Accelerated Message Passing for Entropy-Regularized MAP Inference

Jonathan N. Lee [1] **Aldo Pacchiano** [2] **Peter Bartlett** [2 3] **Michael I. Jordan** [2 3]

Abstract

Maximum a posteriori (MAP) inference in discrete-valued Markov random fields is a fundamental problem in machine learning that involves identifying the most likely configuration of random variables given a distribution. Due to the difficulty of this combinatorial problem, linear programming (LP) relaxations are commonly used to derive specialized message passing algorithms that are often interpreted as coordinate descent on the dual LP. To achieve more desirable computational properties, a number of methods regularize the LP with an entropy term, leading to a class of smooth message passing algorithms with convergence guarantees. In this paper, we present randomized methods for accelerating these algorithms by leveraging techniques that underlie classical accelerated gradient methods. The proposed algorithms incorporate the familiar steps of standard smooth message passing algorithms, which can be viewed as coordinate minimization steps. We show that these accelerated variants achieve faster rates for finding ϵ-optimal points of the unregularized problem, and, when the LP is tight, we prove that the proposed algorithms recover the true MAP solution in fewer iterations than standard message passing algorithms.

1. Introduction

Discrete undirected graphical models are extensively used in machine learning since they provide a versatile and powerful way of modeling dependencies between variables (Wainwright & Jordan, 2008). In this work we focus on the important class of discrete-valued pairwise models. Efficient inference in these models has multiple applications, ranging from computer vision (Jegelka & Bilmes, 2011), to statistical physics (Mezard & Montanari, 2009), information theory (MacKay, 2003) and even genome research (Torada et al., 2019).

In this paper we study and propose efficient methods for maximum a posteriori (MAP) inference in pairwise, discrete-valued Markov random fields. The MAP problem corresponds to finding a configuration of all variables achieving a maximal probability and is a key problem that arises when using these undirected graphical models. There exists a vast literature on MAP inference spanning multiple communities, where it is known as constraint satisfaction (Schiex et al., 1995) and energy minimization (Kappes et al., 2013). Even in the binary case, the MAP problem is known to be NP-hard to compute exactly or even to approximate (Kolmogorov & Zabin, 2004; Cooper, 1990).

As a result, there has been much emphasis on devising methods that may work on settings under which the problem becomes tractable. A popular way to achieve this goal is to express the problem as an integer program and then relax this to a linear program (LP). If the LP constraints are set to the convex hull of marginals corresponding to all global settings, also known as the marginal polytope (Wainwright & Jordan, 2008), then the LP would yield the optimal integral solution to the MAP problem. Unfortunately, writing down this polytope would require exponentially many constraints and therefore it is not tractable. We can consider larger polytopes defined over subsets of the constraints required to define the marginal polytope. This is a popular approach that underpins the family of LP relaxations known as the Sherali-Adams (SA) hierarchy (Sherali & Adams, 1990). Instead of enforcing global consistency, we enforce only pairwise consistency via the local polytope, thus yielding pseudo-marginals that are pairwise consistent but may not correspond to any true global distribution. Despite the local polytope requiring a number of constraints that is linear in the number of edges of the input graph, the runtime required for solving this linear program for large graphs may be prohibitive in practice (Yanover et al., 2006). These limitations have motivated the design and theoretical analysis of message passing algorithms that exploit the structure of the problem. In this paper we study a class of smooth message passing algorithms, derived from a regularized version of the

[1]Department of Computer Science, Stanford University, USA [2]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, USA [3]Department of Statistics, University of California, Berkeley, USA. Correspondence to: Jonathan Lee <jnl@stanford.edu>, Aldo Pacchiano <pacchiano@berkeley.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

local polytope LP relaxation (Ravikumar et al., 2010; Meshi et al., 2012; Savchynskyy et al., 2011; Hazan & Shashua, 2008).

The technique of using entropy penalties to regularize linear programs has a long and successful history. It has been observed, both practically and in theory that, in some problems, solving a regularized linear program yields algorithms with computational characteristics that make them preferable to simply using an LP solver, particularly with large scale problems. Previous work has studied and analyzed convergence rates, and even rounding guarantees for simple message passing algorithms (Ravikumar et al., 2010; Lee et al., 2020; Meshi et al., 2012) based on iterative Bregman projections onto the constraints. These algorithms are sometimes described as being *smooth*, as the dual objective is smooth in the dual variable as a result of the entropy regularization. Inspired by accelerated methods in optimization (Lee & Sidford, 2013; Nesterov, 2012) we propose and analyze two new variants of accelerated message passing algorithms Accel-EMP and Accel-SMP. In this paper, we are able to show our methods drastically improve upon the convergence rate of previous message passing algorithms.

1.1. Related Work

MAP Inference The design and analysis of convergent message passing algorithms has attracted a great deal of attention over the years. Direct methods of deriving asymptotically convergent algorithms have been extensively explored. Examples include tree-reweighted message passing (Kolmogorov, 2006), max-sum diffusion (Werner, 2007), MP-LP (Globerson & Jaakkola, 2008), and other general block-coordinate ascent methods (Kappes et al., 2013; Sontag et al., 2011). Our work builds upon regularized inference problems that directly regularize the linear objective with strongly convex terms, often leading to "smooth" message passing (Savchynskyy et al., 2011; 2012; Hazan & Shashua, 2008; Weiss et al., 2007; Meshi et al., 2015). This formalization has led to a number of computationally fast algorithms, but often with asymptotic guarantees.

The focus of this paper is non-asymptotic convergence guarantees for families of these algorithms. Ravikumar et al. (2010) provided one of the first directions towards this goal leveraging results for proximal updates, but ultimately the rates were not made explicit due to the approximation at every update. Meshi et al. (2012) provided a comprehensive analysis of a message passing algorithm derived from the entropy-regularized objective and later a gradient-based one for the quadratically regularized objective (Meshi et al., 2015). However, the convergence rates were only given for the regularized objective, leaving the guarantee on the unregularized problem unknown. Furthermore, the objective studied by Meshi et al. (2015) did not ultimately yield a message passing (or coordinate minimization) algorithm, which could be more desirable from a computational standpoint. Lee et al. (2020) provided rounding guarantees for a related message passing scheme derived from the entropy regularized objective, but did not consider convergence on the unregularized problem either. Savchynskyy et al. (2011) studied the direct application of the acceleration methods of Nesterov (2018); however, this method also forwent the message passing scheme and convergence on the unregularized problem was only shown asymptotically. Jojic et al. (2010) gave similar results for a dual decomposition method that individually smooths subproblems. Acceleration was applied to get fast convergence but on the dual problem.

In addition to problem-specific message passing algorithms, there are numerous general purpose solvers that can be applied to MAP inference to solve the LP with strong theoretical guarantees. Notably, interior point methods (Karmarkar, 1984; Renegar, 1988; Gondzio, 2012) offer a promising alternative for faster algorithms. For example recent work provides a $\widetilde{O}(\sqrt{\mathrm{rank}})$ iteration complexity by Lee & Sidford (2014), where rank is the rank of the constraint matrix. In this paper, we only consider comparisons between message passing algorithms; however, it would be interesting to compare both empirical and theoretical differences between message passing and interior point methods in the future.

Accelerating Entropy-Regularized Linear Programs We also highlight a connection with similar problems in other fields. Notably, optimal transport also admits an LP form and has seen a surge of interest recently. As in MAP inference, these approximations are conducive to empirically fast algorithms, such as the celebrated *Sinkhorn* algorithm, that outperforms generic solvers (Cuturi, 2013; Benamou et al., 2015; Genevay et al., 2016). In theory, Altschuler et al. (2017) showed convergence guarantees for Sinkhorn and noted that it can be viewed as a block-coordinate descent algorithm on the dual, similar to the MAP problem. Since this work, several methods have striven to obtain faster rates (Lin et al., 2019; Dvurechensky et al., 2018), which can be viewed as building on the seminal acceleration results of Nesterov (2018) for general convex functions. It is interesting to note that the entropy-regularized objectives in optimal transport and MAP inference effectively become softmax minimization problems, which have also been studied generally in the context of smooth approximations (Nesterov, 2005) and maximum flow (Sidford & Tian, 2018).

1.2. Contributions

For the case of MAP inference from entropy-regularized objectives, we address the question: is it possible to directly accelerate message passing algorithms with faster non-asymptotic convergence and improved rounding guarantees? We answer this question affirmatively from a theo-

retical standpoint. We propose a method to directly accelerate standard message passing schemes, inspired by Nesterov. We prove a convergence guarantee for standard schemes on the unregularized MAP objective over $\mathbb{L}_2$, showing convergence on the order of $\widetilde{O}(m^5/\epsilon^3)$ iterations where m is the number of edges, assuming the number of vertices and labels and the potential functions are fixed. We then prove that the accelerated variants converge in expectation on the order of $\widetilde{O}(m^{9/2}/\epsilon^2)$ iterations. We conclude by showing that the accelerated variants recover the true MAP solution with high probability in fewer iterations compared to prior message passing analyses when the LP relaxation is tight and the solution is unique (Lee et al., 2020).

1.3. Notation

Let $\mathbb{R}_+$ denote the set of non-negative reals. The d-dimensional probability simplex over the finite set χ is $\Sigma^d := \left\{p \in \mathbb{R}_+^d \ : \ \sum_{x\in\chi} p(x) = 1\right\}$. A joint distribution, $P \in \Sigma^{d\times d}$, is indexed by $x_c = (x_p, x_q) \in \chi^2$. The transportation polytope of $p, q \in \Sigma^d$ is defined as the set of pairwise joint distributions that marginalize to p and q, written as $\mathcal{U}_d(p,q) = \{P \in \mathbb{R}_+^{d\times d} \ : \ \sum_{x_p} P(x_p, x) = q(x),\ \sum_{x_q} P(x, x_q) = p(x)\}$. For any vector $p \in \mathbb{R}_+^d$, we write the entropy as $H(p) := -\sum_x p(x)(\log p(x) - 1)$. While this is a somewhat unusual definition, it simplifies terms later and has been used by (Benamou et al., 2015; Lee et al., 2020). We will use $\langle\cdot,\cdot\rangle$ generally to mean the sum of the elementwise products between two equal-length indexable vectors. For any two vectors $p, q \in \mathbb{R}_+^d$, the Hellinger distance is $h(p,q) := \frac{1}{\sqrt{2}}\|\sqrt{p} - \sqrt{q}\|_2$. The vector $\mathbf{1}_d \in \mathbb{R}^d$ consists of all ones.

2. Coordinate Methods for MAP Inference

2.1. Smooth Approximations

For the pairwise undirected graph $G = (V, E)$ with $n := |V|$ and $m := |E|$[1], we associate each vertex $i \in V$ with the random variable X_i on the finite set of labels χ of size $d = |\chi| \geq 2$ and a distribution that factorizes as $p(x_V) = \frac{1}{Z(\phi)} \prod_{e\in E} \phi_e(x_e) \prod_{i\in V} \phi_i(x_i)$ where $\phi_i \in \mathbb{R}^d$ and $\phi_e \in \mathbb{R}^{d^2}$. For any edge $e \in E$, we use $i \in e$ to denote that i is one of the two endpoints of e. For $i \in V$, we define $N_i := \{e \in E \ : \ i \in e\}$, as the set of all incident edges to i. We assume that each vertex has at least one edge. MAP inference refers to the combinatorial problem of identifying the configuration that maximizes this probability distribution. In our case, we cast it as the following minimization problem:

$$\underset{x_V \in \chi^{|V|}}{\text{minimize}} \quad \textstyle\sum_{i\in V} C_i(x_i) + \sum_{e\in E} C_e(x_e), \qquad \text{(MAP)}$$

where $C = -\log\phi$, i.e., we view C as indexable by vertices, edges, and their corresponding labels. It can be shown (Wainwright & Jordan, 2008) that (MAP) is equivalent to the following linear program:

$$\min_{\mu\in\mathcal{M}} \langle C, \mu\rangle \quad \text{s.t.} \quad \mu \in \mathcal{M},$$

where $\mu \in \mathbb{R}^{r_P}$ for $r_P = nd + md^2$ is known as a marginal vector, and $\langle C, \mu\rangle = \sum_{i\in V}\sum_{x_i\in\chi} C_i(x_i)\mu_i(x_i) + \sum_{e\in E}\sum_{x_e\in\chi^2} C_e(x_e)\mu_e(x_e)$, and $\mathcal{M}$ is the marginal polytope defined as

$$\mathcal{M} := \left\{\mu \ : \ \exists\, \mathbb{P} \ \text{ s.t. } \begin{array}{l} \mathbb{P}_{X_i}(x_i) = \mu_i(x_i),\ \forall i, x_i \\ \mathbb{P}_{X_iX_j}(x_e) = \mu_e(e),\ \forall e, x_e \end{array}\right\}.$$

Here, $\mathbb{P}$ is any valid distribution over the random variables $\{X_i\}_{i\in V}$. Since $\mathcal{M}$ is described by exponentially many constraints in the graph size, outer-polytope relaxations are a standard paradigm to approximate the above problem by searching instead over the local polytope:

$$\mathbb{L}_2 := \left\{\mu \ : \ \begin{array}{ll} \mu_i \in \Sigma^d & \forall i \in V \\ \mu_e \in \mathcal{U}_d(\mu_i, \mu_j) & \forall e = ij \in E \end{array}\right\}.$$

The local polytope $\mathbb{L}_2$ enforces only pairwise consistency between variables while $\mathcal{M}$ requires the marginal vector to be generated from a globally consistent distribution of $\{X_i\}_{i\in V}$, so that $\mathcal{M} \subseteq \mathbb{L}_2$. We refer the reader to the survey of Wainwright & Jordan (2008, §3) for details. Thus, our primary objective throughout the paper will be finding solutions to the approximate problem

$$\text{minimize} \quad \langle C, \mu\rangle \quad \text{s.t.} \quad \mu \in \mathbb{L}_2. \qquad \text{(P)}$$

Let $\epsilon > 0$. We say that a point $\widehat{\mu} \in \mathbb{L}_2$ is ϵ-optimal for (P) if it satisfies $\langle C, \widehat{\mu}\rangle \leq \min_{\mu\in\mathbb{L}_2}\langle C, \mu\rangle + \epsilon$. For a random $\widehat{\mu}$, we say that it is expected ϵ-optimal if

$$\mathbb{E}\left[\langle C, \widehat{\mu}\rangle\right] \leq \min_{\mu\in\mathbb{L}_2}\langle C, \mu\rangle + \epsilon.$$

Despite the simple form of the linear program, it has been observed to be difficult to solve in practice for large graphs even with state-of-the-art solvers (Yanover et al., 2006), motivating researchers to study an approximate version with entropy regularization:

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

where $\eta \in \mathbb{R}_+$ controls the level of regularization. Intuitively, the regularization encourages μ_i and μ_e to be closer to the uniform distribution for all vertices and edges.

[1] In this paper we study pairwise models with only vertices and edges, implying only pairwise interaction between variables. However, our results can be extended to more general graphs.

The dual problem takes on the succinct form of an unconstrained log-sum-exp optimization problem. Thus, when combined, the local polytope relaxation and entropy-regularizer result in a smooth approximation.

Proposition 1. *The dual objective of (Reg-P) can be written as*

$$\underset{\boldsymbol{\lambda}}{\text{minimize}} \quad L(\boldsymbol{\lambda}), \tag{Reg-D}$$

where L is defined as

$$L(\boldsymbol{\lambda}) = \frac{1}{\eta}\sum_{i\in V}\log\sum_{x\in\chi}\exp\left(-\eta C_i(x) + \textstyle\sum_{e\in N_i}\boldsymbol{\lambda}_{e,i}(x)\right) + \frac{1}{\eta}\sum_{e\in E}\log\sum_{x\in\chi^2}\exp\left(-\eta C_e(x) - \textstyle\sum_{i\in e}\boldsymbol{\lambda}_{e,i}(x_i)\right).$$

Furthermore, primal variables can be recovered directly by

$$\mu_i^{\boldsymbol{\lambda}}(x_i) \propto \exp\left(-\eta C_i(x_i) + \eta\textstyle\sum_{e\in N_i}\boldsymbol{\lambda}_{e,i}(x_i)\right)$$
$$\mu_e^{\boldsymbol{\lambda}}(x_e) \propto \exp\left(-\eta C_e(x_e) - \eta\textstyle\sum_{i\in e}\boldsymbol{\lambda}_{e,i}((x_e)_i)\right).$$

For convenience we let $r_D = 2md$ denote the dimension of the dual variables $\boldsymbol{\lambda} \in \mathbb{R}^{r_D}$. This is in contrast to the dimension r_P of the primal marginal vectors defined earlier. We use $\Lambda^* \subseteq \mathbb{R}^{r_D}$ to denote the set of solutions to (Reg-D).

There is a simple interpretation to dual optimality coming directly from the Lagrangian: a dual variable $\boldsymbol{\lambda}$ is optimal if the candidate primal solution is primal feasible: $\mu^{\boldsymbol{\lambda}} \in \mathbb{L}_2$. It can be seen that the derivative of the dual function $L(\boldsymbol{\lambda})$ captures the slack of a $\mu^{\boldsymbol{\lambda}}$:

$$\frac{\partial L(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}_{e,i}(x_i)} = \mu_i^{\boldsymbol{\lambda}}(x_i) - S_{e,i}^{\boldsymbol{\lambda}}(x_i) \tag{1}$$

where we define $S_{e,i}^{\boldsymbol{\lambda}}(x_i) := \sum_{x_j\in\chi}\mu_e^{\boldsymbol{\lambda}}(x_i, x_j)$. The gradient captures the amount and direction of constraint violation in $\mathbb{L}_2$ by $\mu^{\boldsymbol{\lambda}}$. In order to discuss this object concisely and intuitively, we formally define the notion of a slack vector, which is simply the negative of the gradient, and a slack polytope (Lee et al., 2020), which describes the same polytope as $\mathbb{L}_2$ if the constraints were offset by exactly the amount by which $\mu^{\boldsymbol{\lambda}}$ is offest.

Definition 1 (Slack vector and slack polytope)**.** *For $\boldsymbol{\lambda} \in \mathbb{R}^{r_D}$, the slack vector $\nu^{\boldsymbol{\lambda}} \in \mathbb{R}^{r_D}$ of $\boldsymbol{\lambda}$ is defined as $\nu_{e,i}^{\boldsymbol{\lambda}}(x) = S_{e,i}^{\boldsymbol{\lambda}}(x) - \mu_i^{\boldsymbol{\lambda}}(x)$ for all $e \in E$, $i \in e$, and $x \in \chi$.*

The slack polytope for a slack vector ν is defined as

$$\mathbb{L}_2^{\nu} := \left\{\mu \in \mathbb{R}_+^{r_P} \;:\; \begin{array}{l}\mu_i \in \Sigma^d \\ \mu_e \in \mathcal{U}_d(\mu_i + \nu_{e,i}, \mu_j + \nu_{e,j})\end{array}\right\}$$

2.2. Entropy-Regularized Message Passing

The results in this paper will be primarily concerned with algorithms that approximately solve (MAP) by directly solving (Reg-D). For solving this objective, message passing algorithms can effectively be viewed as block-coordinate descent, except that a full minimization is typically taken at each step. Here we outline two variants.

Algorithm 1 Standard-MP(Update, η, P, K)

$\boldsymbol{\lambda}^{(0)} = 0$
for $k = 0, 1, \ldots, K-1$ **do**
  Set $\boldsymbol{\lambda}^{(k+1)} = \boldsymbol{\lambda}^{(k)}$
  Sample block-coordinate $b_k \sim P$
  Set $\boldsymbol{\lambda}_{b_k}^{(k+1)} = \mathsf{Update}_{b_k}^{\eta}(\boldsymbol{\lambda}^{(k)})$
end for
return $\arg\min_{\boldsymbol{\lambda}\in\{\boldsymbol{\lambda}^{(k)}\}}\sum_{e\in E, i\in e}\|\nu_{e,i}^{\boldsymbol{\lambda}}\|_1^2$

2.2.1. Edge Message Passing

Edge message passing (EMP) algorithms reduce to block-coordinate methods that minimize (Reg-D) for a specific edge $e = \{i, j\} \in E$ and endpoint vertex $i \in e$, while keeping all other dual variables fixed. Let $L_{e,i}(\cdot; \boldsymbol{\lambda}) : \mathbb{R}^d \to \mathbb{R}$ denote the block-coordinate loss of L fixed at $\boldsymbol{\lambda}$ except for free variables $\{\boldsymbol{\lambda}_{e,i}(x)\}_{x\in\chi}$. For each $i \in V$, $e \in N_i$ we define the EMP operator for $\boldsymbol{\lambda} \in \mathbb{R}^{r_D}$:

$$\mathsf{EMP}_{e,i}^{\eta}(\boldsymbol{\lambda}) \in \underset{\boldsymbol{\lambda}'_{e,i}\in\mathbb{R}^d}{\arg\min}\, L_{e,i}(\boldsymbol{\lambda}'_{e,i}(\cdot); \boldsymbol{\lambda}).$$

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

In the entropy-regularized setting, this update rule has been studied by Lee et al. (2020); Ravikumar et al. (2010). Non-regularized versions based on max-sum diffusion have much earlier roots and also been studied by Werner (2007; 2009); however, we do not consider these particular unregularized variants. EMP offers the following improvement on L.

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

2.2.2. Star Message Passing

Star message passing (SMP) algorithms consider block-coordinates that include all edges incident to a particular vertex $i \in V$. For $\boldsymbol{\lambda} \in \mathbb{R}^{r_D}$ and $i \in V$ let $L_i(\cdot; \boldsymbol{\lambda}) : \mathbb{R}^d \to \mathbb{R}$ denote the block-coordinate loss of L fixed at $\boldsymbol{\lambda}$. For a given $i \in V$ and arbitrary $\boldsymbol{\lambda} \in \mathbb{R}^{r_D}$, we define the SMP operator:

$$\mathsf{SMP}_i^{\eta}(\boldsymbol{\lambda}) \in \underset{\boldsymbol{\lambda}'_{\cdot,i}\in\mathbb{R}^{d|N_i|}}{\arg\min}\, L_i(\boldsymbol{\lambda}'_{\cdot,i}(\cdot); \boldsymbol{\lambda})$$

That is, SMP is the minimization over the block-coordinate at vertex i for all edges incident to i in N_i and all possible labels in χ.

Algorithm 2 Accel-EMP(G, C, η, K)

$\boldsymbol{\lambda}^{(0)} = 0, \mathbf{v}^{(0)} = 0, \theta_{-1} = 1$
for $k = 0, 1, \ldots, K-1$ **do**
  $\theta_k = \frac{-\theta_{k-1}^2 + \sqrt{\theta_{k-1}^4 + 4\theta_{k-1}^2}}{2}$
  $\mathbf{y}^{(k)} = \theta_k \mathbf{v}^{(k)} + (1-\theta_k)\boldsymbol{\lambda}^{(k)}$
  Sample $(e_k, i_k) \sim \text{Unif}\{(e,i) \,:\, e \in E, i \in e\}$.
  Set $\boldsymbol{\lambda}^{(k+1)} = \boldsymbol{\lambda}^{(k)}$
  $\boldsymbol{\lambda}_{e,i}^{(k+1)}(\cdot) = \text{EMP}_{e_k,i_k}^{\eta}(\mathbf{y}^{(k)})$
  $\mathbf{v}^{(k+1)} = \mathbf{v}^{(k)}$
  $\mathbf{v}_{e_k,i_k}^{(k+1)} = \mathbf{v}_{e_k,i_k}^{(k)} + \frac{1}{2m\eta\theta_k} \nu_{e_k,i_k}^{\mathbf{y}^{(k)}}$
end for
return $\boldsymbol{\lambda}^{(K)}$

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

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

The proof is similar to the previous one and is deferred to the appendix. Meshi et al. (2012) gave a concise definition and analysis of algorithms from this update rule, and similar star-based algorithms have existed much earlier (Wainwright & Jordan, 2008), such as MP-LP (Globerson & Jaakkola, 2008). Due to Meshi et al. (2012), SMP also has an improvement guarantee.

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

2.3. Randomized Standard Algorithms

The message passing updates described in the previous subsection can be applied to each block-coordinate in many different orders. In this paper, we consider using the updates in a randomized manner, adhering to the generalized procedure presented in Algorithm 1. The algorithm takes as input the update rule Update, which could be EMP or SMP, and a regularization parameter η. It also requires a distribution P over block-coordinates b_k for each iteration $k \leq K$. In this paper, we will use the uniform distribution over edge-vertex pairs for EMP:

$$b_k = (e_k, i_k) \sim \text{Unif}(\{(e,i) \,:\, e \in E,\ i \in e\}). \quad (2)$$

For SMP, we use a categorical distribution over vertices based on the number of neighbors of each vertex:

$$b_k = i_k \sim \text{Cat}(V, \{p_i\}_{i \in V}), \quad (3)$$

where $p_i = \frac{|N_i|}{\sum_{j \in V}|N_j|}$ for each $i \in V$.

Algorithm 3 Accel-SMP(G, C, η, K)

$\boldsymbol{\lambda}^{(0)} = 0, \mathbf{v}^{(0)} = 0, \theta_{-1} = 1$
for $k = 0, 1, \ldots, K-1$ **do**
  $\theta_k = \frac{-\theta_{k-1}^2 + \sqrt{\theta_{k-1}^4 + 4\theta_{k-1}^2}}{2}$
  $\mathbf{y}^{(k)} = \theta_k \mathbf{v}^{(k)} + (1-\theta_k)\boldsymbol{\lambda}^{(k)}$
  Sample $i_k \sim \{p_i\}_{i \in V}$
  Set $\boldsymbol{\lambda}^{(k+1)} = \boldsymbol{\lambda}^{(k)}$
  $\boldsymbol{\lambda}_{\cdot,i}^{(k+1)}(\cdot) = \text{SMP}_{i_k}^{\eta}(\mathbf{y}^{(k)})$
  $\mathbf{v}^{(k+1)} = \mathbf{v}^{(k)}$
  for $e \in N_{i_k}$ **do**
    $\mathbf{v}_{e,i_k}^{(k+1)} = \mathbf{v}_{e,i_k}^{(k)} + \frac{\min_j |N_j|}{2p_{i_k}\theta_k\eta N} \nu_{e,i_k}^{\mathbf{y}^{(k)}}$
  end for
end for
return $\boldsymbol{\lambda}^{(K)}$

3. Accelerating Entropy-Regularized Message Passing

We now present a scheme for accelerating message passing algorithms in the entropy-regularized formulation. The key idea is to leverage the block-coordinate nature of standard message passing algorithms. We draw inspiration from both the seminal work of Nesterov (2018) on accelerating gradient methods and accelerating randomized coordinate gradient methods similar to Lee & Sidford (2013); Lu & Xiao (2015); however, the presented method is specialized to incorporate full block-coordinate minimization at each round, so as to be consistent with existing message passing algorithms used in practice. Furthermore, we can leverage the same simple sampling procedures for selecting the block-coordinates. The presented scheme can thus be viewed as a direct method of acceleration in that standard message passing algorithms can be plugged in.

The scheme is presented in Algorithm 2 for EMP and Algorithm 3 for SMP. In Accel-EMP, at each round k, a random coordinate block is sampled uniformly. That coordinate block for $\boldsymbol{\lambda}$ is then updated with a step of $\text{EMP}_{e_k,i_k}^{\eta}$ evaluated at $\mathbf{y}^{(k)}$. A block-coordinate gradient step evaluated at $\mathbf{y}^{(k)}$ in the form of the slack vector $\nu_{e_k,i_k}^{\mathbf{y}^{(k)}}$ is also applied to $\mathbf{v}^{(k)}$. Accel-SMP works similarly but we instead sample from the non-uniform distribution defined in (3). The choice of distributions ultimately determines the step size for $\mathbf{v}$.

As in the case of the standard smooth message passing algorithms, which optimize the dual function (Reg-D), the returned solutions may not be primal feasible in finite iterations. To obtain feasible solutions, we consider a projection operation Proj, shown in Algorithm 4, that is simply a repeated application of Algorithm 2 of Altschuler et al. (2017), originally designed for optimal transport. The method ef-

Algorithm 4 $\mathsf{Proj}(\mu, \nu)$

Set $\widehat{\mu}_i = \mu_i$ for all $i \in V$
for $(i,j) = e \in E$ **do**
  Compute $\widehat{\mu}_e$ by applying Algorithm 2 of (Altschuler et al., 2017) on μ_e with transportation polytope $\mathcal{U}_d(\mu_i + \nu_{e,i}, \mu_j + \nu_{e,j})$
end for
return $\widehat{\mu}$

fectively finds an edge marginal $\widehat{\mu}_e$ that sums to the given vertex marginals μ_i and μ_j for $e = (i,j)$ plus some optional slack ν. For all practical purposes, we would always set the slack to be $\nu = 0$, ensuring that Proj outputs a point in $\mathbb{L}_2$; however, it will become useful to project points from $\mathbb{L}_2$ into a particular slack polytope $\mathbb{L}_2^{\nu_\lambda}$ for the analysis. When projecting onto $\mathbb{L}_2$, Proj does not require modifying the vertex marginals, so there is no ambiguity if the approximate solution is ultimately rounded to an integral solution using a simple vertex rounding scheme[2].

4. Main Results

We now present iteration complexities for the above algorithms for finding ϵ-optimal solutions to the original unregularized problem (P) over $\mathbb{L}_2$. These guarantees make it easy to compare various algorithms as they do not inherently depend on the tightness of the relaxation, rounding heuristics to find integral solutions, or arbitrary choices of η.

4.1. Standard Algorithms

Our first result bounds the number of iterations required to compute ϵ-optimal solutions to (P). To recapitulate, prior work by Lee et al. (2020) for EMP only provided an iteration guarantee for bounding the norm of the slack vector. They also gave guarantees for the number of iterations required to round to the optimal MAP solution when it is available. Meshi et al. (2012) gave a guarantee on both the primal and dual *regularized* problems (Reg-P) and (Reg-D), but not the original (P) and without rounding or tuning of η. Additionally, both works focused mostly on the ϵ dependence in the rate rather than actually specifying the graph parameters m, n, and d.

In contrast to these prior works, we give a guarantee on optimality for (P) for the standard randomized algorithms, specifying exactly the dependence on graph parameters. The purpose of this extension is to standardize convergence guarantees for the true relaxed problem, which will ultimately be handy for comparing to our primary contribution on the accelerated algorithms.

[2]Ambiguity could arise for more sophisticated rounding schemes but we do not consider those here.

Theorem 1. *Let $\widehat{\boldsymbol{\lambda}}$ be the result of running Algorithm 1 with EMP, uniform sampling distribution (2) and $\eta = \frac{4(m+n)\log d}{\epsilon}$. Let $\widehat{\mu} = \mathsf{Proj}(\mu^{\widehat{\boldsymbol{\lambda}}}, 0)$ be its projection onto $\mathbb{L}_2$. Then, the number of iterations sufficient for $\widehat{\mu}$ to be expected ϵ-optimal is*

$$O\left(md^2(m+n)^4\|C\|_\infty^3\epsilon^{-3}\log d\right).$$

If $\widehat{\boldsymbol{\lambda}}$ is the output of Algorithm 1 using SMP and sampling distribution (3), and $\widehat{\mu} := \mathsf{Proj}(\mu^{\widehat{\boldsymbol{\lambda}}}, 0)$ then $\widehat{\mu}$ is expected ϵ-optimal in the same order of iterations.

The rate harbors a $O(1/\epsilon^3)$ dependence, which at first appears to be worse those of Meshi et al. (2012) and Lee et al. (2020); however, their convergence guarantees hold only for the regularized objective. The extra $O(1/\epsilon)$ in our guarantee occurs in the conversion to the original unregularized problem (P), which is a stronger result. It is interesting to observe that the guarantees are effectively the same for both variants, despite having somewhat different analyses. We hypothesize that this is due to the fact that the "smoothness" constant for SMP in Lemma 2 is greater than that of EMP in Lemma 1. Therefore, the larger block-coordinate size is effectively canceled by the smaller improvement per step.

We now describe the proof briefly here since the first part is fairly standard while the second will be covered in the proof of our main acceleration result. The full proof is found in Appendix D. The basic idea is to use Lemma 1 to lower bound the expected improvement each iteration, which can be done in terms of the average squared norms of the slack $\|\nu_{e,i}\|_1^2$. We can guarantee improvement on L by at least $(\epsilon')^2$ until the norms are on average below ϵ'. Knowing that the slack norms are small, we can prove that the projection $\widehat{\mu}$ of $\mu^{\widehat{\boldsymbol{\lambda}}}$ onto $\mathbb{L}_2$ is not too far from $\mu^{\widehat{\boldsymbol{\lambda}}}$ and so the expected value of $\langle C, \widehat{\mu}\rangle$ is not much worse than that of $\langle C, \mu^{\boldsymbol{\lambda}}\rangle$. We then prove that $\langle C, \mu^{\boldsymbol{\lambda}}\rangle$ is small with respect to the slack norms up to some entropy term, and we set η so the entropy term is sufficiently small with respect to a given $\epsilon > 0$.

4.2. Accelerated Algorithms

Our primary result gives improved iteration complexities for the accelerated versions of EMP and SMP. To do so, we rely on the classic estimate sequence method initially developed by Nesterov. In particular, we turn to a randomized variant, which has appeared before in the literature on randomized coordinate gradient methods by Lee & Sidford (2013); Lu & Xiao (2015) for the strongly convex and general cases respectively. Our main contributions are both extending these results for the full minimization of message passing to achieve the fast rates and also proving the accelerating guarantee on the original relaxed problem (P) rather than the regularized problems.

Theorem 2. *Let $\widehat{\boldsymbol{\lambda}}$ be the output of Algorithm 2 with $\eta = \frac{4(m+n)\log d}{\epsilon}$. Let $\hat{\mu} = \mathsf{Proj}(\mu^{\hat{\boldsymbol{\lambda}}}, 0)$ be its projection onto $\mathbb{L}_2$. Then, the number of iterations sufficient for $\hat{\mu}$ to be expected ϵ-optimal is*

$$O\left(m^{3/2}d^2(m+n)^3\|C\|_\infty^2\epsilon^{-2}\log d\right).$$

If $\widehat{\boldsymbol{\lambda}}$ is the output of Algorithm 3 and $\widehat{\mu} := \mathsf{Proj}(\mu^{\widehat{\boldsymbol{\lambda}}}, 0)$, then $\widehat{\mu}$ is expected ϵ-optimal in the same order of iterations.

The primary difference between the iteration complexities for the standard algorithms and the accelerated ones is the dependence on ϵ. For the accelerated algorithms, we are left with only a $O(1/\epsilon^2)$ dependence versus the $O(1/\epsilon^3)$ dependence for the standard algorithms. This can lead to far fewer iterations to get the same level of accuracy on the original relaxed problem (P). In addition, the bounds in Theorem 2 are strictly better in dependence on the number of edges as well for both EMP and SMP. For $m+n \approx m$, the accelerated algorithms shave off a $m^{1/2}$ factor. It is interesting to observe that these improved guarantees come with virtually no extra computation per iteration. For example, to update the sequences $\boldsymbol{\lambda}^{(k)}$, $\mathbf{v}^{(k)}$, and $\mathbf{y}^{(k)}$ in EMP at each iteration, we need only compute the primal variables $\mu_{i_k}^{\boldsymbol{\lambda}^{(k)}}$ and $S_{e_k,i_k}^{\boldsymbol{\lambda}^{(k)}}$ once to use in both the slack vector $\nu_{e_k,i_k}^{\boldsymbol{\lambda}^{(k)}}$ and the update rule $\mathsf{EMP}_{e_k,i_k}^{\eta}(\boldsymbol{\lambda}^{(k)})$.

We will give the proof for Accel-EMP to convey the main idea. The analogous Accel-SMP case can be found in the appendix. First, we will derive a faster convergence rate on the dual objective, which in turn implies that we can bound the slack norms by the same $\epsilon' > 0$ in fewer iterations. In the second part, we will bound the approximation error caused by the entropy regularization. Finally, we put these pieces together to determine the appropriate choice of ϵ' and η in terms of ϵ to recover the final rate.

4.2.1. Faster Convergence on the Dual

The first steps will involve defining a randomized estimate sequence to work with and then using this sequence to prove the faster convergence rate on the dual objective.

Definition 2. *Let $\phi_0 : \mathbb{R}^{r_D} \to \mathbb{R}$ be an arbitrary deterministic function. A sequence $\{\phi_k, \delta_k\}_{k=0}^K$ of random functions $\phi_k : \mathbb{R}^{r_D} \to \mathbb{R}$ for $k \geq 1$ and deterministic real values $\delta_k \in \mathbb{R}_+$ is a randomized sequence for $L(\boldsymbol{\lambda})$ if it satisfies $\delta_k \xrightarrow{k} 0$ and, for all k, $\mathbb{E}[\phi_k(\boldsymbol{\lambda})] \leq (1-\delta_k)L(\boldsymbol{\lambda}) + \delta_k\phi_0(\boldsymbol{\lambda})$.*

If we are given a random estimate sequence $\{\phi_k, \delta_k\}_{k=0}^K$ and a random sequence $\{\boldsymbol{\lambda}^{(k)}\}_{k=0}^K$ that satisfies $\mathbb{E}[L(\boldsymbol{\lambda}^{(k)})] \leq \min_{\boldsymbol{\lambda}} \mathbb{E}\left[\phi_k(\boldsymbol{\lambda})\right]$, then

$$\begin{aligned}\mathbb{E}[L(\boldsymbol{\lambda}^{(k)})] - L(\boldsymbol{\lambda}^*) &\leq \min_{\boldsymbol{\lambda}} \mathbb{E}[\phi_k(\boldsymbol{\lambda})] - L(\boldsymbol{\lambda}^*) \\ &\leq \delta_k(\phi_0(\boldsymbol{\lambda}^*) - L(\boldsymbol{\lambda}^*))\end{aligned} \tag{4}$$

This expected error converges to zero since $\delta_k \xrightarrow{k} 0$. We now identify a candidate estimate sequence. Let $\boldsymbol{\lambda}^{(0)} = 0$, $\delta_0 = 1$, $\boldsymbol{\lambda}^* \in \Lambda^*$, and $q := 2m$. Let the sequence $\{\theta_k\}_{k=0}^K$ be as it is defined in Algorithm 2 and let $\{\mathbf{y}^{(k)}\}_{k=0}^K \subset \mathbb{R}^{r_D}$ be arbitrary. Consider $\{\phi_k, \delta_k\}_{k=0}^K$ defined recursively as

$$\begin{aligned}(e_k, i_k) &\sim \mathrm{Unif}\{(e,i) : e \in E, i \in e\} \\ \delta_{k+1} &= (1-\theta_k)\delta_k \\ \phi_{k+1}(\boldsymbol{\lambda}) &= (1-\theta_k)\phi_k(\boldsymbol{\lambda}) + \theta_k L(\mathbf{y}^{(k)}) \\ &\quad - q\theta_k\langle \nu_{e_k,i_k}^{\mathbf{y}^{(k)}}, \boldsymbol{\lambda}_{e_k,i_k} - \mathbf{y}_{e_k,i_k}^{(k)}\rangle\end{aligned} \tag{5}$$

where $\phi_0(\boldsymbol{\lambda}) = L(\boldsymbol{\lambda}^{(0)}) + \frac{\gamma_0}{2}\|\boldsymbol{\lambda}^{(0)} - \boldsymbol{\lambda}\|_2^2$ for $\gamma_0 = 2q^2\eta$.

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*

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

The proof is similar to what is given by Lee & Sidford (2013), but since we consider the non-strongly convex case and slightly different definitions, we give a full proof in Appendix C for completeness. We can use this fact to show a rate of convergence on the dual objective.

Lemma 4. *For the random estimate sequence in (5), let $\{\boldsymbol{\lambda}^{(k)}\}_{k=0}^K$ and $\{\mathbf{y}^{(k)}\}_{k=0}^K$ be defined as in Algorithm 2 with $\boldsymbol{\lambda}^{(0)} = 0$. Then, the dual objective error in expectation can be bounded as $\mathbb{E}[L(\boldsymbol{\lambda}^{(k)}) - L(\boldsymbol{\lambda}^*)] \leq \frac{G(\eta)^2}{(k+2)^2}$, where $G(\eta) := 24md(m+n)(\sqrt{\eta}\|C\|_\infty + \frac{\log d}{\sqrt{\eta}})$.*

Proof. It suffices to show that the sequence in (5) with the definitions of $\{\boldsymbol{\lambda}^{(k)}\}_{k=0}^K$ and $\{\mathbf{y}^{(k)}\}_{k=0}^K$ satisfies $\mathbb{E}[L(\boldsymbol{\lambda}^{(k)})] \leq \min_{\boldsymbol{\lambda}} \mathbb{E}\left[\phi_k(\boldsymbol{\lambda})\right]$. To do this, we will use induction and Lemma 1. Note that $\min_{\boldsymbol{\lambda}} \phi_k(\boldsymbol{\lambda}) = \omega_k$. The base case holds trivially with $\mathbb{E}[\omega_0] = L(\boldsymbol{\lambda}^{(0)})$. Suppose $\mathbb{E}\left[L(\boldsymbol{\lambda}^{(k)})\right] \leq \mathbb{E}\left[\omega_k\right]$ at iteration k. For $k+1$, we have

$$\begin{aligned}&\mathbb{E}[\omega_{k+1}] \\ &\geq (1-\theta_k)\mathbb{E}[L(\boldsymbol{\lambda}^{(k)})] + \theta_k\mathbb{E}[L(\mathbf{y}^{(k)})] \\ &\quad - \mathbb{E}\left[\frac{(\theta_k q)^2}{2\gamma_{k+1}}\|\nu_{i_k,e_k}^{\mathbf{y}^{(k)}}\|_2^2 - \theta_k q\langle \nu_{e_k,i_k}^{\mathbf{y}^{(k)}}, \mathbf{v}^{(k)} - \mathbf{y}^{(k)}\rangle\right].\end{aligned}$$

The above inequality uses the inductive hypothesis. Then,

$$\begin{aligned}
&\geq \mathbb{E}\left[L(\mathbf{y}^{(k)}) - \frac{(\theta_k q)^2}{2\gamma_{k+1}}\|\nu^{\mathbf{y}^{(k)}}_{i_k,e_k}\|_2^2\right] \\
&\quad + (1-\theta_k)\mathbb{E}\left[\langle \nabla L(\mathbf{y}^{(k)}), \boldsymbol{\lambda}^{(k)} - \mathbf{y}^{(k)}\rangle\right] \\
&\quad + \theta_k \mathbb{E}\left[\langle \nabla L(\mathbf{y}^{(k)}), \mathbf{v}^{(k)} - \mathbf{y}^{(k)}\rangle\right] \\
&= \mathbb{E}\left[L(\mathbf{y}^{(k)}) - \frac{\theta_k^2 q}{2\gamma_{k+1}}\|\nabla L(\mathbf{y}^{(k)})\|_2^2\right].
\end{aligned}$$

This second inequality uses convexity of L and then applies the law of total expectation to the dot products, conditioning on randomness $\{e_s, i_s\}_{s=1}^{k-1}$. The last identity uses the definition of $\mathbf{y}^{(k)}$ and total expectation again on the norm.

Using Lemma 1 and the definition of $\boldsymbol{\lambda}^{(k+1)}$ from Algorithm 2, we have

$$\begin{aligned}
E[L(\boldsymbol{\lambda}^{(k+1)}] &\leq \mathbb{E}[L(\mathbf{y}^{(k)}) - \frac{1}{4\eta}\|\nu^{\mathbf{y}^{(k)}}_{e_k,i_k}\|_2^2] \\
&= \mathbb{E}[L(\mathbf{y}^{(k)}) - \frac{1}{4q\eta}\|\nabla L(\mathbf{y}^{(k)})\|_2^2]
\end{aligned}$$

The equality uses the law of total expectation. Therefore, taking $\theta_k^2 := \frac{\gamma_{k+1}}{2q^2\eta}$ ensures that $\min_{\boldsymbol{\lambda}} \mathbb{E}[\phi_{k+1}(\boldsymbol{\lambda})] \geq \mathbb{E}[\omega_{k+1}] \geq E[L(\boldsymbol{\lambda}^{(k+1)}]$. By setting $\gamma_0 = 2q^2\eta$, we ensure that θ_k need only satisfy $\theta_k^2 = (1-\theta_k)\theta_{k-1}^2$, which occurs in Algorithm 2. From Nesterov (2018, Lemma 2.2.4), this choice of γ_0 and θ_k ensures $\delta_k \leq \left(1 + \frac{k}{q}\sqrt{\frac{\gamma_0}{8\eta}}\right)^{-2} = \frac{4}{(k+2)^2}$. From (4) and the definition of ϕ_0, we get

$$\mathbb{E}[L(\boldsymbol{\lambda}^{(k)}) - L(\boldsymbol{\lambda}^*)] \leq \frac{4L(0) - 4L(\boldsymbol{\lambda}^*) + 16m^2\eta\|\boldsymbol{\lambda}^*\|_2^2}{(k+2)^2}.$$

The proof that the numerator can be bounded by $G(\eta)^2$ is deferred to the appendix, Lemma 7. □

4.2.2. Approximation Error due to Entropy

Recall that our end goal is to ensure that $\widehat{\mu}$, the projection of $\mu^{\widehat{\boldsymbol{\lambda}}}$ onto $\mathbb{L}_2$, is expected ϵ-optimal for $\epsilon > 0$. To show this, we need to develop some relations which we outline here for brevity but state formally and prove in Appendix C.2. The first relation is how close $\widehat{\mu} \in \mathbb{L}_2$ is to the algorithm's output $\mu^{\widehat{\boldsymbol{\lambda}}}$, which is in the slack polytope $\mathbb{L}_2^{\nu^{\widehat{\boldsymbol{\lambda}}}}$. This is essentially a direct extension of Altschuler et al. (2017, Lemma 7), which tells us that we can bound the projection by the norm of the slacks. Then, we show that there exists a point $\widehat{\mu}^*$ in the slack polytope $\mathbb{L}_2^{\nu^{\widehat{\boldsymbol{\lambda}}}}$ that is close to the optimal point $\mu^* \in \mathbb{L}_2$ with respect to the norm of the slacks also. We use the relations to conclude a bound on the original relaxed problem (P) for any realization of $\widehat{\mu}$.

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:*

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

4.2.3. Completing the Proof

Proof of Theorem 2 for EMP. Let $\widehat{\boldsymbol{\lambda}}$ be the output from Algorithm 2 after K iterations. From Lemma 1, we can lower bound the result in Lemma 4 with $\frac{1}{4\eta}\mathbb{E}[\|\nu^{\widehat{\boldsymbol{\lambda}}}_{e,i}\|_1^2] \leq \frac{G(\eta)}{(K+2)^2}$ for all $e \in E, i \in e$. Then, for $\epsilon' > 0$, we can ensure that

$$\mathbb{E}[\|\nu^{\widehat{\boldsymbol{\lambda}}}_{e,i}\|_1] \leq \epsilon' \quad \text{and} \quad \mathbb{E}\sum_{e\in E, i\in e}\|\nu^{\widehat{\boldsymbol{\lambda}}}_{e,i}\|_1^2 \leq 2m(\epsilon')^2$$

in $K = \frac{\sqrt{4\eta}G(\eta)}{\epsilon'}$ iterations. Let $\widehat{\mu} \in \mathbb{L}_2$ be the projected version of $\mu^{\widehat{\boldsymbol{\lambda}}}$. Taking the expectation of both sides of the result in Proposition 4 gives us

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

where $\mathbb{E}[\delta]^2 \leq \mathbb{E}[\delta^2] \leq \mathbb{E}\sum_{e\in E, i\in e}\|\nu^{\widehat{\boldsymbol{\lambda}}}_{e,i}\|_1^2 \leq 2m(\epsilon')^2$. Then we can conclude

$$\begin{aligned}
\mathbb{E}[\langle C, \widehat{\mu} - \mu^*\rangle] &\leq 16\sqrt{2m}(m+n)d\|C\|_\infty\epsilon' \\
&\quad + 8m\|C\|_\infty\epsilon' + \frac{n\log d + 2m\log d}{\eta} \\
&\leq 24\sqrt{2m}(m+n)d\|C\|_\infty\epsilon' \\
&\quad + \frac{n\log d + 2m\log d}{\eta}.
\end{aligned}$$

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

5. Rounding to Integral Solutions

Inspired by recent results regarding the approximation error achieved by entropy regularization in linear programming (Weed, 2018), we are able to derive rounding guarantees for our algorithms under the assumption that the LP relaxation is tight and the solution is unique. We use a simple rounding scheme: for any μ that may not lie in $\mathbb{L}_2$, $(\text{round}(\mu))_i = \arg\max_x \mu_i(x)$. The main challenge in achieving the results we present here is surpassing the difficulty in obtaining bounds for the l_1 distance between $\widehat{\mu} = \mu^{\widehat{\boldsymbol{\lambda}}}$, the candidate solution obtained from the final iterate $\widehat{\boldsymbol{\lambda}}$ resulting from our algorithms, and μ^*, the optimal solution of (P). Define $\mu^{\boldsymbol{\lambda}^*}$ be the solution to the regularized problem where $\boldsymbol{\lambda}^* \in \Lambda^*$.

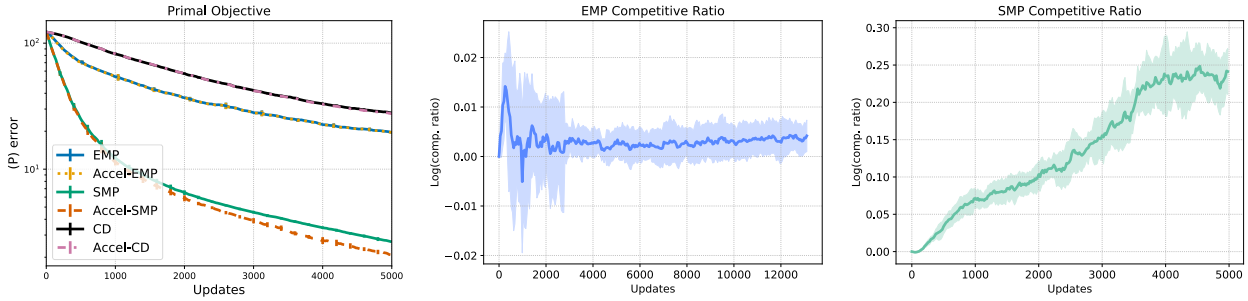

Figure 1. (Left column): The original primal objective (P) on a log scale is compared for the standard algorithms and their accelerated variants, as well as standard (accelerated) coordinate descent, over 10 trials on an Erdős-Rényi random graph with $n = 100$. Error bars denote standard deviation. *(Center and right columns)*: The log-competitive ratio of EMP and SMP with respect to their accelerated variants, Accel-EMP and Accel-SMP, on the primal objective (P).

We proceed in two steps. First, we bound the approximation error $\|\mu^{\boldsymbol{\lambda}^*} - \mu^*\|_1$ using recent results on the quality of solutions for entropy regularized $\mathbb{L}_2$ (Lee et al., 2020). Then we bound the optimization error of $\widehat{\mu}$ using the results derived in the previous section. The proof and a comparison to standard EMP are in Appendix F. Let deg denote the maximum degree of the graph G and define Δ as the suboptimality gap of the LP over $\mathbb{L}_2$.

Theorem 3. *Let $\delta \in (0, 1)$. If $\mathbb{L}_2$ is tight for potential vector C and there exist a unique solution to the MAP problem and $\eta = \frac{16(m+n)(\log(m+n)+\log(d))}{\Delta}$ then with probability $1 - \delta$ in at most*

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

iterations of Accel-EMP*, $\mathrm{round}(\widehat{\mu})$ is the optimal MAP assignment.*

6. Numerical Experiments

Our goal now is to understand the empirical differences between the above algorithms and also where certain theoretical guarantees can likely be improved, if at all. To this end, we compare the convergence rates of EMP and SMP and their accelerated variants on several synthesized Erdős-Rényi random graphs. First, we constructed a graph with $n = 100$ vertices and then generated edges between each pair of vertices with probability $1.1\frac{\log n}{n}$. We considered the standard multi-label Potts model with $d = 3$ labels. The cost vector C was initialized randomly in the following way: $C_i(x_i) \sim \mathrm{Unif}([-0.01, 0.01]), \forall x_i \in \chi$ and $C_{ij}(x_i, x_j) \sim \mathrm{Unif}(\{-1.0, 1.0\}), \forall x_i, x_j \in \chi$.

We consider two different metrics. (1) The first is the original primal objective value (P). This metric computes the objective value of the projection $\widehat{\mu} = \mathsf{Proj}(\mu^{\boldsymbol{\lambda}}, 0)$. (2) The second reports the log-competitive ratio between the standard and accelerated variants. 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. Positive values indicate that the accelerated variant has lower error. We implemented the four message passing algorithms exactly as they are described in Algorithms 1, 2, and 3. We also implemented block-coordinate descent and its accelerated variant as baselines (Lee & Sidford, 2013) with a step size of $1/\eta$. Each algorithm used $\eta = 1000$ over 10 trials, measuring means and standard deviations. We computed the ground-truth optimal value of (P) using a standard solver in CVXPY.

Figure 1 depicts convergence on the primal objective in the left column. SMP achieves convergence in the fewest iterations, and EMP converges faster than coordinate descent. Interestingly, we find that the accelerated variants, including accelerated coordinate descent, appear to have marginal improvement on this metric. However, the competitive ratio figures confirm that the accelerated variants are consistently faster, especially for SMP. These results suggest that, at least for this particular problem, the upper bounds for standard algorithms may be overly conservative. It would be interesting to investigate tighter bounds for the standard algorithms in future work. Further details can be found in the appendix.

7. Conclusion

We analyze the convergence of message passing algorithms on the MAP inference problem over $\mathbb{L}_2$. In addition to providing a novel rate of convergence rate for standard schemes derived from entropy regularization, we show that they can be directly accelerated in the sense of Nesterov with significant theoretical improvement. In future work it would be interesting to consider accelerating greedy message passing algorithms; however, Lu et al. (2018) suggest that, despite empirical success, proving accelerated rates for greedy methods is an open question even in the basic coordinate descent case. The tightness of the presented guarantees is also an open question, motivated by the empirical results here. Finally, we conjecture that reductions from the active area of optimal transport could yield novel, faster algorithms.

References

Altschuler, J., Weed, J., and Rigollet, P. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pp. 1964–1974, 2017.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Cooper, G. F. The computational complexity of probabilistic inference using bayesian belief networks. *Artificial intelligence*, 42(2-3):393–405, 1990.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 2292–2300, 2013.

Dvurechensky, P., Gasnikov, A., and Kroshnin, A. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In *35th International Conference on Machine Learning, ICML 2018*, pp. 2196–2220, 2018.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pp. 3440–3448, 2016.

Globerson, A. and Jaakkola, T. S. Fixing max-product: Convergent message passing algorithms for map lp-relaxations. In *Advances in Neural Information Processing Systems*, pp. 553–560, 2008.

Gondzio, J. Interior point methods 25 years later. *European Journal of Operational Research*, 218(3):587–601, 2012.

Hazan, T. and Shashua, A. Convergent message-passing algorithms for inference over general graphs with convex free energies. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pp. 264–273, 2008.

Jegelka, S. and Bilmes, J. Submodularity beyond submodular energies: coupling edges in graph cuts. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1897–1904. IEEE, 2011.

Jojic, V., Gould, S., and Koller, D. Accelerated dual decomposition for map inference. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 503–510, 2010.

Kappes, J., Andres, B., Hamprecht, F., Schnorr, C., Nowozin, S., Batra, D., Kim, S., Kausler, B., Lellmann, J., Komodakis, N., et al. A comparative study of modern inference techniques for discrete energy minimization problems. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1328–1335, 2013.

Karmarkar, N. A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pp. 302–311, 1984.

Kolmogorov, V. Convergent tree-reweighted message passing for energy minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1568–1583, 2006.

Kolmogorov, V. and Zabin, R. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004.

Lee, J., Pacchiano, A., and Jordan, M. Convergence rates of smooth message passing with rounding in entropy-regularized map inference. In *International Conference on Artificial Intelligence and Statistics*, pp. 3003–3014, 2020.

Lee, Y. T. and Sidford, A. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 147–156. IEEE, 2013.

Lee, Y. T. and Sidford, A. Path finding methods for linear programming: Solving linear programs in o (vrank) iterations and faster algorithms for maximum flow. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 424–433. IEEE, 2014.

Lin, T., Ho, N., and Jordan, M. I. On the acceleration of the sinkhorn and greenkhorn algorithms for optimal transport. *arXiv preprint arXiv:1906.01437*, 2019.

Lu, H., Freund, R., and Mirrokni, V. Accelerating greedy coordinate descent methods. In *International Conference on Machine Learning*, pp. 3257–3266, 2018.

Lu, Z. and Xiao, L. On the complexity analysis of randomized block-coordinate descent methods. *Mathematical Programming*, 152(1-2):615–642, 2015.

MacKay, D. J. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Meshi, O., Globerson, A., and Jaakkola, T. S. Convergence rate analysis of map coordinate minimization algorithms. In *Advances in Neural Information Processing Systems*, pp. 3014–3022, 2012.

Meshi, O., Mahdavi, M., and Schwing, A. Smooth and strong: Map inference with linear convergence. In *Advances in Neural Information Processing Systems*, pp. 298–306, 2015.

Mezard, M. and Montanari, A. *Information, physics, and computation*. Oxford University Press, 2009.

Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

Nesterov, Y. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Nesterov, Y. *Lectures on convex optimization*, volume 137. Springer, 2018.

Ravikumar, P., Agarwal, A., and Wainwright, M. J. Message-passing for graph-structured linear programs: Proximal methods and rounding schemes. *Journal of Machine Learning Research*, 11(Mar):1043–1080, 2010.

Renegar, J. A polynomial-time algorithm, based on newton's method, for linear programming. *Mathematical programming*, 40(1-3):59–93, 1988.

Savchynskyy, B., Kappes, J., Schmidt, S., and Schnörr, C. A study of nesterov's scheme for lagrangian decomposition and map labeling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1817–1823. IEEE, 2011.

Savchynskyy, B., Schmidt, S., Kappes, J., and Schnörr, C. Efficient mrf energy minimization via adaptive diminishing smoothing. *arXiv preprint arXiv:1210.4906*, 2012.

Schiex, T., Fargier, H., and Verfaillie, G. Valued constraint satisfaction problems: hard and easy problems. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 631–637. Morgan Kaufmann Publishers Inc., 1995.

Sherali, H. D. and Adams, W. P. A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. *SIAM Journal on Discrete Mathematics*, 3(3):411–430, 1990.

Sidford, A. and Tian, K. Coordinate methods for accelerating ℓ_∞ regression and faster approximate maximum flow. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 922–933. IEEE, 2018.

Sontag, D., Globerson, A., and Jaakkola, T. Introduction to dual composition for inference. In *Optimization for Machine Learning*. MIT Press, 2011.

Torada, L., Lorenzon, L., Beddis, A., Isildak, U., Pattini, L., Mathieson, S., and Fumagalli, M. Imagene: a convolutional neural network to quantify natural selection from genomic data. *BMC bioinformatics*, 20(9):337, 2019.

Wainwright, M. J. and Jordan, M. I. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Weed, J. An explicit analysis of the entropic penalty in linear programming. In *Conference On Learning Theory*, pp. 1841–1855, 2018.

Weiss, Y., Yanover, C., and Meltzer, T. Map estimation, linear programming and belief propagation with convex free energies. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pp. 416–425, 2007.

Werner, T. A linear programming approach to max-sum problem: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(7):1165–1179, 2007.

Werner, T. Revisiting the linear programming relaxation approach to gibbs energy minimization and weighted constraint satisfaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1474–1488, 2009.

Yanover, C., Meltzer, T., and Weiss, Y. Linear programming relaxations and belief propagation–an empirical study. *Journal of Machine Learning Research*, 7(Sep): 1887–1907, 2006.