyya coefficient which satisfies $BC \in [0, 1]$. Therefore,

$$\begin{aligned}\sum_{e \in N_i} \left( 1 - \left( \sum_{x_i} \sqrt{\mu_i^{\boldsymbol{\lambda}}(x_i) S_{e,i}^{\boldsymbol{\lambda}}(x_i)} \right)^2 \right) &\geq \sum_{e \in N_i} \left( 1 - \left( \sum_{x_i} \sqrt{\mu_i^{\boldsymbol{\lambda}}(x_i) S_{e,i}^{\boldsymbol{\lambda}}(x_i)} \right) \right) \\ &= \sum_{e \in N_i} h^2(\mu_i^{\boldsymbol{\lambda}}, S_{e,i}^{\boldsymbol{\lambda}})\end{aligned}$$

Then,

$$\left( \sum_{x_i} \left( \mu_i^{\boldsymbol{\lambda}} \prod_{e \in N_i} S_{e,i}^{\boldsymbol{\lambda}}(x_i) \right)^{\frac{1}{|N_i|+1}} \right)^{|N_i|+1} \leq 1 - \frac{1}{N_i} \sum_{e \in N_i} h^2(\mu_i^{\boldsymbol{\lambda}}, S_{e,i}^{\boldsymbol{\lambda}})$$

Finally, we lower bound the original difference of values

$$\begin{aligned} L(\boldsymbol{\lambda}) - L(\boldsymbol{\lambda}') &\geq -\frac{1}{\eta}\log\left(1 - \frac{1}{N_i}\sum_{e\in N_i} h^2(\mu_i^{\boldsymbol{\lambda}}, S_{e,i}^{\boldsymbol{\lambda}})\right) \\ &\geq \frac{1}{N_i\eta}\sum_{e\in N_i} h^2(\mu_i^{\boldsymbol{\lambda}}, S_{e,i}^{\boldsymbol{\lambda}}) \\ &\geq \frac{1}{8N_i\eta}\sum_{e\in N_i}\|S_{e,i}^{\boldsymbol{\lambda}} - \mu_i^{\boldsymbol{\lambda}}\|_1^2 \end{aligned}$$

□

**Lemma 11.** *Let $L^* = \min_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda})$ and let $\widehat{\boldsymbol{\lambda}}$ be the output of Algorithm 1 after $K$ iterations with* SMP *and distribution (3). Define $N = \sum_{j\in V}|N_j|$. For any $e \in E$ and $i \in e$, the expected norm of the constraint violation in $\mathbb{L}_2$ is bounded as*

$$\mathbb{E}\sum_{e\in E, i\in e}\|S_{e,i}^{\widehat{\boldsymbol{\lambda}}} - \mu_i^{\widehat{\boldsymbol{\lambda}}}\|_1^2 \leq \frac{8N\eta(L(0) - L^*)}{K}$$

*Proof.* Lemma 2 gave us the following lower bound on the improvement:

$$\mathbb{E}\left[L(\boldsymbol{\lambda}^{(k)}) - L(\boldsymbol{\lambda}^{(k+1)})\right] \geq \mathbb{E}\left[\frac{1}{8|N_{i_k}|\eta}\sum_{e\in N_{i_k}}\|\nu_{e,i_k}^{\boldsymbol{\lambda}^{(k)}}\|_1^2\right]$$

Then, since $i_k$ is chosen with probability $p_i = \frac{|N_i|}{N}$, we can apply the bound for $k = 1, 2, \ldots, K$ and expand the expectations:

$$\begin{aligned} L(0) - L^* &\geq \sum_{k=0}^{K-1}\mathbb{E}\left[\frac{1}{8|N_{i_k}|\eta}\sum_{e\in N_{i_k}}\|\nu_{e,i_k}^{\boldsymbol{\lambda}^{(k)}}\|_1^2\right] \\ &= \frac{1}{8N\eta}\sum_{k=0}^{K-1}\mathbb{E}\sum_{e\in E, i\in e}\|\nu_{e,i}^{\boldsymbol{\lambda}^{(k)}}\|_1^2 \\ &\geq \frac{1}{8N\eta}\sum_{k=0}^{K-1}\sum_{e\in E, i\in e}\mathbb{E}\left[\|\nu_{e,i}^{\widehat{\boldsymbol{\lambda}}}\|_1^2\right] \end{aligned}$$

The equality uses the law of total expectation, conditioning on $\boldsymbol{\lambda}^{(k)}$. The second inequality uses the fact that $\widehat{\boldsymbol{\lambda}}$ is chosen to minimize the sum of squared constraint violations. □

The rest of the proof for SMP proceeds in an identical manner to the case for EMP; however, we simply replace the $8m\eta$ with $8N\eta$ everywhere. This stems from the fact that we can now guarantee

$$\mathbb{E}\sum_{e\in E, i\in e}\|S_{e,i}^{\widehat{\boldsymbol{\lambda}}} - \mu_i^{\widehat{\boldsymbol{\lambda}}}\|_1^2 \leq (\epsilon')^2$$

in $\frac{8N\eta(L(0)-L^*)}{(\epsilon')^2}$ iterations instead. We can then use the same upper bound from Lemma 10 and substitute in the same choices of $\epsilon'$ and $\eta$ as in EMP to get the result.

## E. Proof of Theorem 2 for SMP

The proof for SMP essentially follows the same structure, but it requires defining the estimate sequence in slightly different way. Define the probability distribution $\{p_i\}_{i\in V}$ over $V$ with $p_i = \frac{|N_i|}{\sum_{j\in V}|N_j|}$. We propose the candidate:

$$\begin{aligned} i_k &\sim \text{Cat}(V, \{p_i\}_{i\in V}) \\ \delta_{k+1} &= (1-\theta_k)\delta_k \\ \phi_{k+1}(\boldsymbol{\lambda}) &= (1-\theta_k)\phi_k(\boldsymbol{\lambda}) + \theta_k L(\mathbf{y}^{(k)}) - \frac{\theta_k}{p_{i_k}}\langle \nu_{\cdot,i_k}^{\mathbf{y}^{(k)}}, \boldsymbol{\lambda}_{\cdot,i_k} - \mathbf{y}_{\cdot,i_k}^{(k)}\rangle \end{aligned} \tag{15}$$

Then, we show that this is indeed an estimate sequence with a conducive structure.

**Lemma 12.** *The sequence* $\{\phi_k, \delta_k\}_{k=0}^K$ *defined in (15) is a random estimate sequence. Furthermore, it maintains the form* $\phi_k(\boldsymbol{\lambda}) = \omega_k + \frac{\gamma_k}{2}\|\boldsymbol{\lambda} - \mathbf{v}^{(k)}\|$ *for all* $k$ *where*

$$\gamma_{k+1} = (1-\theta_k)\gamma_k$$

$$\mathbf{v}_{\cdot,i}^{(k+1)} = \begin{cases} \mathbf{v}_{\cdot,i}^{(k)} + \frac{\theta_k}{p_i \gamma_{k+1}} \nu_{\cdot,i}^{\mathbf{y}^{(k)}} & \textit{if } i = i_k \\ \mathbf{v}_{\cdot,i}^{(k)} & \text{otherwise} \end{cases}$$

$$\omega_{k+1} = (1-\theta_k)\omega_k + \theta_k L(\mathbf{y}^{(k)}) - \frac{\theta_k^2}{2\gamma_{k+1} p_{i_k}^2}\|\nu_{\cdot,i_k}^{\mathbf{y}^{(k)}}\|_2^2 - \frac{\theta_k}{p_{i_k}}\langle \nu_{\cdot,i_k}^{\mathbf{y}^{(k)}}, \mathbf{v}_{\cdot,i_k}^{(k)} - \mathbf{y}_{\cdot,i_k}^{(k)}\rangle$$

*Proof.* To show that this is an estimate sequence, the proof is essentially identical to the EMP case. The only exception is that we take expectation over $V$ with distribution $\{p_i\}_{i\in V}$. However, this ensures that

$$\mathbb{E}[\frac{\theta_k}{p_{i_k}}\langle \nu_{\cdot,i_k}^{\mathbf{y}^{(k)}}, \boldsymbol{\lambda}_{\cdot,i_k} - \mathbf{y}_{\cdot,i_k}^{(k)}\rangle] = \theta_k \mathbb{E}[\langle \nabla L(\mathbf{y}^{(k)}), \boldsymbol{\lambda} - \mathbf{y}^{(k)}\rangle]$$

by the law of total expectation. So the the proof that this is an estimate sequence remains the same.

To show that it retains the desired quadratic structure, we again analyze all terms of interest

- $\gamma_{k+1}$ is identical to the EMP case so the result holds.
- Taking the gradient with respect to $\boldsymbol{\lambda}_{\cdot,i}$, we have that the optimality conditions, for $i = i_k$, are

  $$\gamma_{k+1}(\boldsymbol{\lambda}_{\cdot,i_k} - \mathbf{v}_{\cdot,i_k}^{(k)}) - \frac{\theta_k}{p_{i_k}}\nu_{\cdot,i_k}^{\mathbf{y}^{(k)}} = 0$$

  and, for all other $i$, they are

  $$\gamma_{k+1}(\boldsymbol{\lambda}_{\cdot,i_k} - \mathbf{v}_{\cdot,i_k}^{(k)}) = 0.$$

  These conditions imply the given construction for $\mathbf{v}^{(k+1)}$.
- We can then compute $\omega_{k+1}$ by plugging in the choice for $\mathbf{v}^{(k+1)}$ again:

$$\begin{aligned}
\omega_{k+1} &:= \min_{\boldsymbol{\lambda}} \phi_{k+1} \\
&= \phi_{k+1}(\mathbf{v}^{(k+1)}) \\
&= (1-\theta_k)\omega_k + \frac{\gamma_{k+1}}{2}\|\mathbf{v}^{(k)} - \mathbf{v}^{(k+1)}\|_2^2 + \theta_k L(\mathbf{y}^{(k)}) - \frac{\theta_k}{p_{i_k}}\langle \nu_{\cdot,i_k}^{\mathbf{y}^{(k)}}, \mathbf{v}_{\cdot,i_k}^{(k+1)} - \mathbf{y}_{\cdot,i_k}^{(k)}\rangle \\
&= (1-\theta_k)\omega_k + \frac{\theta_k^2}{2\gamma_{k+1}p_{i_k}^2}\|\nu_{\cdot,i_k}^{\mathbf{y}^{(k)}}\|_2^2 + \theta_k L(\mathbf{y}^{(k)}) - \frac{\theta_k}{p_{i_k}}\langle \nu_{\cdot,i_k}^{\mathbf{y}^{(k)}}, \mathbf{v}_{\cdot,i_k}^{(k)} + \frac{\theta_k}{\gamma_{k+1}p_{i_k}}\nu_{\cdot,i_k}^{\mathbf{y}^{(k)}} - \mathbf{y}_{\cdot,i_k}^{(k)}\rangle \\
&= (1-\theta_k)\omega_k + \theta_k L(\mathbf{y}^{(k)}) - \frac{\theta_k^2}{2\gamma_{k+1}p_{i_k}^2}\|\nu_{\cdot,i_k}^{\mathbf{y}^{(k)}}\|_2^2 - \frac{\theta_k}{p_{i_k}}\langle \nu_{\cdot,i_k}^{\mathbf{y}^{(k)}}, \mathbf{v}_{\cdot,i_k}^{(k)} - \mathbf{y}_{\cdot,i_k}^{(k)}\rangle
\end{aligned}$$

□

We now provide a faster convergence guarantee on the dual objective function for SMP which depends on $N = \sum_{j\in V}|N_j|$.

**Lemma 13.** *For the random estimate sequence in (15), let* $\{\boldsymbol{\lambda}^{(k)}\}_{k=0}^K$ *and* $\{\mathbf{y}^{(k)}\}_{k=0}^K$ *be defined as in Algorithm 3 with* $\boldsymbol{\lambda}^{(0)} = 0$*. Then, the dual objective error in expectation can be bounded as*

$$\mathbb{E}[L(\boldsymbol{\lambda}^{(k)}) - L(\boldsymbol{\lambda}^*)] \le \frac{G_{\mathit{SMP}}(\eta)^2}{(k+2)^2},$$

*where* $G_{\mathit{SMP}}(\eta) := 24Nd(m+n)(\sqrt{\eta}\|C\|_\infty + \frac{\log d}{\sqrt{\eta}})$ *and* $N = \sum_{j\in V}|N_j|$.

*Proof.* As in the EMP proof, it suffices to show that $\mathbb{E}[\omega_{k+1}] \geq \mathbb{E}[L(\boldsymbol{\lambda}^{(k+1)})]$ by induction. As before we have

$$\begin{aligned}\mathbb{E}[\omega_{k+1}] &\geq (1-\theta_k)\mathbb{E}[L(\boldsymbol{\lambda}^{(k)})] + \theta_k \mathbb{E}[L(\mathbf{y}^{(k)})] - \mathbb{E}\left[\frac{\theta_k^2}{2\gamma_{k+1} p_{i_k}^2}\|\nu_{\cdot,i_k}^{\mathbf{y}^{(k)}}\|_2^2 - \frac{\theta_k}{p_{i_k}}\langle \nu_{\cdot,i_k}^{y^{(k)}}, \mathbf{v}^{(k)} - \mathbf{y}^{(k)}\rangle\right] \\ &\geq \mathbb{E}\left[L(\mathbf{y}^{(k)}) - \frac{\theta_k^2}{2\gamma_{k+1} p_{i_k}^2}\|\nu_{\cdot,i_k}^{\mathbf{y}^{(k)}}\|_2^2\right] + (1-\theta_k)\mathbb{E}\left[\langle \nabla L(\mathbf{y}^{(k)}), \boldsymbol{\lambda}^{(k)} - \mathbf{y}^{(k)}\rangle\right] + \theta_k \mathbb{E}\left[\langle \nabla L(\mathbf{y}^{(k)}), \mathbf{v}^{(k)} - \mathbf{y}^{(k)}\rangle\right] \\ &= \mathbb{E}\left[L(\mathbf{y}^{(k)}) - \sum_{i \in V}\frac{\theta_k^2}{2\gamma_{k+1} p_i}\|\nu_{\cdot,i}^{\mathbf{y}^{(k)}}\|_2^2\right],\end{aligned}$$

where the last line comes from the definition of $\mathbf{y}^{(k)}$. Choosing $\theta_k$ such that $\theta_k^2 = \frac{\gamma_{k+1}\min_j |N_j|}{4\eta N^2}$ results in

$$\begin{aligned}\mathbb{E}[\omega_{k+1}] &\geq \mathbb{E}\left[L(\mathbf{y}^{(k)}) - \sum_{i \in V}\frac{1}{8N\eta}\|\nu_{\cdot,i}^{\mathbf{y}^{(k)}}\|_2^2\right] \\ &= \mathbb{E}\left[L(\mathbf{y}^{(k)}) - \frac{1}{8N\eta}\|\nabla L(\mathbf{y}^{(k)}\|_2^2\right]\end{aligned}$$

Recall, from the improvement in Lemma 2, we have

$$\begin{aligned}\mathbb{E}[L(\boldsymbol{\lambda}^{(k+1)})] &\leq \mathbb{E}[L(\mathbf{y}^{(k)})] - \mathbb{E}\left[\frac{1}{8|N_{i_k}|\eta}\|\nu_{\cdot,i_k}^{\mathbf{y}^{(k)}}\|_2^2\right] \\ &= \mathbb{E}[L(\mathbf{y}^{(k)})] - \mathbb{E}\left[\frac{1}{8N\eta}\|\nabla L(\mathbf{y}^{(k)})\|_2^2\right]\end{aligned}$$

Therefore, by this induction, the inequality $\mathbb{E}[L(\boldsymbol{\lambda}^{(k+1)})] \leq \mathbb{E}[\omega_{k+1}]$ holds for all $k$. Furthermore, by choosing $\gamma_0 = \frac{4N^2\eta}{\min_j |N_j|}$, we ensure that $\theta_k$ can be updated recursively as in Algorithm 3 and the update equation for $\mathbf{v}$ is simplified to

$$\mathbf{v}_{\cdot,i}^{(k+1)} = \begin{cases} \mathbf{v}_{\cdot,i}^{(k)} + \frac{\min_j |N_j|}{4p_{i_k}\theta_k \eta N}\nu_{\cdot,i}^{\mathbf{y}^{(k)}} & \text{if } i = i_k \\ \mathbf{v}_{\cdot,i}^{(k)} & \textit{otherwise} \end{cases}.$$

Using the property of randomized estimate sequences derived in Section 4, we can bound the expected error in the dual norm as

$$\begin{aligned}\mathbb{E}(L(\boldsymbol{\lambda}^{(k)})] - L^* &\leq \frac{4}{(k+2)^2}\left(L(0) - L^* + \frac{\gamma_0}{2}\|\boldsymbol{\lambda}\|_2^2\right) \\ &= \frac{4}{(k+2)^2}\left(L(0) - L^* + \frac{2N^2\eta}{\min_j |N_j|}\|\boldsymbol{\lambda}^*\|_2^2\right) \\ &\leq \frac{4}{(k+2)^2}\left(L(0) - L^* + 2N^2\eta\|\boldsymbol{\lambda}^*\|_2^2\right)\end{aligned}$$

The numerator can then be bounded in an identical manner to the EMP proof by replacing $4m^2$ with $2N^2$ in Lemma 7, instead yielding $G_{\mathsf{SMP}}(\eta) = 40md(m+n)(\sqrt{\eta}\|C\|_\infty + \frac{\log d}{\sqrt{\eta}})$, which is only different by a constant. We then have

$$\mathbb{E}(L(\boldsymbol{\lambda}^{(k)})] - L^* \leq \frac{G_{\mathsf{SMP}}(\eta)}{(k+2)^2}$$

□

With these tools, we are ready to present the proof of Theorem 2 for SMP.

*Proof of Theorem 2 for* SMP. . Let $\widehat{\boldsymbol{\lambda}}$ be the output from Algorithm 3 after $K$ iterations. From Lemma 2, we can lower bound the result in Lemma 13 with

$$\begin{aligned}\frac{1}{8\eta|N_i|}\mathbb{E}\left[\sum_{e\in N_i}\|\nu_{e,i}^{\widehat{\boldsymbol{\lambda}}}\|_1^2\right] &\leq \mathbb{E}[L(\widehat{\boldsymbol{\lambda}})] - L^* \\ &\leq \frac{G_{\mathsf{SMP}}(\eta)}{(K+2)^2}\end{aligned}$$

for all $i \in V$. This further implies that

$$\frac{1}{8\eta|N_i|}\mathbb{E}\left[\|\nu_{e,i}^{\widehat{\boldsymbol{\lambda}}}\|_1^2\right] \leq \frac{G_{\mathsf{SMP}}(\eta)}{(K+2)^2}$$

for all $e \in E$ and $i \in e$. Then, for $\epsilon' > 0$, we can ensure that

$$\begin{aligned}\mathbb{E}[\|\nu_{e,i}^{\widehat{\boldsymbol{\lambda}}}\|_1] &\leq |N_i|\epsilon' \\ \mathbb{E}\sum_{i\in V, e\in N_i}\|\nu_{\cdot,i}^{\widehat{\boldsymbol{\lambda}}}\|_1^2 &\leq N(\epsilon')^2\end{aligned}$$

in $K = \frac{\sqrt{8\eta}G(\eta)}{\epsilon'}$ iterations. Letting $\widehat{\mu} \in \mathbb{L}_2$ be the projected version of $\mu^{\widehat{\boldsymbol{\lambda}}}$,

$$\langle C, \widehat{\mu} - \mu^*\rangle \leq \|C\|_\infty\left(16(m+n)d\delta + \sum_{e,i}4\|\nu_{e,i}^{\widehat{\boldsymbol{\lambda}}}\|_1\right) + \frac{n\log d + 2m\log d}{\eta}.$$

Taking the expectation of both sides gives us

$$\mathbb{E}[\langle C, \widehat{\mu} - \mu^*\rangle] \leq \|C\|_\infty\left(16(m+n)d\mathbb{E}[\delta] + 4N\epsilon'\right) + \frac{n\log d + 2m\log d}{\eta},$$

where

$$\mathbb{E}\left[\delta\right]^2 \leq \mathbb{E}[\delta^2] \leq \mathbb{E}\sum_{e\in E, i\in e}\|\nu_{e,i}^{\widehat{\boldsymbol{\lambda}}}\|_1^2 \leq N(\epsilon')^2.$$

Then we can conclude

$$\begin{aligned}\mathbb{E}[\langle C, \widehat{\mu} - \mu^*\rangle] &\leq 16\sqrt{N}(m+n)d\|C\|_\infty\epsilon' + 4N\|C\|_\infty\epsilon' + \frac{n\log d + 2m\log d}{\eta} \\ &\leq 24\sqrt{N}(m+n)d\|C\|_\infty\epsilon' + \frac{n\log d + 2m\log d}{\eta}.\end{aligned}$$

The last inequality uses the fact that $N = 2m$. Therefore, $\widehat{\mu}$ is expected $\epsilon$-optimal with $\eta$ as defined in the statement and $\epsilon' = \frac{\epsilon}{48\sqrt{N}(m+n)d\|C\|_\infty}$. Substituting these values into $K$ and $G(\eta)$ yields the result. □

## F. Rounding to Integral Solutions Proofs

In this section, we prove the bound on the number of iterations sufficient to recover the MAP solution using Accel-EMP and rounding the output of the algorithm. We then compare with standard EMP.

### F.1. Approximation Error

Let $\mathcal{V}_2$ be the set of vertices of $\mathbb{L}_2$ and $\mathcal{V}_2^*$ be the set of optimal vertices with respect to $C$. Denote by $\Delta = \min_{V_1\in\mathcal{V}_2\setminus\mathcal{V}_2^*, V_2\in\mathcal{V}_2^*}\langle C, V_1 - V_2\rangle$ the suboptimality gap. Let $\mathcal{R}_1 = \max_{\mu\in\mathbb{L}_2}\|\mu\|_1$, and $\mathcal{R}_H = \max_{\mu,\mu'\in\mathbb{L}_2} H(\mu) - H(\mu')$. Define $\deg$ to be the maximum degree of the graph. The following holds:

**Theorem 4** (Theorem 1 of (Lee et al., 2020))**.** *If $\mathbb{L}_2$ is tight, $|\mathcal{V}_2^*| = 1$ and $\eta \geq \frac{2\mathcal{R}_1\log(64\mathcal{R}_1)+2\mathcal{R}_1+2\mathcal{R}_H}{\Delta}$, then $\|\mu_\eta^* - \mu^*\|_1 \leq \frac{1}{8}$ and therefore the rounded solution $\mathrm{round}(\mu_\eta^*)$ is a MAP assignment.*

### F.2. Estimation Error for Accelerated Message Passing

To bound the estimation error, we invoke the accelerated convergence guarantees presented in the previous section. In particular, we showed that

$$\mathbb{E}\left[\|\nu^{\widehat{\boldsymbol{\lambda}}}_{e,i}\|_1\right] \leq \epsilon' \quad \forall e \in E, i \in e$$

after $K = \frac{\sqrt{4\eta}G(\eta)}{\epsilon'}$ iterations for $\mathsf{Accel\text{-}EMP}$. Markov's inequality implies that with probability $1-\delta$, $\|\nu^{\widehat{\boldsymbol{\lambda}}}_{e,i}\|_1 \leq \frac{2m\epsilon'}{\delta} := \epsilon$ for all $e \in E$ and $i \in e$. From Theorem 3 of Lee et al. (2020), we require

$$\epsilon < O\left(d^{-2}m^{-2}\deg^{-2}\max(1, \eta\|C\|_\infty)^{-1}\right)$$

Furthermore, the theorem of the previous subsection implies we can set

$$\eta = \frac{16(m+n)(\log(m+n)+\log(d))}{\Delta}$$

Then, by setting

$$\epsilon' \leq O\left(d^{-2}m^{-4}\delta\deg^{-2}\max(1, \|C\|_\infty/\Delta)^{-1}(\log dm)^{-1}\right)$$

the condition is satisfied. Therefore, plugging into $\sqrt{4\eta}G(\eta)$ yields

$$\sqrt{4\eta}G(\eta) = O\left(\frac{dm^3\|C\|_\infty\log dm}{\Delta}\right)$$

which implies, with probability $1-\delta$,

$$K = O\left(\frac{d^3m^7\deg^2\|C\|^2_\infty\log^2 dm}{\delta\Delta}\right)$$

These conditions of $\epsilon'$ and $\eta$ guarantee that the $\mathrm{round}(\mu^{\widehat{\boldsymbol{\lambda}}})$ is the MAP solution by invoking Theorem 3 of Lee et al. (2020).

### F.3. Comparison to Standard Methods

Using standard $\mathsf{EMP}$, we require the same conditions be satisfied on $\epsilon'$ and $\eta$ to guarantee recover of the MAP solution. However, the rate of convergence differs, requiring $K = \frac{L(0)-L(\boldsymbol{\lambda}^*)}{(\epsilon')^2}$ iterations, as seen previously. Note that

$$L(0) - L(\boldsymbol{\lambda}^*) = O\left(\frac{m^3\|C\|_\infty\log dm}{\Delta}\right)$$

. Note that there is no additional $d$ dependence. It holds that with probability $1-\delta$, the MAP solution is recovered by $\mathsf{EMP}$ in at most

$$K = O\left(\frac{d^4m^{11}\deg^4\|C\|^3_\infty\log^3 dm}{\delta^2\Delta}\right)$$

iterations. We emphasize that this iteration bound is only a sufficient condition by directly applying the technique developed in this section. We suspect it can be greatly improved.