The Benefits of Learning with Strongly Convex Approximate Inference

Ben London BLONDON@CS.UMD.EDU
University of Maryland, College Park, MD 20742 USA

Bert Huang BHUANG@VT.EDU
Virginia Tech, Blacksburg, VA 24061 USA

Lise Getoor GETOOR@SOE.UCSC.EDU
University of California, Santa Cruz, CA 95064 USA

Abstract

We explore the benefits of strongly convex free energies in variational inference, providing both theoretical motivation and a new meta-algorithm. Using the duality between strong convexity and stability, we prove a high-probability bound on the error of learned marginals that is inversely proportional to the modulus of convexity of the free energy, thereby motivating free energies whose moduli are constant with respect to the size of the graph. We identify sufficient conditions for $\Omega(1)$-strong convexity in two popular variational techniques: tree-reweighted and counting number entropies. Our insights for the latter suggest a novel counting number optimization framework, which guarantees strong convexity for any given modulus. Our experiments demonstrate that learning with a strongly convex free energy, using our optimization framework to guarantee a given modulus, results in substantially more accurate marginal probabilities, thereby validating our theoretical claims and the effectiveness of our framework.

1. Introduction

Though marginal inference in general graphical models is an intractable problem, many approximations have been proposed using the *variational free energy*. Much of this research has focused on the convexity of the free energy. When it is convex, convergence to a global minimum is guaranteed. Less attention has been paid to when the free energy is *strongly* convex (i.e., has curvature), and what benefits this offers. In this work, we show that learning with a strongly convex free energy results in more accurate marginal probabilities. Our contributions include: a theoretical motivation for using strongly convex free energies, a framework for optimizing the strength of convexity in many variational methods, and experimental evaluation.

We frame our theoretical analysis in stability, which measures the inference algorithm's robustness to perturbation. One way to characterize stability is the *Lipschitz gradient* condition (Hiriart-Urruty & Lemaréchal, 2001), which is the dual of strong convexity. Using this duality and the variational form of the log-partition function, we show that strongly convex free energies result in more stable marginals. Further, we argue that a simply convex free energy cannot satisfy this stability guarantee. Using our stability result, we prove an error bound for the marginals of a model that is learned using strongly convex variational inference. The error bound is inversely proportional to the *modulus* of convexity (i.e., amount of curvature) of the free energy, thereby highlighting an important consideration for strongly convex free energies: the modulus should be constant with respect to the size of the graph, $|G|$.

Based on the above insights, we aim to identify free energies that are strongly convex, and when their respective moduli of convexity are constant with respect to $|G|$. We consider two popular variational methods: tree-reweighted (Wainwright et al., 2005) and counting number (Heskes, 2006) entropies. Using the notion of *contraction*, we give model-dependent conditions under which the negative tree-reweighted entropy is $\Omega(1)$-strongly convex. We then propose new sufficient conditions to characterize the modulus of convexity for counting number entropies. We use this to derive a novel counting number optimization that yields κ-strongly convex free energies, for any $\kappa > 0$, independent of the model parameters. This optimization can "strongly convexify" any entropy approximation that can be expressed via counting numbers, which includes many

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

used in practice (e.g., Bethe and tree-reweighted).

We demonstrate the practical impact of our theory in a set of experiments on challenging grid-structured models. Our empirical results suggest that strongly convex free energies can dramatically improve the quality of marginal inference, and that our counting number optimization reduces the error of learned marginals by over 40%. These findings indicate that having a tunable modulus can offer substantial benefit in practice.

1.1. Related Work

Our theoretical motivation, which connects strong convexity and stability to error bounds, is primarily related to two previous studies. Wainwright (2006) argued that, when approximate inference is necessary, using an inconsistent M-estimator is sometimes better than using the true model. He motivated this conclusion with an error bound that leverages the stability of strongly convex variational inference. Whereas Wainwright's bound is asymptotic, and tied to a specific generative process, our error bound is general and holds with high probability over draws of a finite training set. Moreover, our results are more explicit about the role of the modulus of convexity, which highlights the importance of it being independent of the graph size. The other related line of work is from London et al. (2013; 2014), who derived PAC learning bounds for structured prediction. Their bounds crucially rely on a form of "collective" stability that is guaranteed by L^1 strongly convex free energies. We distinguish our error bound from these by our proof technique, which uses L^2 strong convexity, and that we are interested in learning accurate marginals, not just maximizing the marginal probability of the correct label.

The study of convex free energies in approximate inference has a long history. Approaches can be broadly categorized by their approximation of the negative entropy term.[1] Wainwright et al.'s (2005) tree-reweighted approximation decomposes the entropy into a convex combination of tree entropies, each of which is convex. Wainwright (2006) later showed that this approximation is in fact strongly convex, though his lower bound on the modulus decreases as a function of the size of the graph. Another decomposition approach, due to Globerson & Jaakkola (2007), replaces the entropy with a sum of conditional entropies. This approximation is provably convex, but not strongly convex. Heskes (2006) proposed general sufficient conditions, based on counting (or, "over-counting") numbers, to establish the convexity of the Bethe and Kikuchi approximations. This work inspired a wave of research in counting number-based approximations (e.g., Weiss et al., 2007; Hazan & Shashua, 2008; Meltzer et al., 2009; Meshi et al., 2009). Hazan & Shashua (2008) used a slight modification of Heskes's conditions to guarantee strict convexity, which guarantees a unique global minimum, but does not identify a modulus. To our knowledge, our sufficient conditions are the first to identify when the counting number entropy is strongly convex, with a known modulus.

2. Background and Notation

We first introduce notation and review some concepts that will be used in our analysis. We consider the following class of Markov random fields (MRFs). Let $\mathcal{Y} \triangleq \{\mathbf{e}_1, \ldots, \mathbf{e}_\ell\}$ denote a set of ℓ labels, represented by the ℓ-dimensional standard basis (a.k.a. "one-hot") vectors. Let $\mathbf{Y} \triangleq (Y_1, \ldots, Y_n)$ denote a set of random variables, each with domain $\mathcal{Y}$. Let $G \triangleq (\mathcal{V}, \mathcal{E})$ denote an undirected graph, whose edges correspond to interactions between variables. We refer to $|G| \triangleq |\mathcal{V}| + |\mathcal{E}|$ as the *size* of the graph. The model is parameterized by a set of *potential functions*, organized according to the nodes and edges of G. Given an assignment, $\mathbf{y} \in \mathcal{Y}^n$, let $\theta_v(y_v)$ denote the potential for node $v \in \mathcal{V}$ being in state $y_v \in \mathcal{Y}$, and let $\theta_e(y_e)$ denote the potential for edge $e = \{u, v\} \in \mathcal{E}$ being in state $y_e = y_u \otimes y_v$. Since y_v and y_e are standard basis vectors, we can represent the potentials as vectors, such that $\theta_v(y_v) = \theta_v \cdot y_v$ and $\theta_e(y_e) = \theta_e \cdot y_e$. With

$$\boldsymbol{\theta} \triangleq ((\theta_v)_{v\in\mathcal{V}}, (\theta_e)_{e\in\mathcal{E}}) \quad \text{and} \quad \hat{\mathbf{y}} \triangleq ((y_v)_{v\in\mathcal{V}}, (y_e)_{e\in\mathcal{E}}),$$

we can then express the aggregate potential for $\mathbf{y}$ as a dot product, $\boldsymbol{\theta} \cdot \hat{\mathbf{y}} = \sum_{v\in\mathcal{V}} \theta_v(y_v) + \sum_{e\in\mathcal{E}} \theta_e(y_e)$. This describes a log-linear distribution,

$$p(\mathbf{Y} = \mathbf{y}; \boldsymbol{\theta}) \triangleq \exp\big(\boldsymbol{\theta} \cdot \hat{\mathbf{y}} - \Phi(\boldsymbol{\theta})\big),$$

where $\Phi(\boldsymbol{\theta}) \triangleq \log \sum_{\mathbf{y}'} \exp(\boldsymbol{\theta} \cdot \hat{\mathbf{y}}')$ is a normalizing function known as the *log-partition* function.

The log-partition is convex in $\boldsymbol{\theta}$, and has a well-known variational form (Wainwright & Jordan, 2008), $\Phi(\boldsymbol{\theta}) = \max_{\boldsymbol{\mu}\in\mathcal{M}} \boldsymbol{\theta} \cdot \boldsymbol{\mu} - \Phi^*(\boldsymbol{\mu})$, where $\mathcal{M}$ is the *marginal polytope*—the set of all consistent marginal vectors—and Φ^* is the *convex conjugate* of Φ. In the model we consider, $\Phi^*(\boldsymbol{\mu})$ is equal to the negative entropy of the distribution consistent with marginals $\boldsymbol{\mu}$.[2] The negative of the quantity being maximized is often referred to as the *free energy*, $E(\boldsymbol{\mu}; \boldsymbol{\theta}) \triangleq -\boldsymbol{\theta} \cdot \boldsymbol{\mu} + \Phi^*(\boldsymbol{\mu})$. The gradient of $\Phi(\boldsymbol{\theta})$ is the maximizing $\boldsymbol{\mu}$ (i.e., minimizer of E), which corresponds to the marginal distributions of $Y_1, \ldots, Y_n$. We denote this by

$$\boldsymbol{\mu}(\boldsymbol{\theta}) \triangleq \operatorname*{arg\,min}_{\boldsymbol{\mu}\in\mathcal{M}} E(\boldsymbol{\mu}; \boldsymbol{\theta}) = \nabla\Phi(\boldsymbol{\theta}).$$

Unfortunately, for general graph structures, $\mathcal{M}$ may require an exponential number of constraints, and Φ^* may

[1] Since most of these approximations use the same local relaxation of the marginal polytope, we focus on the entropy.

[2] See Wainwright & Jordan (2008) for a precise definition.

lack an explicit form. Many variational methods address these problems by relaxing $\mathcal{M}$ to an outer bound that uses a polynomial number of "local" constraints, and replacing Φ^* with a tractable approximation, $\tilde{\Phi}^*$. The *local* marginal polytope, $\tilde{\mathcal{M}} \supseteq \mathcal{M}$, is typically defined as follows:

$$\tilde{\mathcal{M}} \triangleq \left\{ \tilde{\boldsymbol{\mu}} : \begin{array}{c} \forall v \in \mathcal{V},\ \sum_{j=1}^{\ell} \tilde{\mu}_v^j = 1\ ; \\ \forall e \in \mathcal{E}, \forall v \in e,\ \sum_{i=1}^{\ell} \tilde{\mu}_e^{ij} = \tilde{\mu}_v^j \end{array} \right\}.$$

We call each $\tilde{\boldsymbol{\mu}} \in \tilde{\mathcal{M}}$ a set of *pseudomarginals*. With a slight abuse of notation, let $\tilde{E}(\tilde{\boldsymbol{\mu}};\boldsymbol{\theta}) \triangleq -\boldsymbol{\theta} \cdot \tilde{\boldsymbol{\mu}} + \tilde{\Phi}^*(\tilde{\boldsymbol{\mu}})$ denote a variational free energy for $\tilde{\Phi}^*$ and $\tilde{\mathcal{M}}$, let $\tilde{\Phi}(\boldsymbol{\theta}) \triangleq \max_{\tilde{\boldsymbol{\mu}} \in \tilde{\mathcal{M}}} -\tilde{E}(\tilde{\boldsymbol{\mu}};\boldsymbol{\theta})$ denote the convex conjugate of $\tilde{\Phi}^*$ (i.e., the approximate log-partition), and let

$$\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}) \triangleq \operatorname*{arg\,min}_{\tilde{\boldsymbol{\mu}} \in \tilde{\mathcal{M}}} \tilde{E}(\tilde{\boldsymbol{\mu}};\boldsymbol{\theta}) = \nabla \tilde{\Phi}(\boldsymbol{\theta})$$

denote the pseudomarginals of the variational distribution,

$$\tilde{p}(\mathbf{Y} = \mathbf{y};\boldsymbol{\theta}) \triangleq \exp\big(\boldsymbol{\theta} \cdot \hat{\mathbf{y}} - \tilde{\Phi}(\boldsymbol{\theta})\big).$$

It is common in structured prediction to condition the distribution of $\mathbf{Y}$ on some observed variables, $\mathbf{X}$. Evidence $\mathbf{X} = \mathbf{x}$ is incorporated into the potential functions, so that $p\left(\mathbf{Y} = \mathbf{y} \,|\, \mathbf{X} = \mathbf{x};\boldsymbol{\theta}\right) \triangleq \exp\left(\boldsymbol{\theta}(\mathbf{x}) \cdot \hat{\mathbf{y}} - \Phi(\boldsymbol{\theta}(\mathbf{x}))\right)$. For simplicity of exposition, we will not discuss conditional distributions, though most of our analysis also holds for conditional distributions with small modifications.

3. A Case for Strong Convexity

Because the dot product is linear, the convexity of the free energy is determined by the convexity of the conjugate function, Φ^*, or $\tilde{\Phi}^*$ for approximations. Some approximations are known to be convex, yet few studies discuss the *strength* of convexity, by which we mean the following.

Definition 1. A differentiable function, $\varphi : \mathcal{S} \to \mathbb{R}$, of a convex set, $\mathcal{S}$, is κ*-strongly convex* w.r.t. a norm[3], $\|\cdot\|$, if and only if, for all $s, s' \in \mathcal{S}$,

$$\frac{\kappa}{2} \left\| s - s' \right\|^2 + \langle \nabla\varphi(s),\, s' - s \rangle \le \varphi(s') - \varphi(s). \tag{1}$$

The *modulus* of convexity, κ, measures the curvature of φ.

The true conjugate function, Φ^*, is a strongly convex function of the full probability table. Since the marginals are a linear function of the probability table, Φ^* is also a strongly convex function of $\mathcal{M}$—albeit with an unknown modulus. Approximations of Φ^* that are simply convex ignore this fact, and may result in less accurate marginals.

The purpose of this section is to motivate the use of strongly convex free energies. We start by connecting strong convexity to stability, showing that strong convexity is both sufficient (Section 3.1) and necessary (Section 3.2) for uniform stability, which can be used to derive bounds on the quality of learned marginals (Section 3.3). More importantly, the theory suggests that the modulus of convexity is crucial, and that one should prefer moduli that are independent of the size of the graph (Section 3.4). Proofs from this section are deferred to Appendix B.

3.1. Strong Convexity Guarantees Stability

There is a well-known duality between strong convexity and the *Lipschitz continuity* of the gradient.

Definition 2. A differentiable function, $\varphi : \mathcal{S} \to \mathbb{R}$, has a λ*-Lipschitz continuous gradient* if and only if, for all $s, s' \in \mathcal{S}$,

$$\left\| \nabla\varphi(s) - \nabla\varphi(s') \right\|_2 \le \lambda \left\| s - s' \right\|_2. \tag{2}$$

Lemma 1 (Hiriart-Urruty & Lemaréchal, 2001, Theorem 4.2.1)**.** *Let $\varphi : \mathcal{S} \to \mathbb{R}$ denote a differentiable function, and $\varphi^\star : \mathcal{S}^\star \to \mathbb{R}$ its convex conjugate. If $\varphi^\star$ is κ-strongly convex, then φ has a $(1/\kappa)$-Lipschitz continuous gradient.*

Since the gradient of $\tilde{\Phi}$ corresponds to the pseudomarginals of the distribution, a strongly convex conjugate function lets us bound the stability of approximate marginal inference. This is summarized in the following lemma.[4]

Lemma 2. *Assume that $\tilde{E}$ uses a κ-strongly convex conjugate function, $\tilde{\Phi}^*$. Then, for any $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$,*

$$\frac{1}{\sqrt{|G|}} \left\| \tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}) - \tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}') \right\|_2 \le \frac{1}{\kappa\sqrt{|G|}} \left\| \boldsymbol{\theta} - \boldsymbol{\theta}' \right\|_2. \tag{3}$$

Lemma 2 upper-bounds the root-mean-squared difference between the respective pseudomarginals of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. Observe that one can trivially upper-bound this quantity by $\sqrt{2}$ by assuming that the marginals are completely different. In contrast, the right-hand side of Eq. 3 shrinks as a function of the size of the graph, $|G|$, and the L^2 distance between the potentials, $\left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|_2$, provided κ is lower-bounded by a function that is independent of these terms. Of course, since the potentials have length $\mathrm{O}(|G|)$, their L^2 distance could be $\mathrm{O}(\sqrt{|G|})$; but there are some cases in which the distance could be small. In Section 3.3, we discuss one such scenario and use it to derive a bound on the root-mean-squared error (RMSE) of learned pseudomarginals.

3.2. Convexity Alone Does Not Guarantee Stability

Strong convexity is central to Lemma 2. In fact, there is good reason to believe that strong convexity is a *necessary* condition for *uniform* stability. To understand why, we return to the relationship between strong convexity and Lipschitz gradients. Lemma 1 states that the former property implies the latter; however, the converse is also true.

[3] Unless specified, assume strong convexity w.r.t. the 2-norm.

[4] Wainwright derived a similar result (2006, Lemma 6). Our lemma is more explicit about the role of the modulus of convexity.

Lemma 3 (Hiriart-Urruty & Lemaréchal, 2001, Theorem 4.2.2)**.** *Let $\varphi : \mathcal{S} \to \mathbb{R}$ denote a differentiable function, and $\varphi^\star : \mathcal{S}^\star \to \mathbb{R}$ its convex conjugate. If φ has a λ-Lipschitz continuous gradient, then $\varphi^\star$ is $(1/\lambda)$-strongly convex.*

This establishes an equivalence between strong convexity and Lipshitz gradients: φ has a $(1/\kappa)$-Lipschitz continuous gradient *if and only if* $\varphi^\star$ is κ-strongly convex. In the context of variational inference, this means that Eq. 3 holds if and only if $\tilde{\Phi}^*$ is strongly convex. Mere convexity (i.e., $\kappa = 0$) is insufficient for guaranteeing stability. In fact, for any simply convex $\tilde{\Phi}^*$, it may be possible to construct an example in which marginal inference is not stable.

For instance, consider the extreme case in which $\tilde{\Phi}^*$ is linear in $\tilde{\mathcal{M}}$. This means that $\tilde{E}$ is also linear. Mangasarian & Shiau (1987) prove by counterexample that solutions to linear programs are not *Lipschitz continuous* (a form of stability) with respect to perturbations in the objective coefficients (in this case, the potentials). Therefore, inference with a linear conjugate function cannot have non-trivial uniform stability.

3.3. Stability Yields Learning Guarantees

Eq. 3 is especially meaningful in the context of learning. Suppose we are trying to learn a distribution, $p(\mathbf{Y}; \boldsymbol{\theta}^\star)$, parameterized by some potentials, $\boldsymbol{\theta}^\star$. We assume that the class of models to which $\boldsymbol{\theta}^\star$ belongs is known, and that the variable interactions, defined by a graph G, are fixed. Our goal is to estimate $\boldsymbol{\theta}^\star$ given m independent draws from the distribution, $(\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(m)})$. To do so, we minimize the negative log-likelihood (NLL) of the variational distribution, $\tilde{p}$, induced by an approximate log-partition, $\tilde{\Phi}$. The approximation is for efficiency, since we make repeated evaluations of the objective during learning. Assume that $\tilde{\Phi}^*$, the convex conjugate of $\tilde{\Phi}$, is κ-strongly convex. Let $\mathcal{L}(\mathbf{Y}; \boldsymbol{\theta}) \triangleq -\ln \tilde{p}(\mathbf{Y}; \boldsymbol{\theta})$ denote the NLL under $\tilde{p}$, and let

$$\mathcal{L}_m(\boldsymbol{\theta}) \triangleq \frac{1}{m} \sum_{j=1}^{m} \mathcal{L}(\mathbf{y}^{(j)}; \boldsymbol{\theta}). \tag{4}$$

Let

$$\bar{\boldsymbol{\theta}} \triangleq \arg\min_{\boldsymbol{\theta}} \mathbb{E}\left[\mathcal{L}(\mathbf{Y}; \boldsymbol{\theta})\right], \tag{5}$$

$$\text{and} \qquad \hat{\boldsymbol{\theta}}_m \triangleq \arg\min_{\boldsymbol{\theta}} \mathcal{L}_m(\boldsymbol{\theta}) + \Lambda_m \left\|\boldsymbol{\theta}\right\|_2^2. \tag{6}$$

If $\Lambda_m \to 0$ as $m \to \infty$, then $\bar{\boldsymbol{\theta}} = \lim_{m\to\infty} \hat{\boldsymbol{\theta}}_m$.

Because $\mathcal{L}_m$ uses the approximate log-partition, $\hat{\boldsymbol{\theta}}_m$ is not a *consistent* estimator. In other words, in the limit of infinite data, $\hat{\boldsymbol{\theta}}_m$ may be different from $\boldsymbol{\theta}^\star$. Nonetheless, we have that $\boldsymbol{\mu}(\boldsymbol{\theta}^\star) = \tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{\theta}})$, as shown in Appendix B.2. In light of this, substituting $\hat{\boldsymbol{\theta}}_m$ and $\bar{\boldsymbol{\theta}}$ into Eq. 3, we have that the RMSE of the learned marginals, $\tilde{\boldsymbol{\mu}}(\hat{\boldsymbol{\theta}}_m)$, with respect to the true marginals, $\boldsymbol{\mu}(\boldsymbol{\theta}^\star)$, is proportional to the distance between $\hat{\boldsymbol{\theta}}_m$ and $\bar{\boldsymbol{\theta}}$, divided by the modulus of convexity, κ. As $\hat{\boldsymbol{\theta}}_m$ converges to $\bar{\boldsymbol{\theta}}$, the RMSE decreases at a rate that is inversely proportional to κ.

Convergence of *M-estimators* has been studied extensively. Many of these works (e.g., Bickel et al., 2009; Kakade et al., 2010; Ravikumar et al., 2011; Negahban et al., 2012; Bradley & Guestrin, 2012; Meng et al., 2014) rely on a *restricted eigenvalue (RE)* assumption. Essentially, this assumes that the eigenvalues of $\nabla^2 \mathcal{L}(\,\cdot\,; \boldsymbol{\theta})$—which is independent of $\mathbf{Y}$, and therefore the same as $\nabla^2 \mathcal{L}_m(\boldsymbol{\theta})$—evaluated in the vicinity of $\bar{\boldsymbol{\theta}}$, are bounded away from zero; meaning, the NLL is strongly convex in a region around $\bar{\boldsymbol{\theta}}$. We will further assume that, with probability $\geq 1 - \delta$ over draws of the training set, both $\bar{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_m$ (which is a random variable) are contained in a convex set within which $\nabla^2 \mathcal{L}(\,\cdot\,; \boldsymbol{\theta})$ is positive definite, thereby implying that the NLL is strongly convex in this set. The minimum eigenvalue of the Hessian (hence, the modulus of convexity) may depend on δ, m and G, but should be bounded away from zero by a constant as $m \to \infty$. This requirement will always be met if $\nabla^2 \mathcal{L}(\,\cdot\,; \bar{\boldsymbol{\theta}})$ is positive definite.

Assumption 1. Assume that there exists a constant, $\bar{\gamma} > 0$, such that the minimum eigenvalue of $\nabla^2 \mathcal{L}(\,\cdot\,; \bar{\boldsymbol{\theta}})$ is at least $\bar{\gamma}$. Further, for any $\delta \in (0, 1)$ and $m \geq 1$, there exists a convex set, $\mathcal{S} \subseteq \mathbb{R}^{|\boldsymbol{\theta}|}$, encompassing both $\bar{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_m$, and a function, $\gamma(\delta, m, G) = \Omega(1)$, such that, with probability $\geq 1 - \delta$ over draws of m i.i.d. examples, the minimum eigenvalue of $\nabla^2 \mathcal{L}(\,\cdot\,; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{S}$ is at least $\gamma(\delta, m, G)$.

Combining Assumption 1 and Lemma 2, we can prove a high-probability error bound on the marginals of a model learned with strongly convex variational inference.

Proposition 1. *Let $\Lambda_m \triangleq 1/\sqrt{m}$. Assume that $\tilde{\Phi}^*$ is κ-strongly convex, that Assumption 1 holds, and that $\left\|\bar{\boldsymbol{\theta}}\right\|_\infty \leq 1$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - 2\delta$ over draws of m i.i.d. examples,*

$$\frac{\left\|\tilde{\boldsymbol{\mu}}(\hat{\boldsymbol{\theta}}_m) - \boldsymbol{\mu}(\boldsymbol{\theta}^\star)\right\|_2}{\sqrt{|G|}} \leq \frac{\ell\left(2 + \sqrt{\frac{1}{2}\ln\frac{2\ell^2|G|}{\delta}}\right)}{\kappa\,\gamma(\delta, m, G)\sqrt{m}}. \tag{7}$$

Like most error bounds, Eq. 7 has an inverse dependence on the square root of m, so the bound decreases as the training set grows. What is interesting about our bound is that it incorporates the modulus of convexity, κ, of the variational free energy. Because of the inverse dependence on κ, the bound tightens as κ grows. Note that the upper bound for $\left\|\bar{\boldsymbol{\theta}}\right\|_\infty$ can be replaced with any constant. We also note that Proposition 1 is easily adapted for the mean-absolute error (MAE), since the RMSE upper-bounds the MAE.

3.4. Prefer a Constant Modulus

Eqs. 3 and 7 have an inverse dependence on the modulus of convexity. We should therefore prefer higher values, leading to sharper bounds. However, stronger convexity might mean that the approximation is looser. For instance, one can trivially boost the modulus by scaling the conjugate function with a temperature parameter. This reduces the bounds, but creates a totally entropic distribution. One therefore wonders whether there is a "right" amount of convexity that trades off stability for marginal accuracy.

One criterion stands out: *the modulus should not have an inverse dependence on* $|G|$. This insight is the most important takeaway of this section. When learning large graphical models, it is usually the case that the number of examples is small relative to the size of the graph. In this setting, κ can have great impact. If $\kappa = \Omega(1/|G|)$, then the learning rate (Eq. 7) is $\tilde{\mathrm{O}}\left(|G|/\sqrt{m}\right)$, which is vacuous for $|G| > \sqrt{m}$. In contrast, if $\kappa = \Omega(1)$, then the learning rate is $\tilde{\mathrm{O}}\left(1/\sqrt{m}\right)$. This observation motivates the study of $\Omega(1)$-strongly convex free energies in the next section.

4. Strongly Convex Variational Inference

In light of Section 3.4, we would like to identify strongly convex free energies for which the modulus of convexity is lower-bounded by a function that does not decrease with $|G|$. In this section, we present new guarantees for two popular variational methods. First, we provide model-dependent conditions under which the tree-reweighted negative entropy is $\Omega(1)$-strongly convex (Section 4.1). To prove this result, we prove a similar claim for the negative entropy of a tree-structured model (given in Appendix C.1). We also analyze the class of counting number entropies (which subsumes tree-reweighting), proving an interesting relationship between the counting numbers and the modulus of convexity (Section 4.2). Using this insight, we then provide a counting number optimization that guarantees κ-strong convexity, for any $\kappa > 0$, independent of the model.

4.1. Tree-Reweighting

The tree-reweighted entropy (Wainwright et al., 2005) is a convex combination of tree entropies. In this section, we give conditions under which its modulus of convexity is lower-bounded by a function of the parameters and structural properties, independent of graph size.

Fix a graph, G, and let $\mathcal{T}(G)$ denote its spanning trees. For a tree $T \triangleq (\mathcal{V}, \mathcal{E}_T) \in \mathcal{T}(G)$, its entropy is given by

$$H_T(\tilde{\boldsymbol{\mu}}) \triangleq \sum_{v\in\mathcal{V}} (1-\deg(v)) H_v(\tilde{\mu}_v) + \sum_{e\in\mathcal{E}_T} H_e(\tilde{\mu}_e), \quad (8)$$

where $\deg(v)$ is the degree of node v, and $H_v(\tilde{\mu}_v) \triangleq -\sum_{j=1}^{\ell} \tilde{\mu}_v^j \log \tilde{\mu}_v^j$ and $H_e(\tilde{\mu}_e) \triangleq -\sum_{i,j=1}^{\ell} \tilde{\mu}_e^{ij} \log \tilde{\mu}_e^{ij}$ are the node and edge local entropies. (Eq. 8 is also the Bethe entropy.) For a distribution, ρ, over $\mathcal{T}(G)$, the tree-reweighted entropy is given by

$$\begin{aligned} H^{\mathrm{TR}}(\tilde{\boldsymbol{\mu}}) &\triangleq \sum_{T\in\mathcal{T}(G)} \rho(T)\, H_T(\tilde{\boldsymbol{\mu}}) \quad (9) \\ &= \sum_{v\in\mathcal{V}} \Big(1 - \sum_{e:v\in e} \rho(e)\Big) H_v(\mu_v) + \sum_{e\in\mathcal{E}} \rho(e) H_e(\mu_e). \end{aligned}$$

Wainwright (2006) showed that if each edge, $e \in \mathcal{E}$, has positive marginal probability, $\rho(e) > 0$ (i.e., e appears in at least one tree, T, with $\rho(T) > 0$), then $-H^{\mathrm{TR}}$ is at least $\Omega(1/|G|)$-strongly convex. Unfortunately, this modulus decreases as a function of the size of the graph. This is partly because Wainwright's analysis considers all models in the exponential family. Here, we prove a more optimistic lower bound for models that exhibit good *contraction*.

Definition 3. Fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, and potentials, $\boldsymbol{\theta}$, which induce a probability density, p. For any $(u, v) : \{u, v\} \in \mathcal{E}$, define the *contraction coefficient* as

$$\vartheta_{\boldsymbol{\theta}}(u, v) \triangleq \sup_{y, y' \in \mathcal{Y}} \left\| p\left(Y_u \,|\, Y_v = y; \boldsymbol{\theta}\right) - p\left(Y_u \,|\, Y_v = y'; \boldsymbol{\theta}\right) \right\|_{\mathrm{TV}}.$$

Denote the maximum of the contraction coefficients by

$$\vartheta^{\star}_{\boldsymbol{\theta}} \triangleq \sup_{(u,v):\{u,v\}\in\mathcal{E}} \vartheta_{\boldsymbol{\theta}}(u, v).$$

The contraction coefficients measure the dependence between adjacent variables in a graphical model. A contraction coefficient of 1 implies determinism, and 0 implies independence. In Appendix C.2, we describe an efficient procedure for computing the contraction coefficients in a tree-structured model.

Roughly speaking, the contraction coefficients are determined by the ratio of "local" signal to "relational" signal. If the local signal is strong, Y_v has little influence on Y_u. For models with a sufficiently high ratio of local-to-relational signal, dependence decays with graph distance at a geometric rate. In this case, one can show that $-H_T$ is $\Omega(1)$-strongly convex (see Appendix C.1). Using this result, we obtain the following.

Proposition 2. *Fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, with maximum degree independent of $|\mathcal{V}|$. Fix a distribution, ρ, over the spanning trees, $\mathcal{T}(G)$, such that there exists a constant, $C > 0 : \forall\, e \in \mathcal{E},\ \rho(e) \geq C$, that lower-bounds the edge probabilities. Let $\Theta \subseteq \mathbb{R}^{|\boldsymbol{\theta}|}$ denote the set of potentials such that each tree $T \in \mathcal{T}(G) : \rho(T) > 0$, with maximum degree Δ_T, has maximum contraction coefficient $\vartheta^{\star}_{\boldsymbol{\theta},T} \leq 1/\Delta_T$. Let $\tilde{\mathcal{M}}(\Theta) \triangleq \{\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ denote the set of pseudomarginals realizable under any $\boldsymbol{\theta} \in \Theta$. Then, $-H^{\mathrm{TR}}$ is $\Omega(1)$-strongly convex in $\tilde{\mathcal{M}}(\Theta)$.*

The proof is given in Appendix D.1. See Appendix D.2 for implications of Proposition 2 for a grid-graph model.

Proposition 2 guarantees $\Omega(1)$-strong convexity, but it still does not identify the modulus. Further, it is model-dependent, and may not hold for certain potentials. Therefore, applying Proposition 1 to tree-reweighted variational inference is only meaningful when learning in a constrained model space that admits good contraction. In the next section, we describe a technique to tune the modulus to any specified value, regardless of the model.

4.2. Counting Number Optimization

Counting number techniques decompose the entropy into a weighted sum of node and edge local entropies. For $\mathbf{c} \triangleq ((c_v)_{v\in\mathcal{V}}, (c_e)_{e\in\mathcal{E}})$, the counting number entropy is

$$H^{\mathbf{c}}(\tilde{\boldsymbol{\mu}}) \triangleq \sum_{v\in\mathcal{V}} c_v H_v(\tilde{\mu}_v) + \sum_{e\in\mathcal{E}} c_e H_e(\tilde{\mu}_e). \tag{10}$$

Note that $H^{\mathbf{c}}$ generalizes the Bethe entropy (Eq. 8), which is given by $c_v = 1 - \deg(v)$ and $c_e = 1$. We can also recreate the tree-reweighted entropy (Eq. 9) with $c_v = 1 - \sum_{e:v\in e} \rho(e)$ and $c_e = \rho(e)$. In this section, we show how to find counting numbers that preserve strong convexity, with a modulus that is lower-bounded by a given value.

Since $-H_v$ and $-H_e$ are convex, it is clear from Eq. 10 that $-H^{\mathbf{c}}$ is convex for nonnegative counting numbers. Heskes (2006) derived more sophisticated sufficient conditions for convexity by reparameterizing the counting numbers. Specifically, $-H^{\mathbf{c}}$ is convex if there exist nonnegative *auxiliary* counting numbers, $(\alpha_v \geq 0)_{v\in\mathcal{V}}$, $(\alpha_e \geq 0)_{e\in\mathcal{E}}$ and $(\alpha_{v,e} \geq 0)_{e\in\mathcal{E}, v\in e}$, such that

$$\forall v \in \mathcal{V}, \;\; c_v = \alpha_v - \sum_{e:v\in e} \alpha_{v,e}, \tag{11}$$

$$\text{and} \quad \forall e \in \mathcal{E}, \;\; c_e = \alpha_e + \sum_{v:v\in e} \alpha_{v,e}. \tag{12}$$

The effect of the auxiliary counting numbers, in particular, $\alpha_{v,e}$, is to shift weight between the regular counting numbers, c_v and c_e. Heskes' conditions mean that c_v can be negative and still guarantee convexity. We can further show that $-H^{\mathbf{c}}$ is *strongly* convex whenever α_e is uniformly lower-bounded; α_v and $\alpha_{v,e}$, however, are only required to be nonnegative.

Proposition 3. *Fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, and assume that every node is in at least one edge. If $\mathbf{c}$ satisfies Eqs. 11 and 12 for some $\kappa > 0$, $(\alpha_v \geq 0)_{v\in\mathcal{V}}$, $(\alpha_e \geq \kappa)_{e\in\mathcal{E}}$ and $(\alpha_{v,e} \geq 0)_{e\in\mathcal{E}, v\in e}$, then $-H^{\mathbf{c}}$, is $(\kappa/3)$-strongly convex.*

The proof is given in Appendix E.1.

Proposition 3 lets us characterize the strong convexity of a range of algorithms that optimize counting numbers. For example, observing that the Bethe approximation often outperformed tree-reweighting in practice, Meshi et al. (2009) proposed a "convexified" Bethe approximation. Their algorithm finds a set of counting numbers that best approximates the Bethe counting numbers, $\mathbf{c}^{\text{B}}$, while satisfying Heskes' convexity conditions (Eqs. 11 and 12). They also proposed incorporating a constraint that, for all $v \in \mathcal{V}$, $c_v + \sum_{e:v\in e} c_e = 1$; this ensures that the counting numbers are *variable-valid* for a fully factored (i.e., edgeless) model. Via Proposition 3, adding a constraint that $\alpha_e \geq 3\kappa$ ensures that the resulting negative entropy is κ-strongly convex. This yields the following constrained quadratic program (QP), which we refer to as the *strongly convexified Bethe* approximation:

$$\begin{aligned} \min_{\mathbf{c},\, \boldsymbol{\alpha}\geq\mathbf{0}} \quad & \|\mathbf{c} - \mathbf{c}^{\text{B}}\|_2^2 \\ \text{s.t.} \;\; & \forall v \in \mathcal{V}, \; c_v + \sum_{e:v\in e} \alpha_{v,e} \geq 0\,; \\ & \forall e \in \mathcal{E}, \; c_e - \sum_{v:v\in e} \alpha_{v,e} \geq 3\kappa\,; \\ & \forall v \in \mathcal{V}, \; c_v + \sum_{e:v\in e} c_e = 1. \end{aligned} \tag{13}$$

Note that Eq. 13 only depends on the graph structure; it is independent of the potentials. Thus, the QP only needs to be solved once, prior to learning, for each example in the training set. Moreover, examples that have the same structure can use the same counting numbers. Certain graphs, such as regular graphs, may admit an analytic solution to Eq. 13, thereby avoiding numerical optimization.

We can strongly convexify any desired counting numbers. For instance, Hazan & Shashua (2008) proposed a convex counting number optimization that encourages $c_e = 1$ uniformly. With a small modification to Eq. 13, we can make Hazan & Shashua's method strongly convex. We can also optimize the tree-reweighted entropy. Though $-H^{\text{TR}}$ is already $\Omega(1)$-strongly convex for certain models (per Proposition 2), it may be difficult to identify the modulus. By substituting the tree-reweighted counting numbers for $\mathbf{c}^{\text{B}}$ in the objective, we can ensure that $-H^{\text{TR}}$ is at least κ-strongly convex, for any given κ, independent of the model.

For certain graphs and values of κ, the variable validity constraint may make the optimization infeasible. In these cases, we propose switching to a slackened QP, described in Appendix E.2. This QP adds a free parameter, C, that trades off between fitting the target counts and satisfying variable validity. We explore this trade-off in Section 5.3.

5. Experiments

Our empirical evaluation tests the hypothesis that strongly convex free energies result in better learned marginals, as suggested by Proposition 1. Evaluations of approximate inference techniques typically use the *true* model to measure the discrepancy in the marginals. That is, given the model that generated the data, $\boldsymbol{\theta}^\star$, most studies measure $\|\boldsymbol{\mu}(\boldsymbol{\theta}^\star) - \tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}^\star)\|$, where $\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}^\star)$ uses the true model with approximate inference. While this isolates the quality of the approximation, it ignores the fact that approximate inference is typically used both at train and test time. It is therefore valuable to test the quality of the approximation using a model that is *learned* with said approximation. Wainwright (2006) called this "learning the 'wrong' graphical model," since the learned model may not converge to the true model. We prefer to call it "learning the 'right' graphical model for the 'wrong' inference," since it finds the best parameters for the given variational method. We therefore report scores for both true and learned models.

5.1. Data Generator

Our synthetic data generator is based on those used in prior work (e.g., Hazan & Shashua, 2008; Meshi et al., 2009) to evaluate approximate marginal inference. We generate data from an (8×8) non-toroidal grid-structured model, in which each node, v, is associated with a binary variable, $Y_v \in \{\mathbf{e}_1, \mathbf{e}_2\}$. The model is defined by the following process, for either "attractive" or "mixed" potentials. First, we fix $\omega_s > 0$ and $\omega_p > 0$. For each node, we flip a fair coin, $c_v \in \{\pm 1\}$, and let $w_v \triangleq \omega_s \, c_v \left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]$. If the model is "attractive," we uniformly set $w_e \triangleq \omega_p \operatorname{vec}\left(\left[\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right]\right)$, where $\operatorname{vec}(\,\cdot\,)$ converts a matrix to a vector; if "mixed," we flip another fair coin, $c_e \in \{\pm 1\}$, and set $w_e \triangleq \omega_p \, c_e \operatorname{vec}\left(\left[\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right]\right)$. To create local perturbations (i.e., evidence), we draw a uniformly random $x_v \sim \mathbb{U}[0,1]$ for each node. Given these, we let

$$\forall v, \; \theta_v \triangleq w_v \, x_v, \;\text{ and }\; \forall e = \{u, v\}, \; \theta_e \triangleq w_e \left(\frac{x_u + x_v}{2} \right),$$

and define the data distribution as

$$p\left(\mathbf{Y} = \mathbf{y}; \boldsymbol{\theta}\right) \triangleq \sum_{v \in \mathcal{V}} \theta_v \cdot y_v + \sum_{e \in \mathcal{E}} \theta_e \cdot (y_u \otimes y_v).$$

This is equivalent to an Ising model with field potentials $\theta_v \sim \mathbb{U}[-\omega_s, \omega_s]$, and interaction potentials $\theta_e \sim \mathbb{U}[0, \omega_p]$, for attractive, or $\theta_e \sim \mathbb{U}[-\omega_p, \omega_p]$, for mixed.

5.2. Experiment Design

We use four variational methods from the literature:

LBP: The Bethe approximation (i.e., "loopy" BP).

C-Bethe: Meshi et al.'s (2009) convexified Bethe, which is equivalent to Eq. 13 with $\kappa = 0$.

TRBP: Wainwright et al.'s (2005) tree-reweighted BP, with the tree distribution described in Appendix D.2.

C-Unif: Hazan & Shashua's (2008) convex counting number optimization, which prefers $c_e = 1$ uniformly.

Of the four, only the last three are guaranteed to be convex; LBP is not convex on a grid. TRBP is in fact strongly convex, though the true modulus depends on the model, and may be difficult to identify. Hazan & Shashua's method actually enforces *strict* convexity, but since the modulus can be arbitrarily close to zero, we consider it effectively just convex. We also compare strongly convexified versions of C-Bethe, TRBP and C-Unif, using our counting number optimization. This results in counting numbers that are provably κ-strongly convex, for a given $\kappa > 0$. We denote these versions by **SC-Bethe**, **SC-TRBP** and **SC-Unif**, respectively, and indicate the value of κ whenever relevant.

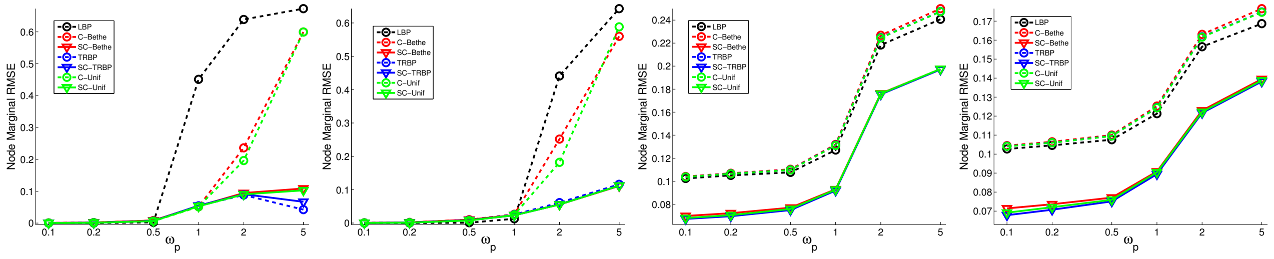
For each value of $\omega_s \in \{0.05, 1\}$ and $\omega_p \in \{0.1, 0.2, 0.5, 1, 2, 5\}$, we generate 20 models using the above synthetic generator. Each model acts as a learning trial. For each model, we compute the true marginal probabilities using exact (junction tree) inference and sample 100 joint assignments to $\mathbf{Y}$. We use these samples to train a model for each variational method (and value of κ), using L-BFGS to minimize the regularized NLL (Eq. 6). The regularization parameter, Λ_m, is set to $1/\sqrt{m}$, per Proposition 1. We then compute the node marginals using variational inference with the true (i.e., generating) and learned models. For each set of approximate marginals, we compute the root-mean-squared error (RMSE) with respect to the true, exact marginals. We report the average RMSE over 20 trials.

Our experiments are implemented in MATLAB, using data structures from Mark Schmidt's *Undirected Graphical Models* (UGM) toolkit (2013b). To optimize the learning objective, we use Schmidt's implementation of L-BFGS with Wolfe line search (2013a). For exact inference and sampling, we use UGM's junction tree implementation. For all variational inference algorithms, we use our own implementation of counting number belief propagation (CBP), based on Schwing et al.'s (2011) message updates; this can optimize any variational method whose entropy can be expressed with counting numbers. To optimize the counting number QP (Eq. 13, or Eq. 23 in Appendix E.2), we use MATLAB's `quadprog`, with the interior point method. To measure statistical significance, we use a paired t-test, with rejection threshold .05.

5.3. Results

Due to space restrictions, we defer the full catalog of figures to Appendix F. Figure 1 highlights select plots.

Strong Convexity Improves Marginal Inference. Figures 3a-d plot the RMSE of the node marginals as a function of the interaction parameter, ω_p. Inference is performed with the true model. The SC methods use the post hoc optimal value of κ (and C) in the counting number optimization. All methods perform about the same for $\omega_s = 1$



(a) Model, Attract, $\omega_s = .05$ (b) Model, Mixed, $\omega_s = .05$ (c) Learned, Attract, $\omega_s = 1$ (d) Learned, Mixed, $\omega_s = 1$

Figure 1. Select plots of RMSE (averaged over 20 trials) of the approximate node marginals w.r.t. the true marginals, as a function of the interaction parameter, ω_p. Data is generated with either "attractive" or "mixed" potentials. Figs. (a)-(b) use the true model for inference, and (c)-(d) use the learned model. The black dotted line is LBP; color dotted lines are the convex baselines, and solid lines are their SC counterparts, using the post hoc optimal value of κ (and C for $\kappa \geq .1$). See Section 5.3 for discussion and Appendix F for all figures.

and $\omega_p \leq 2$. LBP has a slight advantage for mixed potentials with $\omega_p \leq 1$, which concurs with previous conclusions (e.g., Meshi et al., 2009) that LBP performs well when there is strong local signal. Focusing on $\omega_s = .05$, the convex methods offer significant improvement over LBP for $\omega_p \geq 1$ with attractive and $\omega_p \geq 2$ with mixed potentials. This shows that convexity helps when there is low local-to-relational signal. In particular, we note that the strongly convex methods (TRBP and all SC variants) exhibit dramatically lower error in this setting (see Figures 1a-b), with over 10x improvement over LBP.

Strong Convexity Improves Learned Marginals. Figures 3e-h also plot RMSE as a function of ω_p, but using the learned model to compute the marginals. The SC methods yield statistically significant improvements in almost all data models. Figures 1c-d highlight the improvement, which is most prominent when $\omega_s = 1$. In certain cases, SC reduces the error of the convex baselines by over 40%. These results support the hypothesis of Proposition 1, that using a variational free energy that is provably $\Omega(1)$-strongly convex can significantly improve the quality of learned marginals. Moreover, the SC counting number optimization can even improve TRBP—which is already strongly convex, though the modulus is model-dependent.

Tuning κ in the SC Methods. The value of κ used in the SC counting number optimization can have great impact on the quality of the marginals. The theory in Section 3 suggests that increasing the modulus of convexity improves stability and marginal accuracy; however, altering κ affects the quality of the entropy approximation, hence, the marginals. Thus, there is a trade-off that needs to be explored. In Figures 4 and 5, we plot the RMSE of the marginals as a function of κ, using the true and learned models respectively, for select values of ω_s and ω_p. Since values of $\kappa \geq .1$ result in non-variable-valid counting numbers for this grid, we use the slackened QP and report the score for the post hoc optimal C. We learn the following from these plots. When the true potentials are given, and the model has low local-to-relational signal ($\omega_s = .05, \omega_p \geq 2$), any modulus of convexity above a certain threshold yields significant improvement. When using variational inference for training, if there is low local signal ($\omega_s = .05$), use the highest value of κ that supports variable validity. Since the local signal is weak, it is even more important to be variable-valid. If local signal is strong ($\omega_p = 1$), one can relax variable validity and push κ further.

Slackened Variable Validity. When using a value of κ that requires slackening variable validity, this requires selecting a value for the slack parameter, C. The quality of the slackened solution can vary with C, since this parameter controls the trade-off between variable validity and fitting the target counts. Figures 6 and 7 show select plots of RMSE as a function of C, focusing on the Bethe and tree-reweighted approximations. Data is generated using mixed potentials. In general, we find that the optimal value of C depends on κ, with lower values of κ favoring lower values of C. This is likely because lower C makes it easier for the QP solver to reduce the slack variables. When training with $\kappa \geq .1$, a good rule of thumb is to set C fairly high; we found that $C = 100$ works well overall.

6. Conclusion

We have shown, both theoretically and empirically, that variational inference with a strongly convex free energy can improve the accuracy of marginal probabilities. We proved sufficient conditions under which two popular variational methods are strongly convex, and proposed a novel counting number optimization that guarantees κ-strong convexity, for any κ. Our results indicate that using this approach to specify a modulus can dramatically reduce the error of approximate marginal inference, suggesting substantial, tangible benefit to applications of graphical models.

Acknowledgements

This work was supported by the National Science Foundation (NSF), under grant number IIS1218488, and by the Intelligence Advanced Research Projects Activity (IARPA), via Department of Interior National Business Center (DoI/NBC) contract number D12PC00337. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF, IARPA, DoI/NBC, or the U.S. Government.

References

Bickel, P., Ritov, Y., and Tsybakov, A. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4), 2009.

Bradley, J. and Guestrin, C. Sample complexity of composite likelihood. In *Artificial Intelligence and Statistics*, 2012.

Globerson, A. and Jaakkola, T. Approximate inference using conditional entropy decompositions. In *Artificial Intelligence and Statistics*, pp. 130–138, 2007.

Hazan, T. and Shashua, A. Convergent message-passing algorithms for inference over general graphs with convex free energies. In *Uncertainty in Artificial Intelligence*, 2008.

Heskes, T. Convexity arguments for efficient minimization of the Bethe and Kikuchi free energies. *Journal of Artificial Intelligence Research*, 26:153–190, 2006.

Hiriart-Urruty, J. and Lemaréchal, C. *Fundamentals of Convex Analysis*. Grundlehren Text Editions. Springer Berlin Heidelberg, 2001.

Kakade, S., Shamir, O., Sindharan, K., and Tewari, A. Learning exponential families in high-dimensions: Strong convexity and sparsity. In *Artificial Intelligence and Statistics*, 2010.

Kontorovich, A. Obtaining measure concentration from Markov contraction. *Markov Processes and Related Fields*, 18:613–638, 2012.

London, B., Huang, B., Taskar, B., and Getoor, L. Collective stability in structured prediction: Generalization from one example. In *Intl. Conference on Machine Learning*, 2013.

London, B., Huang, B., Taskar, B., and Getoor, L. PAC-Bayesian collective stability. In *Artificial Intelligence and Statistics*, 2014.

Mangasarian, O. and Shiau, T. Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems. *SIAM Journal on Control and Optimization*, 25(3):583–595, 1987.

Meltzer, T., Globerson, A., and Weiss, Y. Convergent message passing algorithms – a unifying view. In *Uncertainty in Artificial Intelligence*, 2009.

Meng, Z., Eriksson, B., and Hero III, A. Learning latent variable Gaussian graphical models. In *International Conference on Machine Learning*, pp. 1269–1277, 2014.

Meshi, O., Jaimovich, A., Globerson, A., and Friedman, N. Convexifying the Bethe free energy. In *Uncertainty in Artificial Intelligence*, 2009.

Negahban, S., Ravikumar, P., Wainwright, M., and Yu, B. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Ravikumar, P., Wainwright, M., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Schmidt, M. minFunc. `http://www.di.ens.fr/~mschmidt/Software/minFunc`, 2013a.

Schmidt, M. UGM: Matlab code for undirected graphical models. `http://www.di.ens.fr/~mschmidt/Software/UGM`, 2013b.

Schwing, A., Hazan, T., Pollefeys, M., and Urtasun, R. Distributed message passing for large scale graphical models. In *Computer Vision and Pattern Recognition*, 2011.

Shalev-Schwartz, S. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University of Jerusalem, 2007.

Wainwright, M. Estimating the "wrong" graphical model: Benefits in the computation-limited setting. *Journal of Machine Learning Research*, 7:1829–1859, 2006.

Wainwright, M. and Jordan, M. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., 2008.

Wainwright, M., Jaakkola, T., and Willsky, A. A new class of upper bounds on the log partition function. *IEEE Trans. on Information Theory*, 51(7):2313–2335, 2005.

Weiss, Y., Yanover, C., and Meltzer, T. MAP estimation, linear programming and belief propagation with convex free energies. In *Uncertainty in Artificial Intelligence*, 2007.

A. Properties of Strong Convexity

Strong convexity can be characterized in a number of ways. The following facts provide some conditions that are equivalent to Definition 1.

Fact 1. *A differentiable function, $\varphi : \mathcal{S} \to \mathbb{R}$, of a convex set, $\mathcal{S}$, is κ-strongly convex w.r.t. a norm, $\|\cdot\|$, if and only if, for all $s, s' \in \mathcal{S}$,*

$$\kappa \left\|s - s'\right\|^2 \leq \langle \nabla\varphi(s) - \nabla\varphi(s'),\, s - s' \rangle .$$

Fact 2. *A twice-differentiable function, $\varphi : \mathcal{S} \to \mathbb{R}$, of a convex set, $\mathcal{S}$, is κ-strongly convex w.r.t. a norm, $\|\cdot\|$, if and only if, for all $s, s' \in \mathcal{S}$,*

$$\kappa \left\|s\right\|^2 \leq \left\langle s, \nabla^2\varphi(s')\, s \right\rangle .$$

For the 2-norm, Fact 2 means that the minimum eigenvalue of the Hessian is lower-bounded by κ.

B. Proofs from Section 3

This section contains all deferred proofs from Section 3.

B.1. Proof of Stability Lemma (Lemma 2)

Recall that $\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}')$ are the gradients of $\tilde{\Phi}(\boldsymbol{\theta})$ and $\tilde{\Phi}(\boldsymbol{\theta}')$, respectively. Since the conjugate function, $\tilde{\Phi}^*$, is assumed to be κ-strongly convex, we have via Lemma 1 and Definition 2 that

$$\begin{aligned} \left\|\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}) - \tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}')\right\|_2 &= \left\|\nabla\tilde{\Phi}(\boldsymbol{\theta}) - \nabla\tilde{\Phi}(\boldsymbol{\theta}')\right\|_2 \\ &\leq \frac{1}{\kappa} \left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|_2 . \end{aligned} \tag{14}$$

Dividing both sides by $\sqrt{|G|}$ completes the proof.

B.2. The Marginals of the Expected NLL Minimizer are the True Marginals

Observe that $\hat{\boldsymbol{\theta}}_m$ effectively fits the empirical marginals of the dataset, $\frac{1}{m}\sum_{j=1}^{m} \hat{\mathbf{y}}^{(j)}$. Thus, as $m \to \infty$, the marginals induced by $\hat{\boldsymbol{\theta}}_m$ and $\boldsymbol{\theta}^\star$ converge. This is formalized in the following lemma.

Lemma 4. *Let $\boldsymbol{\mu}(\boldsymbol{\theta}^\star)$ denote the true marginals of a distribution. Let $\bar{\boldsymbol{\theta}}$ denote the minimizer of the expected NLL, per Eq. 5. Then,*

$$\boldsymbol{\mu}(\boldsymbol{\theta}^\star) = \tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{\theta}}).$$

Proof Expanding the expected NLL, we have

$$\mathbb{E}\left[-\ln\tilde{p}(\mathbf{Y};\boldsymbol{\theta})\right] = \tilde{\Phi}(\boldsymbol{\theta}) - \mathbb{E}[\boldsymbol{\theta}\cdot\hat{\mathbf{Y}}].$$

The gradient of this is

$$\nabla\,\mathbb{E}\left[-\ln\tilde{p}(\mathbf{Y};\boldsymbol{\theta})\right] = \tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}) - \mathbb{E}[\hat{\mathbf{Y}}] = \tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}) - \boldsymbol{\mu}(\boldsymbol{\theta}^\star).$$

Since the NLL is differentiable, the gradient at the minimum is zero. Thus, when $\nabla\,\mathbb{E}\left[-\ln\tilde{p}(\mathbf{Y};\bar{\boldsymbol{\theta}})\right] = 0$, we have $\tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{\theta}}) = \boldsymbol{\mu}(\boldsymbol{\theta}^\star)$. ■

B.3. Proof of Error Bound (Proposition 1)

By Lemma 4, $\boldsymbol{\mu}(\boldsymbol{\theta}^\star) = \tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{\theta}})$. Further, because $\tilde{\Phi}^*$ is assumed to be κ-strongly convex, using Lemma 2, we have that

$$\begin{aligned} \frac{1}{\sqrt{|G|}}\left\|\tilde{\boldsymbol{\mu}}(\hat{\boldsymbol{\theta}}_m) - \boldsymbol{\mu}(\boldsymbol{\theta}^\star)\right\|_2 &= \frac{1}{\sqrt{|G|}}\left\|\tilde{\boldsymbol{\mu}}(\hat{\boldsymbol{\theta}}_m) - \tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{\theta}})\right\|_2 \\ &\leq \frac{1}{\kappa\sqrt{|G|}}\left\|\hat{\boldsymbol{\theta}}_m - \bar{\boldsymbol{\theta}}\right\|_2 . \end{aligned} \tag{15}$$

The rest of the proof involves upper-bounding $\left\|\hat{\boldsymbol{\theta}}_m - \bar{\boldsymbol{\theta}}\right\|_2$.

Assumption 1 states that, with probability at least $1-\delta$, there exists a convex set, $\mathcal{S}$, encompassing $\bar{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_m$, such that the minimum eigenvalue of $\nabla^2\mathcal{L}(\,\cdot\,;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{S}$ is lower-bounded by $\gamma(\delta, m, G)$. By Fact 2, this event implies that the NLL is $\gamma(\delta, m, G)$-strongly convex in $\mathcal{S}$. Since $\nabla^2\mathcal{L}(\,\cdot\,;\boldsymbol{\theta}) = \nabla^2\mathcal{L}_m(\boldsymbol{\theta})$, the same can be said for $\mathcal{L}_m$, so the regularized NLL,

$$\mathcal{L}_m^{\mathrm{R}}(\boldsymbol{\theta}) \triangleq \mathcal{L}_m(\boldsymbol{\theta}) + \Lambda_m \left\|\boldsymbol{\theta}\right\|_2^2 ,$$

is also $\gamma(\delta, m, G)$-strongly convex in $\mathcal{S}$. Therefore, with probability at least $1-\delta$ over draws of m examples,

$$\begin{aligned} \left\|\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m\right\|_2^2 &\leq \frac{\left\langle \nabla\mathcal{L}_m^{\mathrm{R}}(\bar{\boldsymbol{\theta}}) - \nabla\mathcal{L}_m^{\mathrm{R}}(\hat{\boldsymbol{\theta}}_m),\, \bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m \right\rangle}{\gamma(\delta, m, G)} \\ &= \frac{\left\langle \nabla\mathcal{L}_m^{\mathrm{R}}(\bar{\boldsymbol{\theta}}),\, \bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m \right\rangle}{\gamma(\delta, m, G)} \\ &\leq \frac{\left\|\nabla\mathcal{L}_m^{\mathrm{R}}(\bar{\boldsymbol{\theta}})\right\|_2 \left\|\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m\right\|_2}{\gamma(\delta, m, G)} . \end{aligned}$$

The second line follows from the fact that $\hat{\boldsymbol{\theta}}_m$ is the minimizer of $\mathcal{L}_m^{\mathrm{R}}$, which is differentiable, so $\nabla\mathcal{L}_m^{\mathrm{R}}(\hat{\boldsymbol{\theta}}_m) = \mathbf{0}$. The last line uses Cauchy-Schwarz. Dividing both sides by $\left\|\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m\right\|_2$, and combining with Eq. 15, we have that, with probability at least $1-\delta$,

$$\frac{1}{\sqrt{|G|}}\left\|\tilde{\boldsymbol{\mu}}(\hat{\boldsymbol{\theta}}_m) - \boldsymbol{\mu}(\boldsymbol{\theta}^\star)\right\|_2 \leq \frac{\left\|\nabla\mathcal{L}_m^{\mathrm{R}}(\bar{\boldsymbol{\theta}})\right\|_2}{\kappa\,\gamma(\delta, m, G)\sqrt{|G|}} . \tag{16}$$

Using the triangle inequality, the norm of the gradient decomposes as

$$\begin{aligned} \left\|\nabla\mathcal{L}_m^{\mathrm{R}}(\bar{\boldsymbol{\theta}})\right\|_2 &= \left\|\nabla\mathcal{L}_m(\bar{\boldsymbol{\theta}}) + 2\Lambda_m\bar{\boldsymbol{\theta}}\right\|_2 \\ &\leq \left\|\nabla\mathcal{L}_m(\bar{\boldsymbol{\theta}})\right\|_2 + 2\Lambda_m\left\|\bar{\boldsymbol{\theta}}\right\|_2 . \end{aligned} \tag{17}$$

Let $N = \left|\bar{\boldsymbol{\theta}}\right|$, and note that $N = \ell \left|\mathcal{V}\right| + \ell^2 \left|\mathcal{E}\right| \leq \ell^2 \left|G\right|$. Therefore, using the definition of Λ_m, and leveraging the assumption that $\left\|\bar{\boldsymbol{\theta}}\right\|_\infty \leq 1$, we have that

$$2\Lambda_m \left\|\bar{\boldsymbol{\theta}}\right\|_2 \leq 2\sqrt{\frac{N}{m}} \left\|\bar{\boldsymbol{\theta}}\right\|_\infty \leq 2\ell\sqrt{\frac{|G|}{m}}. \qquad (18)$$

Turning now to the gradient of $\mathcal{L}_m$, we can expand Eq. 4 as

$$\mathcal{L}_m(\boldsymbol{\theta}) = \frac{1}{m}\sum_{j=1}^{m} \tilde{\Phi}(\boldsymbol{\theta}) - \boldsymbol{\theta} \cdot \hat{\mathbf{y}}^{(j)}.$$

Since $\tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{\theta}})$ is the gradient of $\tilde{\Phi}(\boldsymbol{\theta})$, and is in fact equal to the true marginals, $\boldsymbol{\mu}(\boldsymbol{\theta}^\star)$, we have that the gradient of $\mathcal{L}_m$ is

$$\begin{aligned}\nabla\mathcal{L}_m(\bar{\boldsymbol{\theta}}) &= \frac{1}{m}\sum_{j=1}^{m} \tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{\theta}}) - \hat{\mathbf{y}}^{(j)} \\ &= \boldsymbol{\mu}(\boldsymbol{\theta}^\star) - \frac{1}{m}\sum_{j=1}^{m} \hat{\mathbf{y}}^{(j)}.\end{aligned}$$

Note that the gradient is a zero-mean random vector; random because it depends on the draw of the training set. We will bound this quantity with high probability, using a technique borrowed from London et al. (2014).

It helps to denote the gradient by a vector, $\nabla\mathcal{L}_m(\bar{\boldsymbol{\theta}}) \triangleq \mathbf{g} \in \mathbb{R}^N$. Fix some value $\epsilon > 0$. For $\mathbf{g}$ to be greater than ϵ, at least one of its coordinates must have magnitude at least $\epsilon/\sqrt{N}$; otherwise, we would have

$$\left\|\mathbf{g}\right\|_2 = \sqrt{\sum_{i=1}^{N} \left|g_i\right|^2} < \sqrt{\sum_{i=1}^{N} \frac{\epsilon^2}{N}} = \epsilon.$$

Thus, using the union bound, we have that

$$\begin{aligned}\Pr\left\{\left\|\mathbf{g}\right\|_2 \geq \epsilon\right\} &\leq \Pr\left\{\exists i : \left|g_i\right| \geq \frac{\epsilon}{\sqrt{N}}\right\} \\ &\leq \sum_{i=1}^{N} \Pr\left\{\left|g_i\right| \geq \frac{\epsilon}{\sqrt{N}}\right\}.\end{aligned}$$

Each g_i is the difference of the mean and sample average of a sufficient statistic for some node variable Y_v (or edge variable Y_e) having label y_v (or y_e). The sufficient statistics are bounded in the interval $[0, 1]$, so $|g_i| \leq 1$. Moreover, the sample average is taken from m i.i.d. draws from the target distribution. Therefore, applying Hoeffding's inequality to each i, we have that

$$\Pr\left\{\left|g_i\right| \geq \frac{\epsilon}{\sqrt{N}}\right\} \leq 2\exp\left(\frac{-2m\epsilon^2}{N}\right).$$

Summing over $i = 1, \ldots, N$, we have

$$\Pr\left\{\left\|\mathbf{g}\right\|_2 \geq \epsilon\right\} \leq 2N\exp\left(\frac{-2m\epsilon^2}{N}\right).$$

Thus, with probability at least $1 - \delta$,

$$\left\|\nabla\mathcal{L}_m(\bar{\boldsymbol{\theta}})\right\|_2 \leq \sqrt{\frac{N \ln\frac{2N}{\delta}}{2m}} \leq \ell\sqrt{\frac{|G| \ln\frac{2\ell^2|G|}{\delta}}{2m}}. \qquad (19)$$

The last inequality uses the fact that $N \leq \ell^2 |G|$.

Substituting Eqs. 18 and 19 into Eq. 17, and rearranging the terms, we have that with probability at least $1 - \delta$,

$$\left\|\nabla\mathcal{L}^{\mathrm{R}}_m(\bar{\boldsymbol{\theta}})\right\|_2 \leq \ell\sqrt{\frac{|G|}{m}}\left(\sqrt{\frac{1}{2}\ln\frac{2\ell^2\left|G\right|}{\delta}} + 2\right).$$

Then, combining the above with Eq. 16, we have that with probability at least $1 - 2\delta$ over draws of the training set,

$$\frac{1}{\sqrt{|G|}}\left\|\tilde{\boldsymbol{\mu}}(\hat{\boldsymbol{\theta}}_m) - \boldsymbol{\mu}(\boldsymbol{\theta}^\star)\right\|_2 \leq \frac{\ell\left(\sqrt{\frac{1}{2}\ln\frac{2\ell^2|G|}{\delta}} + 2\right)}{\kappa\,\gamma(\delta, m, G)\sqrt{m}},$$

which completes the proof.

C. Tree-Structured Models

In this section, we analyze tree-structured models. We show that the negative entropy of a tree-structured model is strongly convex, with a modulus that depends on the contraction coefficients induced by the model. This result is used in the proof of Proposition 2. We also show how the contraction coefficients of a tree-structured model can be measured efficiently.

C.1. Strong Convexity of the Tree Negative Entropy

When the model is structured according to a tree, T, the marginal polytope, $\mathcal{M}$, is exactly equivalent to the local marginal polytope, $\tilde{\mathcal{M}}$. Further, its entropy function, H_T, can be expressed succinctly as a function of the marginals, using the Bethe entropy formula (see Eq. 8). Wainwright (2006) showed that $-H_T$ is $\Omega(1/\left|G\right|)$-strongly convex. This is a pessimistic lower bound, since it considers all models in the exponential family. Indeed, we can show that tree-structured models with good contraction (see Definition 3) and bounded degree induce a negative entropy that is $\Omega(1)$-strongly convex.

Proposition 4. *Fix a tree, T, with maximum degree $\Delta_T = \mathrm{O}(1)$, independent of $|\mathcal{V}|$. Let $\Theta \subseteq \mathbb{R}^{|\boldsymbol{\theta}|}$ denote the set of potentials with maximum contraction coefficient $\vartheta^\star_{\boldsymbol{\theta}} \leq 1/\Delta_T$, and let $\mathcal{M}(\Theta) \triangleq \{\boldsymbol{\mu}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ denote the corresponding set of realizable marginals. Then, the negative entropy, $-H_T$, is $\Omega(1)$-strongly convex in $\mathcal{M}(\Theta)$.*

Proof The Hessian of the log-partition, $\Phi(\boldsymbol{\theta})$, is the *covariance matrix*,

$$\Sigma(\mathbf{Y}; \boldsymbol{\theta}) \triangleq \mathbb{E}\left[\hat{\mathbf{y}}\hat{\mathbf{y}}^\top; \boldsymbol{\theta}\right] - \mathbb{E}\left[\hat{\mathbf{y}}; \boldsymbol{\theta}\right]\mathbb{E}\left[\hat{\mathbf{y}}^\top; \boldsymbol{\theta}\right],$$

where $\mathbb{E}[\,\cdot\,;\boldsymbol{\theta}]$ denotes an expectation over the distribution parameterized by $\boldsymbol{\theta}$. (For a derivation of this fact, see Wainwright & Jordan (2008).) Let $\Sigma^{-1}(\mathbf{Y};\boldsymbol{\theta})$ denote the inverse covariance (i.e., *precision*) matrix. Since Φ is the convex conjugate of the negative entropy, $-H$, the Hessian of one is the inverse Hessian of the other. This insight yields the following lemma.

Lemma 5. *The negative entropy, $-H$, is $(1/\lambda_{max})$-strongly convex in $\mathcal{M}(\Theta)$, where $\lambda_{max} \triangleq \max_{\boldsymbol{\theta}\in\Theta} \|\Sigma(\mathbf{Y};\boldsymbol{\theta})\|_2$ is the maximum eigenvalue of the covariance matrix, over all potentials in Θ.*

Proof Via Fact 2, $-H$ is κ-strongly convex in $\mathcal{M}(\Theta)$ if the eigenvalues of $\nabla^2\left(-H(\boldsymbol{\mu}(\boldsymbol{\theta}))\right)$, for every $\boldsymbol{\mu}(\boldsymbol{\theta}) \in \mathcal{M}(\Theta)$ (i.e., every $\boldsymbol{\theta} \in \Theta$), are bounded away from zero by κ. Via convex conjugacy,

$$\nabla^2\left(-H(\boldsymbol{\mu}(\boldsymbol{\theta}))\right) = \left(\nabla^2\Phi(\boldsymbol{\theta})\right)^{-1} = \Sigma^{-1}(\mathbf{Y};\boldsymbol{\theta}).$$

Therefore, the minimum eigenvalue of $-H(\boldsymbol{\mu}(\boldsymbol{\theta}))$ is equal to the maximum eigenvalue of $\Sigma(\mathbf{Y};\boldsymbol{\theta})$. ■

Thus, to lower-bound the convexity of $-H$, it suffices to uniformly upper-bound the spectral norm of $\Sigma(\mathbf{Y};\boldsymbol{\theta})$, over all $\boldsymbol{\theta} \in \Theta$. A simple way to do this (used by Wainwright, 2006) is to analyze the trace norm (i.e., sum of the diagonal), which upper-bounds the spectral norm. The diagonal elements of the covariance matrix are uniformly upper-bounded by $1/4$, since the sufficient statistics are bounded in $[0,1]$. This yields a (loose) upper bound of $\mathrm{O}(|G|)$. For our purposes, this bound is too loose, since it grows with the size of the graph.

A better approach is to analyze the induced 1-norm (i.e., maximum column sum) or ∞-norm (i.e., maximum row sum), which, for symmetric matrices, are equivalent, and conveniently upper-bound the spectral norm. (This is because $\|\mathbf{A}\|_2 \leq \sqrt{\|\mathbf{A}\|_1\|\mathbf{A}\|_\infty} = \sqrt{\|\mathbf{A}\|_1\|\mathbf{A}\|_1} = \|\mathbf{A}\|_1$.) Intuitively, the 1-norm of the covariance matrix captures the maximum dependence as a function of graph distance. To bound the 1-norm, we will relate each covariance coefficient to a product of contraction coefficients. For contraction less than 1—i.e., without determinism—this product will decrease geometrically with graph distance. This geometric series converges, provided the structure has bounded degree and sufficiently small contraction.

Our proof uses a technical lemma that is often credited to Dobrushin. We use a version of this given by Kontorovich (2012).

Lemma 6 (Kontorovich, 2012, Lemma 2.1)**.** *Let $\nu : \Omega \to \mathbb{R}$ be a signed,* balanced *measure, such that $\sum_{\omega\in\Omega}\nu(\omega) = 0$. Let $K : \Omega\times\Omega \to \mathbb{R}$ be a* Markov kernel*, where $K(\omega\,|\,\omega') \geq 0$, $\sum_\omega K(\omega\,|\,\omega') = 1$, and*

$$(K\nu)(\omega) \triangleq \sum_{\omega'\in\Omega} K(\omega\,|\,\omega')\,\nu(\omega').$$

Then

$$\begin{aligned}\|K\nu\|_{\mathrm{TV}} &= \sum_{\omega}\left|\sum_{\omega'} K(\omega\,|\,\omega')\,\nu(\omega')\right| \\ &\leq \vartheta\sum_{\omega'}|\nu(\omega')| \\ &= \vartheta\,\|\nu\|_{\mathrm{TV}}\,,\end{aligned}$$

where

$$\vartheta \triangleq \sup_{\omega,\omega'\in\Omega} \|K(\cdot\,|\,\omega) - K(\cdot\,|\,\omega')\|_{\mathrm{TV}}\,.$$

is the contraction coefficient of K.

Fix any $\boldsymbol{\theta} \in \Theta$. For the following, we use the shorthand $p_{\boldsymbol{\theta}}(y)$ to denote $p(Y = y;\boldsymbol{\theta})$, and similar probabilities. We also let $\sigma_{\boldsymbol{\theta}}(y_u, y_v)$ denote the entry of the covariance matrix corresponding to $Y_u = y_u$ and $Y_v = y_v$.

Let $\pi(1),\ldots,\pi(l)$ denote the sequence of nodes along a path. Note that π is the unique path connecting its end points, since the model is tree-structured. The covariance entries corresponding to $Y_{\pi(1)} = y_{\pi(1)}$ and $Y_{\pi(l)} = y_{\pi(l)}$ can be written recursively as

$$\begin{aligned}&\sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l)}) \\ &\quad= p_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l)}) - p_{\boldsymbol{\theta}}(y_{\pi(1)})p_{\boldsymbol{\theta}}(y_{\pi(l)}) \\ &\quad= \sum_{y_{\pi(l-1)}} p_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l-1)}, y_{\pi(l)}) \\ &\qquad\quad - p_{\boldsymbol{\theta}}(y_{\pi(1)})p_{\boldsymbol{\theta}}(y_{\pi(l-1)}, y_{\pi(l)}) \\ &\quad= \sum_{y_{\pi(l-1)}} p_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l-1)})p_{\boldsymbol{\theta}}(y_{\pi(l)}\,|\,y_{\pi(l-1)}) \\ &\qquad\quad - p_{\boldsymbol{\theta}}(y_{\pi(1)})p_{\boldsymbol{\theta}}(y_{\pi(l-1)})p_{\boldsymbol{\theta}}(y_{\pi(l)}\,|\,y_{\pi(l-1)}) \\ &\quad= \sum_{y_{\pi(l-1)}} p_{\boldsymbol{\theta}}(y_{\pi(l)}\,|\,y_{\pi(l-1)}) \\ &\qquad\quad \times\left(p_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l-1)}) - p_{\boldsymbol{\theta}}(y_{\pi(1)})p_{\boldsymbol{\theta}}(y_{\pi(l-1)})\right) \\ &\quad= \sum_{y_{\pi(l-1)}} p_{\boldsymbol{\theta}}(y_{\pi(l)}\,|\,y_{\pi(l-1)})\,\sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l-1)}).\end{aligned}$$

Note that the second equality follows from the Markov property; since $Y_{\pi(l)}$ is conditionally independent of $Y_{\pi(1)}$ given $Y_{\pi(l-1)}$, we have that $p_{\boldsymbol{\theta}}(y_{\pi(l)}\,|\,y_{\pi(l-1)}, y_{\pi(l)}) = p_{\boldsymbol{\theta}}(y_{\pi(l)}\,|\,y_{\pi(l-1)})$.

In the righthand expression, the conditional probability under $p_{\boldsymbol{\theta}}$ defines a Markov kernel. Moreover, the covariance with $y_{\pi(1)}$ defines a signed measure,

$$\nu(Y; y_{\pi(1)}) \triangleq \sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, Y),$$

which is balanced, since

$$\begin{aligned}\sum_y \nu(y; y_{\pi(1)}) &= \sum_y \sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, y) \\ &= \sum_y p_{\boldsymbol{\theta}}(y_{\pi(1)}, y) - p_{\boldsymbol{\theta}}(y_{\pi(1)}) p_{\boldsymbol{\theta}}(y) \\ &= p_{\boldsymbol{\theta}}(y_{\pi(1)}) - p_{\boldsymbol{\theta}}(y_{\pi(1)}) = 0.\end{aligned}$$

Therefore, via Lemma 6, we have that

$$\begin{aligned}&\sum_{y_{\pi(l)}} \left|\sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l)})\right| \\ &\quad = \sum_{y_{\pi(l)}} \left| \sum_{y_{\pi(l-1)}} p_{\boldsymbol{\theta}}(y_{\pi(l)} \,|\, y_{\pi(l-1)})\, \sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l-1)}) \right| \\ &\quad \leq \vartheta^\star_{\boldsymbol{\theta}} \sum_{y_{\pi(l-1)}} \left|\sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l-1)})\right|\end{aligned}$$

Applying this identity recursively, we have that

$$\begin{aligned}&\sum_{y_{\pi(l)}} \left|\sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l)})\right| \\ &\qquad \leq \vartheta^\star_{\boldsymbol{\theta}} \sum_{y_{\pi(l-1)}} \left|\sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(l-1)})\right| \\ &\qquad \vdots \\ &\qquad \leq (\vartheta^\star_{\boldsymbol{\theta}})^{l-2} \sum_{y_{\pi(2)}} \left|\sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(2)})\right| \\ &\qquad \leq (\vartheta^\star_{\boldsymbol{\theta}})^{l-1} \sum_{y'_{\pi(1)}} \left|\sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, y'_{\pi(1)})\right| \\ &\qquad \leq \frac{\ell}{4} (\vartheta^\star_{\boldsymbol{\theta}})^{l-1}.\end{aligned}$$

The last inequality follows from the fact that the covariance of any variable assignment is at most $1/4$ in magnitude, and the covariance between any two assignments to the same variable is also at most $1/4$.

Given an upper bound on the covariances of node assignments, we can bound the covariance of edge assignments. Consider edges $\{a, b\}, \{c, d\} \in \mathcal{E}$. Due to the tree structure, the edges lie at opposite ends of a unique path connecting their constituent nodes. Without loss of generality, assume that this path has the order $a, b, \ldots, c, d$, and that the length of the path from b to c is l. By the Markov property, Y_a and Y_d are conditionally independent given Y_b and Y_c. Thus, for any configuration $(Y_a, Y_b) = (y_a, y_b)$ and $(Y_c, Y_d) = (y_c, y_d)$, we have that

$$\begin{aligned}&\sum_{y_c, y_d} \left|\sigma_{\boldsymbol{\theta}}((y_a, y_b), (y_c, y_d))\right| \\ &\quad = \sum_{y_c, y_d} \left|p_{\boldsymbol{\theta}}(y_a, y_b, y_c, y_d) - p_{\boldsymbol{\theta}}(y_a, y_b) p_{\boldsymbol{\theta}}(y_c, y_d)\right| \\ &\quad = \sum_{y_c, y_d} \big|p_{\boldsymbol{\theta}}(y_a, y_d \,|\, y_b, y_c) p_{\boldsymbol{\theta}}(y_b, y_c) \\ &\qquad\quad - p_{\boldsymbol{\theta}}(y_a \,|\, y_b) p_{\boldsymbol{\theta}}(y_b) p_{\boldsymbol{\theta}}(y_d \,|\, y_c) p_{\boldsymbol{\theta}}(y_c)\big| \\ &\quad = \sum_{y_c, y_d} \big|p_{\boldsymbol{\theta}}(y_a \,|\, y_b) p_{\boldsymbol{\theta}}(y_d \,|\, y_c) p_{\boldsymbol{\theta}}(y_b, y_c) \\ &\qquad\quad - p_{\boldsymbol{\theta}}(y_a \,|\, y_b) p_{\boldsymbol{\theta}}(y_d \,|\, y_c) p_{\boldsymbol{\theta}}(y_b) p_{\boldsymbol{\theta}}(y_c)\big| \\ &\quad = \sum_{y_c, y_d} p_{\boldsymbol{\theta}}(y_a \,|\, y_b) p_{\boldsymbol{\theta}}(y_d \,|\, y_c) \left|\sigma_{\boldsymbol{\theta}}(y_b, y_c)\right| \\ &\quad = p_{\boldsymbol{\theta}}(y_a \,|\, y_b) \sum_{y_c} \left|\sigma_{\boldsymbol{\theta}}(y_b, y_c)\right| \sum_{y_d} p_{\boldsymbol{\theta}}(y_d \,|\, y_c) \\ &\quad = p_{\boldsymbol{\theta}}(y_a \,|\, y_b) \sum_{y_c} \left|\sigma_{\boldsymbol{\theta}}(y_b, y_c)\right| \\ &\quad \leq \frac{\ell}{4} (\vartheta^\star_{\boldsymbol{\theta}})^{l-1}.\end{aligned}$$

The same argument can be used to bound the covariance between node and edge variables, where the relevant path length l becomes the length from the node to the closest endpoint of the edge. The base case of covariance between a node or edge state indicator and another state is also at most $1/4$.

Thus far, we have derived upper bounds on the entries of the covariance matrix, which correspond to covariances between three types of pairs: node variables and node variables; node variables and edge variables; and edge variables and edge variables. For a distribution induced by a tree-structured model, with maximum degree Δ_T, the 1-norm of a column corresponding to a node assignment

$Y_u = y_u$ is

$$\begin{aligned}
&\sigma_{\boldsymbol{\theta}}(Y_u = y_u)\\
&\quad = \sum_{y'_u} |\sigma_{\boldsymbol{\theta}}(y_u, y'_u)| + \sum_{v \in \mathcal{V} \setminus u} \sum_{y_v} |\sigma_{\boldsymbol{\theta}}(y_u, y_v)|\\
&\qquad + \sum_{\{v,v'\} \in \mathcal{E}} \sum_{y_v, y_{v'}} |\sigma_{\boldsymbol{\theta}}(y_u, (y_v, y_{v'}))|\\
&\quad \leq \frac{\ell}{4} + \frac{\ell}{4} \sum_{v \in \mathcal{V} \setminus u} (\vartheta^{\star}_{\boldsymbol{\theta}})^{l(u,v)-1}\\
&\qquad + \frac{\ell}{4} \sum_{\{v,v'\} \in \mathcal{E}} (\vartheta^{\star}_{\boldsymbol{\theta}})^{\max\{0, \min\{l(u,v), l(u,v')\}-1\}}\\
&\quad \leq \frac{\ell}{4} + \frac{\ell}{4} \sum_{d=1}^{\infty} \Delta_T^{d} (\vartheta^{\star}_{\boldsymbol{\theta}})^{d-1}\\
&\qquad + \frac{\ell \Delta_T}{4} + \frac{\ell}{4} \sum_{d=1}^{\infty} \Delta_T^{d+1} (\vartheta^{\star}_{\boldsymbol{\theta}})^{d-1}\\
&\quad = \frac{\ell}{4} + \frac{\ell \Delta_T}{4} \sum_{d=1}^{\infty} (\Delta_T \, \vartheta^{\star}_{\boldsymbol{\theta}})^{d-1}\\
&\qquad + \frac{\ell \Delta_T}{4} + \frac{\ell \Delta_T^2}{4} \sum_{d=1}^{\infty} (\Delta_T \, \vartheta^{\star}_{\boldsymbol{\theta}})^{d-1}\\
&\quad = \frac{\ell}{4} + \frac{\ell \Delta_T}{4(1 - \Delta_T \, \vartheta^{\star}_{\boldsymbol{\theta}})} + \frac{\ell \Delta_T}{4} + \frac{\ell \Delta_T^2}{4(1 - \Delta_T \, \vartheta^{\star}_{\boldsymbol{\theta}})}.
\end{aligned}$$

where $l(u, v)$ is the length of the path from node u to v. The second inequality holds because the number of nodes at distance d is at most Δ_T^d, and the maximum number of edges with endpoints at distance d is at most Δ_T^{d+1}, where we adjust for node and edge variables at distance zero. The last line applies the geometric series identity, since $\Delta_T \, \vartheta^{\star}_{\boldsymbol{\theta}} < \Delta_T / \Delta_T = 1$. An analogous argument bounds the 1-norm of any column corresponding to an edge assignment.

Since the 1-norm of every column of the covariance matrix is upper-bounded independently of $|G|$, it follows that the induced 1-norm of $\Sigma(\mathbf{Y}; \boldsymbol{\theta})$ is bounded independently of $|G|$; that is,

$$\|\Sigma(\mathbf{Y}; \boldsymbol{\theta})\|_1 = \mathrm{O}(1).$$

This holds for every $\boldsymbol{\theta} \in \Theta$, though the constant may differ, depending on $\vartheta^{\star}_{\boldsymbol{\theta}}$. Recall that the 1-norm of the covariance matrix upper-bounds the spectral norm, since the covariance matrix is symmetric. Thus, the minimum eigenvalue of $\nabla^2 \left(-H(\boldsymbol{\mu}(\boldsymbol{\theta}))\right)$, for every $\boldsymbol{\mu}(\boldsymbol{\theta}) \in \mathcal{M}(\Theta)$, is lower-bounded by a constant, which means that the negative entropy is $\Omega(1)$-strongly convex in $\mathcal{M}(\Theta)$. ■

C.2. Measuring Contraction

In the previous section, we relate the convexity of $-H_T$ to the model's maximum contraction coefficient. For general graphical models, measuring the contraction coefficients may be intractable. However, when the model is tree-structured, there is an efficient algorithm.

For a tree-structured model, exact inference can be computed efficiently using message passing. Given the node and edge marginals, one can compute the conditional probabilities via

$$p\left(Y_u = y_u \,|\, Y_v = y_v \,; \boldsymbol{\theta}\right) = \frac{p\left(Y_u = y_u, Y_v = y_v \,; \boldsymbol{\theta}\right)}{p\left(Y_v = y_v \,; \boldsymbol{\theta}\right)}.$$

One can then compute the total variation distance; hence, the contraction coefficient. For variables with small domains (e.g., binary), this is efficient. Given the contraction coefficient for each $(u, v) : \{u, v\} \in \mathcal{E}$, computing the maximum contraction coefficient is trivial.

Note that marginal inference only needs to be computed once in this procedure. The time complexity of inference in a tree-structured model, with ℓ labels and $|\mathcal{E}|$ edges is $\mathrm{O}(\ell^2 \, |\mathcal{E}|)$. For each undirected edge, there are two contraction coefficients (one per direction), each of which involves ℓ^2 operations (ℓ additions to compute the total variation distance conditioned on Y_v; and ℓ values of Y_v to condition on to compute the supremum). Since there are $|\mathcal{E}|$ edges, the overall time complexity of computing the contraction coefficients is $\mathrm{O}(\ell^2 \, |\mathcal{E}|)$.

D. Tree-Reweighting

In this section, we prove Proposition 2, which gives a model-dependent lower bound on the modulus of convexity for the tree-reweighted negative entropy. We also explore the ramifications of Proposition 2 for a grid-structured model.

D.1. Proof of $-H^{\mathrm{TR}}$ Strong Convexity (Proposition 2)

The following lemma relates the convexity of $-H^{\mathrm{TR}}$ to the convexity of its constituent tree entropies, as well as the tree distribution.

Lemma 7. *(Wainwright, 2006, Appendix C) Fix a graph, $G \triangleq (\mathcal{V}, \mathcal{E})$, and a distribution, ρ, over the spanning trees, $\mathcal{T}(G)$, such that $\rho(e) > 0$ for all $e \in \mathcal{E}$. Let $\rho^{\star}_e \triangleq \min_{e \in \mathcal{E}} \rho(e)$ denote the minimum edge probability. Let $\kappa^{\star}_T$ denote the minimum convexity of $-H_T$ for any tree $T \in \mathcal{T}(G)$ with positive probability under ρ. Then the tree-reweighted negative entropy, $-H^{\mathrm{TR}}$, is $(\rho^{\star}_e \kappa^{\star}_T)$-strongly convex.*

Thus, to prove $\Omega(1)$-strong convexity, one must show that the minimum edge probability, $\rho^{\star}_e$, and the minimum tree convexity, $\kappa^\star_T$ are both lower-bounded by values that are independent of $|G|$.

In Proposition 2, we assume that $\rho^\star_e$ is lower-bounded by a positive constant, $C > 0$. Since H^{TR} can be defined using *any* distribution over spanning trees, it is usually possible to construct an edge distribution for which this holds. (An example for a grid is given in Appendix D.2.) Therefore, the real challenge is to show that $\kappa^\star_T = \Omega(1)$. For each $T \in \mathcal{T}(G)$, denote the set of admissible potentials by $\Theta_T \subseteq \mathbb{R}^{|\boldsymbol{\theta}|}$, where dimensions corresponding to edges that don't exist in T have unbounded range. Note that

$$\Theta = \bigcap_{T \in \mathcal{T}(G) : \rho(T) > 0} \Theta_T,$$

so

$$\tilde{\mathcal{M}}(\Theta) = \bigcap_{T \in \mathcal{T}(G) : \rho(T) > 0} \tilde{\mathcal{M}}(\Theta_T).$$

Let $\tilde{\mathcal{M}}_T(\Theta)$ denote the projection of $\tilde{\mathcal{M}}(\Theta)$ onto the subspace defined by the nodes and edges in T, and note that $\tilde{\mathcal{M}}_T(\Theta) \subseteq \tilde{\mathcal{M}}_T(\Theta_T)$. In Proposition 4, we showed that, under suitable structural and contraction conditions, $-H_T$ is $\Omega(1)$-strongly convex in $\tilde{\mathcal{M}}_T(\Theta_T)$; hence, in $\tilde{\mathcal{M}}_T(\Theta)$ as well. When combined with Lemma 7, with $\rho^\star_e > C$, this proves that $-H^{\text{TR}}$ is $\Omega(1)$-strongly convex in $\tilde{\mathcal{M}}(\Theta)$.

D.2. Example Tree-Reweighting for a Grid Graph

Suppose the model is structured according to an $m \times n$ grid. This graph can be covered using a set of 4 chains, using the "snake-like" pattern illustrated in Figure 2. Observe that each internal edge is covered by 2 chains, and each boundary edge is covered by 3 chains. Therefore, using a uniform distribution over the chains, we have that each internal edge, e, has probability $\rho(e) = 1/2$, and each boundary edge, e', has probability $\rho(e') = 3/4$.

To apply Proposition 2 to this spanning tree distribution, we take $C = 1/2$ as the minimum edge probability. The maximum degree of a chain is 2, so the maximum contraction coefficient, $\vartheta^\star_{\boldsymbol{\theta},T}$, must be at most $1/2$. It may be possible to upper-bound $\vartheta^\star_{\boldsymbol{\theta},T}$ analytically for all $\boldsymbol{\theta}$ in some space. Alternately, one could map out the space of feasible potentials by measuring $\vartheta^\star_{\boldsymbol{\theta},T}$, using the procedure from Appendix C.2.

E. Counting Numbers

In this section, we prove Proposition 3, which characterizes the modulus of convexity for counting number entropies. We also present a slackened version of the counting number QP, which can be used when the variable validity constraints are not satisfied.

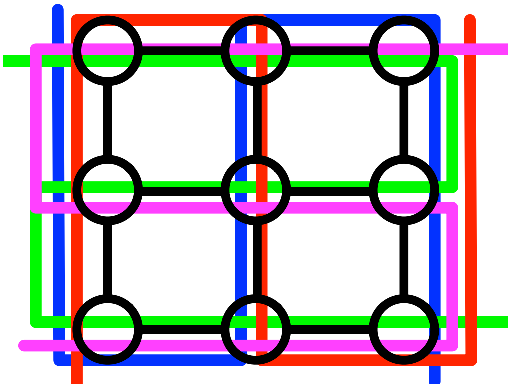


Figure 2. Covering the edges of a grid graph with 4 chains.

E.1. Proof of $-H^{\text{c}}$ Strong Convexity (Proposition 3)

The proof of Proposition 3 requires two technical lemmas.

Lemma 8 (Shalev-Schwartz, 2007, Lemma 16)**.** *The function $\varphi(\mathbf{z}) \triangleq \sum_i^d z_i \log z_i$ is 1-strongly convex in the probability simplex, $\{\mathbf{z} \in [0,1]^d : \|\mathbf{z}\|_1 = 1\}$, w.r.t. the 1-norm.*

Lemma 9 (Heskes, 2006, Lemma A.1)**.** *The difference of entropies, equivalent to the negative conditional entropy, $H_v(\tilde{\mu}_v) - H_e(\tilde{\mu}_e) = -H_{e|v}(\tilde{\mu}_e)$, for $v \in e$, is a convex function of $\tilde{\mu}_e$.*

We now prove Proposition 3.

Proof [Proposition 3] Every edge, e, is composed of exactly two nodes, $\{u, v\}$. By assumption, we have that $\alpha_e \geq \kappa > 0$. Therefore, we can shift $(2\kappa/3)$ weight from α_e to α_u and α_v without affecting the counting numbers or Heskes's convexity conditions. Let:

$$\begin{aligned} \forall e \in \mathcal{E},\ \ \tilde{\alpha}_e &\triangleq \alpha_e - \frac{2\kappa}{3}\,; \\ \forall (v,e) : v \in e,\ \ \tilde{\alpha}_{v,e} &\triangleq \alpha_{v,e} + \frac{\kappa}{3}\,; \\ \forall v \in \mathcal{V},\ \ \tilde{\alpha}_v &\triangleq \alpha_v + \sum_{e : v \in e} \frac{\kappa}{3}. \end{aligned}$$

Observe that the new auxiliary counts satisfy Eqs. 11 and 12:

$$\begin{aligned} \forall v \in \mathcal{V},\ c_v &= \alpha_v - \sum_{e : v \in e} \left(\alpha_{v,e} + \frac{\kappa}{3} - \frac{\kappa}{3} \right) \\ &= \tilde{\alpha}_v - \sum_{e : v \in e} \tilde{\alpha}_{v,e}\,; \qquad (20) \\ \forall e \in \mathcal{E},\ c_e &= \alpha_e + \sum_{v : v \in e} \left(\alpha_{v,e} + \frac{\kappa}{3} - \frac{\kappa}{3} \right) \\ &= \tilde{\alpha}_e + \sum_{v : v \in e} \tilde{\alpha}_{v,e}. \qquad (21) \end{aligned}$$

Now, every e has $\tilde{\alpha}_e \geq \kappa/3$. Further, because we assume that every node is involved in at least one edge, every v

has $\tilde{\alpha}_v \geq \kappa/3$. (We could extend Proposition 3 to arbitrary graphs by assuming that every isolated node has $c_v \geq \kappa/3$.)

Substituting Eqs. 20 and 21 into Eq. 10 and rearranging the terms, we obtain

$$\begin{aligned}
-H^{\mathbf{c}}(\tilde{\boldsymbol{\mu}}) &= -\sum_{v \in \mathcal{V}} \tilde{\alpha}_v H_v(\tilde{\mu}_v) - \sum_{e \in \mathcal{E}} \tilde{\alpha}_e H_e(\tilde{\mu}_e) \\
&\quad + \sum_{e \in \mathcal{E}} \sum_{v : v \in e} \tilde{\alpha}_{v,e} (H_v(\tilde{\mu}_v) - H_e(\tilde{\mu}_e)) \\
&= -\sum_{v \in \mathcal{V}} \tilde{\alpha}_v H_v(\tilde{\mu}_v) - \sum_{e \in \mathcal{E}} \tilde{\alpha}_e H_e(\tilde{\mu}_e) \\
&\quad - \sum_{e \in \mathcal{E}} \sum_{v : v \in e} \tilde{\alpha}_{v,e} H_{e|v}(\tilde{\mu}_e). \qquad (22)
\end{aligned}$$

We will analyze the entropy terms individually, using the gradient definition of (strong) convexity.

Fix any two vectors $\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mu}}' \in \tilde{\mathcal{M}}$, and let $\boldsymbol{\delta} \triangleq \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}'$. Recall that $\forall v$, $\|\tilde{\mu}_v\|_1 = \|\tilde{\mu}'_v\|_1 = 1$ and $\forall e$, $\|\tilde{\mu}_e\|_1 = \|\tilde{\mu}'_e\|_1 = 1$. Via Lemma 8, $-H_v$ and $-H_e$ are 1-strongly convex in the probability simplex with respect to the 1-norm. By Fact 1, this means that every node v satisfies,

$$\langle \nabla(-H_v(\tilde{\mu}_v)) - \nabla(-H_v(\tilde{\mu}'_v)),\, \delta_v \rangle \geq \|\delta_v\|_1^2 .$$

Therefore,

$$\begin{aligned}
\tilde{\alpha}_v \langle \nabla(-H_v(\tilde{\mu}_v)) - \nabla(-H_v(\tilde{\mu}'_v)),\, \delta_v \rangle &\geq \tilde{\alpha}_v \|\delta_v\|_1^2 \\
&\geq \tilde{\alpha}_v \|\delta_v\|_2^2 \\
&\geq \frac{\kappa}{3} \|\delta_v\|_2^2 .
\end{aligned}$$

The same holds for every edge e. Further, by Lemma 9, $H_{e|v}(\tilde{\mu}_e) = H_v(\tilde{\mu}_v) - H_e(\tilde{\mu}_e)$ is convex, meaning

$$\left\langle \nabla(-H_{e|v}(\tilde{\mu}_e)) - \nabla(-H_{e|v}(\tilde{\mu}'_e)),\, \delta_e \right\rangle \geq 0.$$

Thus, taking the gradient of Eq. 22, we have that

$$\begin{aligned}
&\left\langle \nabla(-H^{\mathbf{c}}(\tilde{\boldsymbol{\mu}})) - \nabla(-H^{\mathbf{c}}(\tilde{\boldsymbol{\mu}}')),\, \boldsymbol{\delta} \right\rangle \\
&= \sum_{v \in \mathcal{V}} \tilde{\alpha}_v \langle \nabla(-H_v(\tilde{\mu}_v)) - \nabla(-H_v(\tilde{\mu}'_v)),\, \delta_v \rangle \\
&\quad + \sum_{e \in \mathcal{E}} \tilde{\alpha}_e \langle \nabla(-H_e(\tilde{\mu}_e)) - \nabla(-H_e(\tilde{\mu}'_e)),\, \delta_e \rangle \\
&\quad + \sum_{e \in \mathcal{E}} \sum_{v : v \in e} \tilde{\alpha}_{v,e} \left\langle \nabla(-H_{e|v}(\tilde{\mu}_e)) - \nabla(-H_{e|v}(\tilde{\mu}'_e)),\, \delta_e \right\rangle \\
&\geq \frac{\kappa}{3} \sum_{v \in \mathcal{V}} \|\delta_v\|_2^2 + \frac{\kappa}{3} \sum_{e \in \mathcal{E}} \|\delta_e\|_2^2 + 0 \\
&= \frac{\kappa}{3} \|\boldsymbol{\delta}\|_2^2 ,
\end{aligned}$$

which completes the proof, via Fact 1. ■

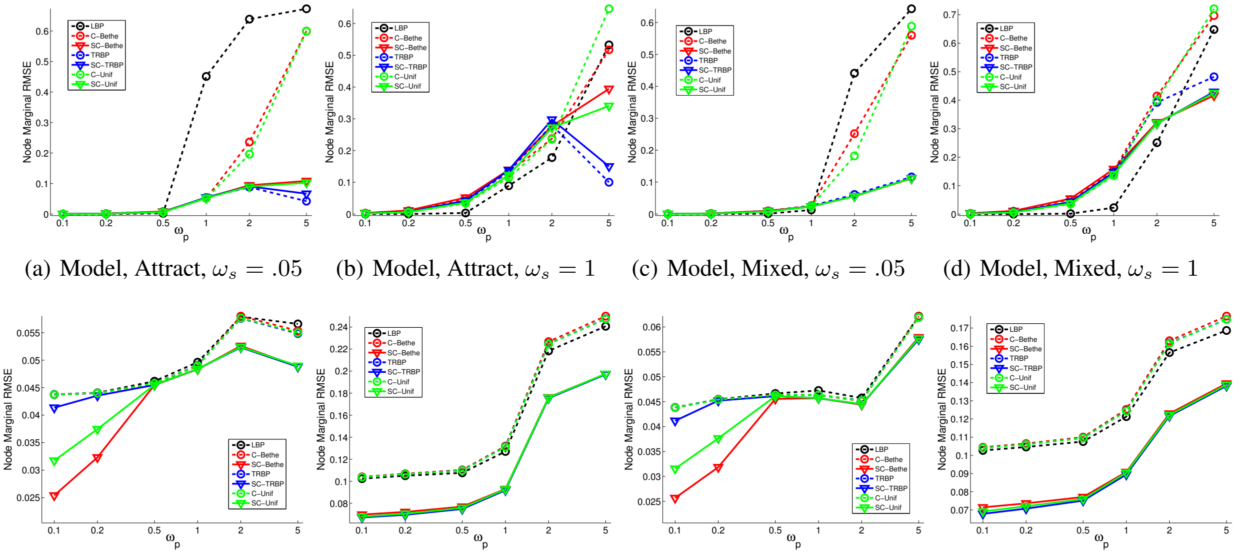
E.2. Slackened Variable-Valid Counting Number Optimization

For certain values of κ, the variable validity constraints in Eq. 13 create an infeasible optimization problem. When this happens, we propose switching to a slackened QP. This introduces a free parameter, $C \geq 0$, that adjusts the trade-off between fitting the target counts (in the equation below, the Bethe counts) and variable validity. The slackened QP is then

$$\begin{aligned}
\min_{\mathbf{c},\, \boldsymbol{\alpha} \geq \mathbf{0},\, \boldsymbol{\xi}} \quad & \|\mathbf{c} - \mathbf{c}^{\mathrm{B}}\|_2^2 + C \|\boldsymbol{\xi}\|_2^2 \qquad (23) \\
\text{s.t.} \quad & \forall v \in \mathcal{V},\ c_v + \sum_{e : v \in e} \alpha_{v,e} \geq 0 \,; \\
& \forall e \in \mathcal{E},\ c_e - \sum_{v : v \in e} \alpha_{v,e} \geq 3\kappa \,; \\
& \forall v \in \mathcal{V},\ c_v + \sum_{e : v \in e} c_e = 1 + \xi_v \,.
\end{aligned}$$

F. Figures for Experimental Results

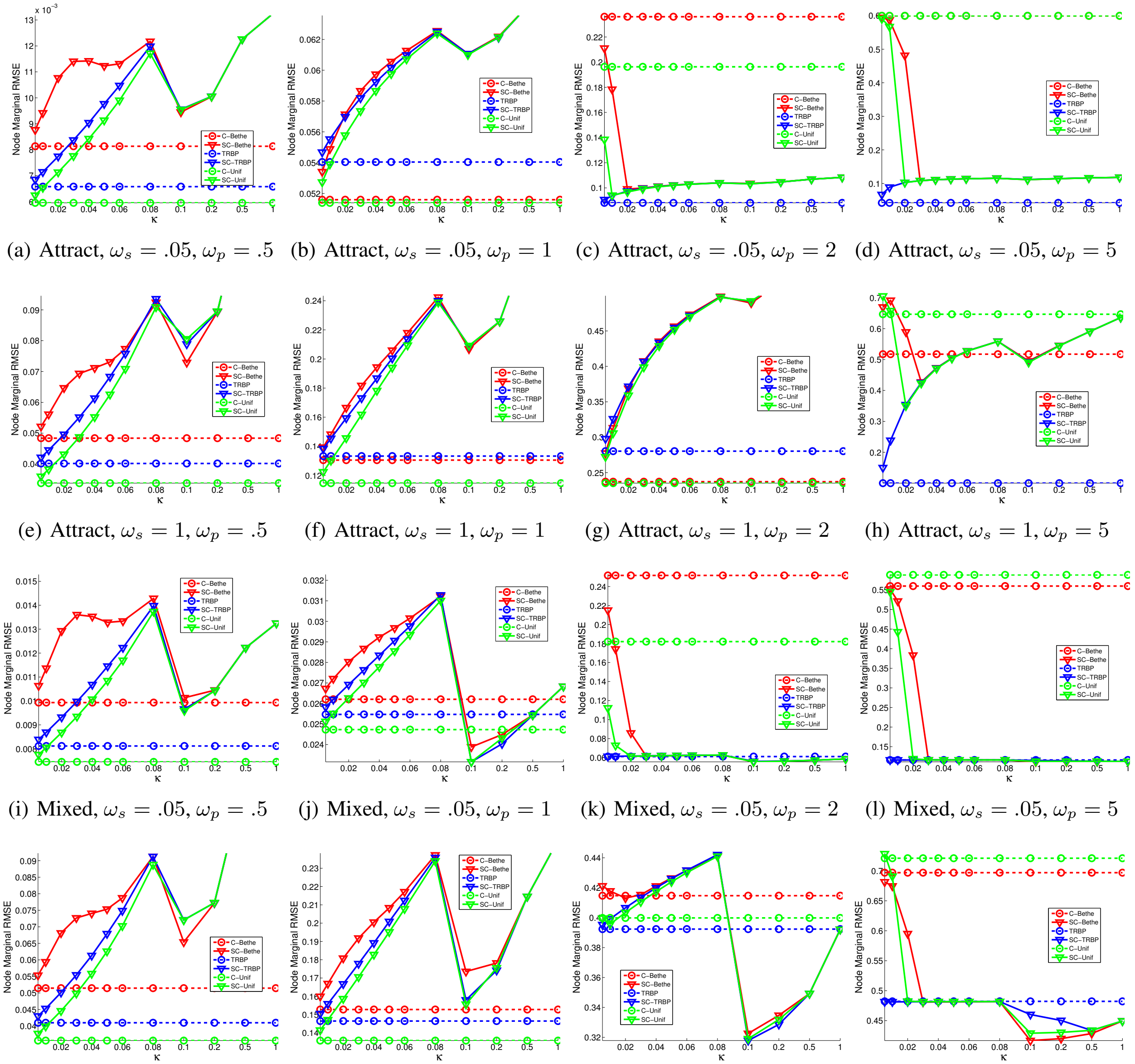
In all plots, results are averaged over 20 trials and the y-axis has been rescaled to fit the data. See Section 5.3 for discussion.



(a) Model, Attract, $\omega_s = .05$ (b) Model, Attract, $\omega_s = 1$ (c) Model, Mixed, $\omega_s = .05$ (d) Model, Mixed, $\omega_s = 1$

(e) Learned, Attract, $\omega_s = .05$ (f) Learned, Attract, $\omega_s = 1$ (g) Learned, Mixed, $\omega_s = .05$ (h) Learned, Mixed, $\omega_s = 1$

Figure 3. Plots of RMSE of the node marginals as a function of the interaction parameter, ω_p. Inference is performed using the true model in (a)-(d), and the learned model in (e)-(h). The first two columns correspond to a model with attractive potentials; the third and fourth to a model with mixed potentials. The black dotted line is LBP; color dotted lines are the convex baselines, and solid lines are their SC counterparts. The SC methods use the post hoc optimal value of κ (and C) in the counting number optimization. For learned marginals, SC offers statistically significant error reduction—sometimes over 40%—for all data models and baselines, except C-Bethe at $\omega_p = .5$ in (g).



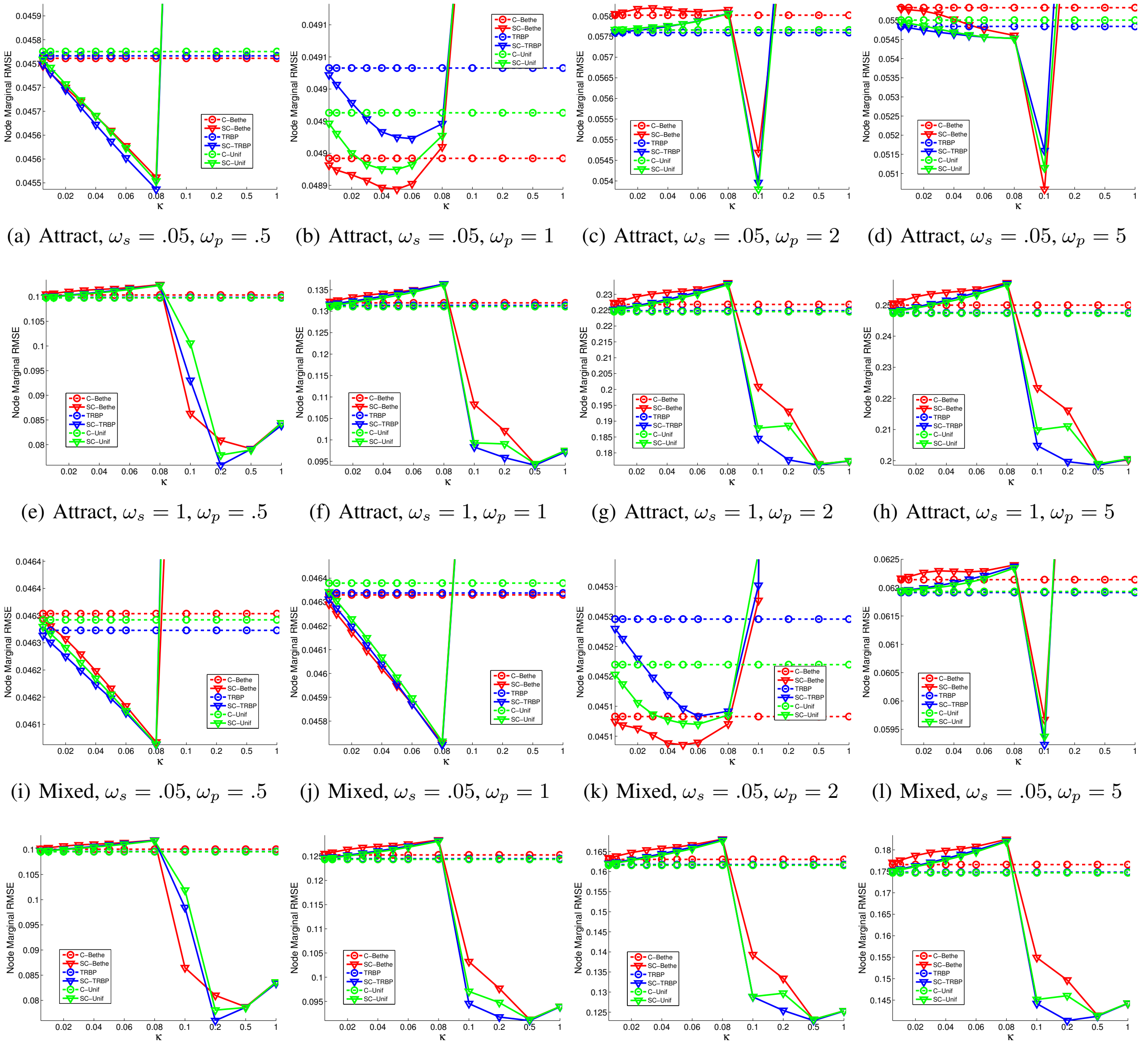
(a) Attract, $\omega_s = .05, \omega_p = .5$ (b) Attract, $\omega_s = .05, \omega_p = 1$ (c) Attract, $\omega_s = .05, \omega_p = 2$ (d) Attract, $\omega_s = .05, \omega_p = 5$

(e) Attract, $\omega_s = 1, \omega_p = .5$ (f) Attract, $\omega_s = 1, \omega_p = 1$ (g) Attract, $\omega_s = 1, \omega_p = 2$ (h) Attract, $\omega_s = 1, \omega_p = 5$

(i) Mixed, $\omega_s = .05, \omega_p = .5$ (j) Mixed, $\omega_s = .05, \omega_p = 1$ (k) Mixed, $\omega_s = .05, \omega_p = 2$ (l) Mixed, $\omega_s = .05, \omega_p = 5$

(m) Mixed, $\omega_s = 1, \omega_p = .5$ (n) Mixed, $\omega_s = 1, \omega_p = 1$ (o) Mixed, $\omega_s = 1, \omega_p = 2$ (p) Mixed, $\omega_s = 1, \omega_p = 5$

Figure 4. Plots of RMSE of the node marginals as a function of the convexity parameter, κ, which determines the minimum modulus of convexity used in the counting number QP. For $\kappa < .1$, we use Eq. 13; for $\kappa \geq .1$, we use Eq. 23 and report the score for the post hoc optimal C. SC algorithms are plotted as solid lines, and their respective counterparts are overlaid as dashed lines. Inference is performed using the true model. The first two rows correspond to a model with attractive potentials; the third and fourth to a model with mixed potentials. In all plots, the x-axis scales logarithmically for $\kappa > .1$. Certain plots have been truncated vertically to better fit the data.



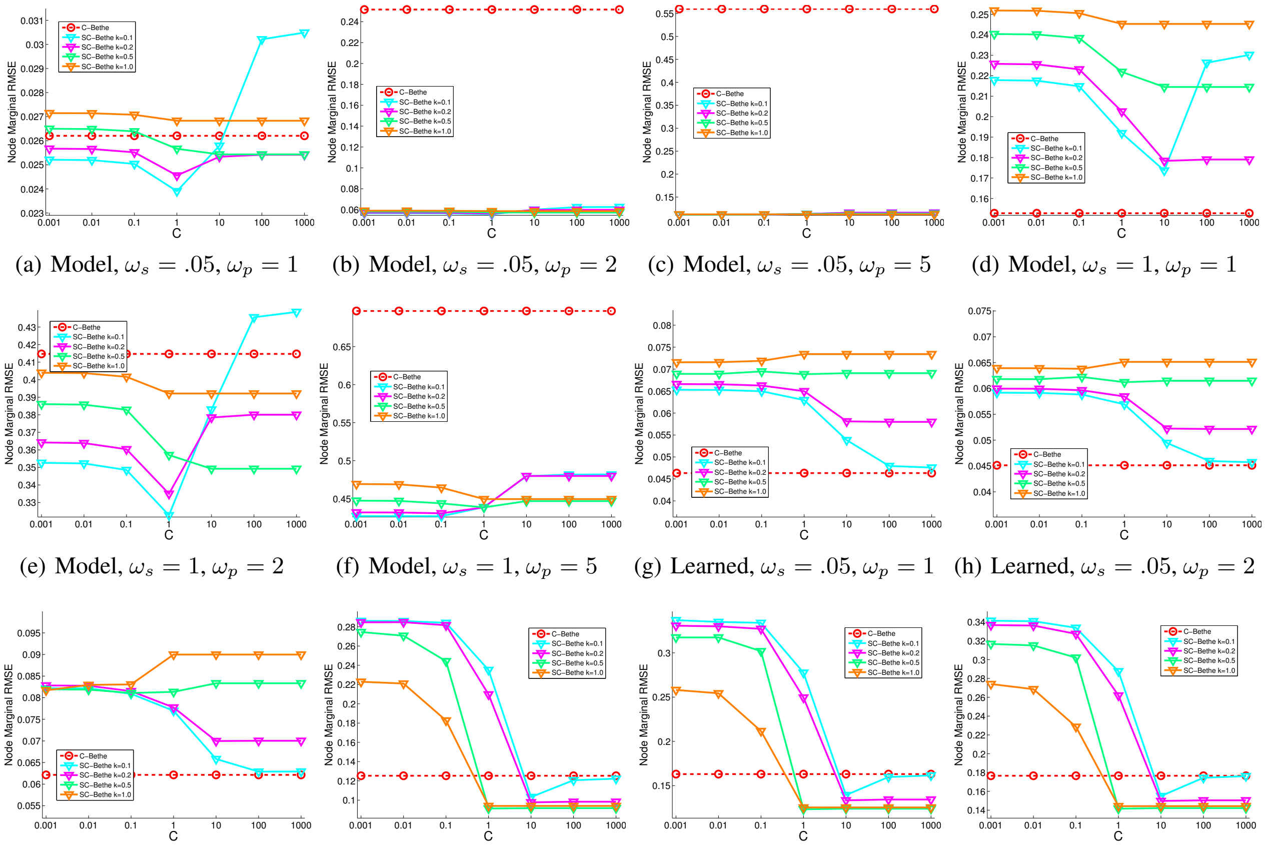
(a) Attract, $\omega_s = .05, \omega_p = .5$
(b) Attract, $\omega_s = .05, \omega_p = 1$
(c) Attract, $\omega_s = .05, \omega_p = 2$
(d) Attract, $\omega_s = .05, \omega_p = 5$

(e) Attract, $\omega_s = 1, \omega_p = .5$
(f) Attract, $\omega_s = 1, \omega_p = 1$
(g) Attract, $\omega_s = 1, \omega_p = 2$
(h) Attract, $\omega_s = 1, \omega_p = 5$

(i) Mixed, $\omega_s = .05, \omega_p = .5$
(j) Mixed, $\omega_s = .05, \omega_p = 1$
(k) Mixed, $\omega_s = .05, \omega_p = 2$
(l) Mixed, $\omega_s = .05, \omega_p = 5$

(m) Mixed, $\omega_s = 1, \omega_p = .5$
(n) Mixed, $\omega_s = 1, \omega_p = 1$
(o) Mixed, $\omega_s = 1, \omega_p = 2$
(p) Mixed, $\omega_s = 1, \omega_p = 5$

Figure 5. Plots of RMSE of the node marginals as a function of the convexity parameter, κ, when using the learned model for inference.

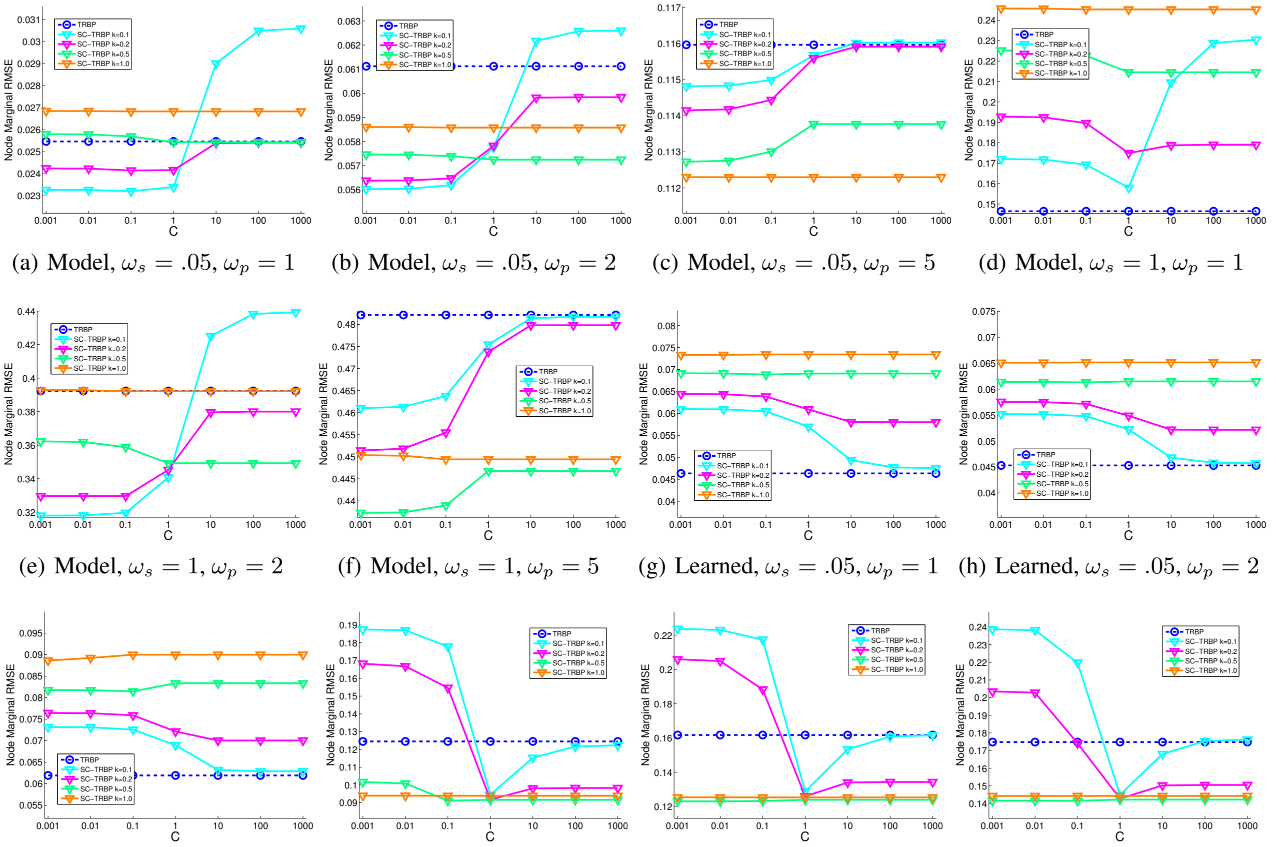


(a) Model, $\omega_s = .05, \omega_p = 1$ (b) Model, $\omega_s = .05, \omega_p = 2$ (c) Model, $\omega_s = .05, \omega_p = 5$ (d) Model, $\omega_s = 1, \omega_p = 1$

(e) Model, $\omega_s = 1, \omega_p = 2$ (f) Model, $\omega_s = 1, \omega_p = 5$ (g) Learned, $\omega_s = .05, \omega_p = 1$ (h) Learned, $\omega_s = .05, \omega_p = 2$

(i) Learned, $\omega_s = .05, \omega_p = 5$ (j) Learned, $\omega_s = 1, \omega_p = 1$ (k) Learned, $\omega_s = 1, \omega_p = 2$ (l) Learned, $\omega_s = 1, \omega_p = 5$

Figure 6. Select plots of RMSE as a function of the slack parameter, C, in the slackened counting number QP (Eq. 23), at higher values of κ. The slack parameter trades off between fitting the target counting numbers and satisfying variable validity. Data is generated using mixed potentials in all plots. These plots focus on the Bethe approximation. SC versions are solid color lines; C-Bethe is overlaid as a dashed red line.



(a) Model, $\omega_s = .05, \omega_p = 1$ (b) Model, $\omega_s = .05, \omega_p = 2$ (c) Model, $\omega_s = .05, \omega_p = 5$ (d) Model, $\omega_s = 1, \omega_p = 1$

(e) Model, $\omega_s = 1, \omega_p = 2$ (f) Model, $\omega_s = 1, \omega_p = 5$ (g) Learned, $\omega_s = .05, \omega_p = 1$ (h) Learned, $\omega_s = .05, \omega_p = 2$

(i) Learned, $\omega_s = .05, \omega_p = 5$ (j) Learned, $\omega_s = 1, \omega_p = 1$ (k) Learned, $\omega_s = 1, \omega_p = 2$ (l) Learned, $\omega_s = 1, \omega_p = 5$

Figure 7. Select plots of RMSE as a function of the slack parameter, C, for the tree-reweighting approximation. SC versions are solid color lines; C-TRBP is overlaid as a dashed blue line.