On the Strong Convexity of Variational Inference

Ben London
University of Maryland
blondon@cs.umd.edu

Bert Huang
University of Maryland
bert@cs.umd.edu

Lise Getoor
University of California, Santa Cruz
getoor@soe.ucsc.edu

1 Introduction

Variational methods provide a tractable approximation to the intractable task of computing inference in probabilistic graphical models. Much of the literature in this field concerns the convexity of the variational objective, since convexity guarantees convergence to a global optimum. However, less attention has focused on the *strength* of the convexity. By this 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* with respect to a norm $\|\cdot\|$ iff, for all $s, s' \in \mathcal{S}$, $\frac{\kappa}{2}\|s - s'\|^2 + \langle \nabla\varphi(s),\, s' - s\rangle \leq \varphi(s') - \varphi(s)$.

Strong convexity enables faster convergence in stochastic optimization [11], and has recently been shown to improve the *stability* of inference—that is, the sensitivity of a predictor to perturbations in the input. This is partially explained by the duality between strong convexity and *strong smoothness* [3]. Wainwright [15] used stability to show that learning with a strongly convex variational method can asymptotically produce a better approximation to the true model than learning with exact inference. Similarly, London et al. [7, 8] showed that predictors that use strongly convex variational inference have improved PAC generalization guarantees, due to the stability of inference. Kakade et al. [4] also related the excess risk (i.e., regret) of exponential families to their moduli of convexity.

We are therefore interested in which variational methods are strongly convex. Since the aforementioned bounds have a $\mathrm{O}(1/\kappa)$ dependence on the modulus of convexity, κ, we would also like to identify cases in which the modulus is a constant. In what follows, we present new strong convexity guarantees for two popular variational methods, *tree-reweighting* and *counting number* approximations. We provide conditions under which their respective objectives are strongly convex, with moduli that do not depend on the number of variables. When combined with existing theory [e.g., 4, 8, 15], this yields more optimistic generalization and regret bounds.

2 Background and Notation

We first introduce some notation and review definitions that will be useful in discussing our main results in the following sections. Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote a compact domain of observations, and $\mathcal{Y} \subset \{0, 1\}^k$ a set of k labels, represented by the k-dimensional standard basis (a.k.a. "one-hot") vectors. A structured example is a tuple, $(\mathbf{x}, \mathbf{y}, G)$, where $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{Y}^n$, and $G \triangleq (\mathcal{V}, \mathcal{E})$ is some implicit graph topology that represents the interactions between variables.

We consider the following class of Markov networks for classification, which includes many popular log-linear models [e.g., 6, 12, 14]. The model's *potential functions* are organized according to the nodes and edges of a graph, and parameterized by a vector of weights, $\mathbf{w}$. In practice, the weights may be *tied* (i.e., *templated*) across node (resp. edge) potentials, though this is not critical to our results. Let $\theta_v(y_v \,|\, \mathbf{x}; \mathbf{w})$ denote the potential for a node v being in state $y_v \in \mathcal{Y}$, conditioned on observations $\mathbf{x} \in \mathcal{X}$. Similarly, let $\theta_e(y_e \,|\, \mathbf{x}; \mathbf{w})$ denote the potential for edge e being in state $y_e \in \mathcal{Y}^2$. Since y_v is a vector, we can organize the potentials for v as a vector, $\boldsymbol{\theta}_v(\mathbf{x}; \mathbf{w})$; then, $\theta_v(y_v \,|\, \mathbf{x}; \mathbf{w}) = \boldsymbol{\theta}_v(\mathbf{x}; \mathbf{w}) \cdot y_v$. Similarly, $\theta_e(y_e \,|\, \mathbf{x}; \mathbf{w})$ and y_e can be vectorized, such that $\theta_e(y_e \,|\, \mathbf{x}; \mathbf{w}) = \boldsymbol{\theta}_e(\mathbf{x}; \mathbf{w}) \cdot y_e$. For brevity, when $\mathbf{x}$ and $\mathbf{w}$ are clear from context, we will simply use $\boldsymbol{\theta}_v$ or $\boldsymbol{\theta}_e$.

Given an example $(\mathbf{x}, \mathbf{y}, G)$, we *ground* the model by instantiating $\boldsymbol{\theta}_v$ and y_v for all nodes, and $\boldsymbol{\theta}_e$ and y_e for all edges. With $\boldsymbol{\theta} \triangleq ((\boldsymbol{\theta}_v)_{v\in\mathcal{V}}, (\boldsymbol{\theta}_e)_{e\in\mathcal{E}})$ and $\hat{\mathbf{y}} \triangleq ((y_v)_{v\in\mathcal{V}}, (y_e)_{e\in\mathcal{E}})$, we can then write the aggregate potential for $(\mathbf{x}, \mathbf{y}, G)$ as a dot product, $\boldsymbol{\theta} \cdot \mathbf{y}$. This describes a log-linear distribution, $p_{\mathbf{w}}(\mathbf{Y} = \mathbf{y} \,|\, \mathbf{X} = \mathbf{x}) = \exp\left(\boldsymbol{\theta} \cdot \hat{\mathbf{y}} - \Phi(\boldsymbol{\theta})\right)$, where $\Phi(\boldsymbol{\theta}) \triangleq \log \sum_{\hat{\mathbf{y}}'} \exp\left(\boldsymbol{\theta} \cdot \hat{\mathbf{y}}'\right)$ is a normalizing function known as the *log-partition*.

The log-partition is convex in $\boldsymbol{\theta}$, and has a well-known variational form [16],

$$\Phi(\boldsymbol{\theta}) = \max_{\boldsymbol{\mu}\in\mathcal{M}} \; \boldsymbol{\theta} \cdot \boldsymbol{\mu} + H(\boldsymbol{\mu}), \tag{1}$$

where $\mathcal{M}$ is the *marginal polytope*—the set of all consistent marginal vectors—and $H(\boldsymbol{\mu})$ is the entropy of the distribution consistent with marginals $\boldsymbol{\mu}$. The maximizing $\boldsymbol{\mu}$ corresponds to the marginal distribution of $\mathbf{Y}$ given $\mathbf{x}$. The negative of the quantity being maximized is often referred to as the *free energy*. The maximizing $\boldsymbol{\mu}$ are the true marginals of $\mathbf{Y}$, given $\mathbf{X} = \mathbf{x}$ and $\mathbf{w}$. Further, MAP inference is achieved by removing the entropy term.

Unfortunately, for general graph structures, $\mathcal{M}$ may require an exponential number of constraints, and H may lack an explicit form. Many variational methods address these problems by: a) relaxing $\mathcal{M}$ to an outer bound using a polynomial set of "local" constraints,

$$\tilde{\mathcal{M}} \triangleq \left\{ \tilde{\boldsymbol{\mu}} : \forall v \in \mathcal{V}, \sum_{j=1}^{k} \tilde{\boldsymbol{\mu}}_v^j = 1 \;;\; \forall e = \{u, v\} \in \mathcal{E}, \sum_{j=1}^{k} \tilde{\boldsymbol{\mu}}_e^{ij} = \tilde{\boldsymbol{\mu}}_u^i, \sum_{i=1}^{k} \tilde{\boldsymbol{\mu}}_e^{ij} = \tilde{\boldsymbol{\mu}}_v^j \right\};$$

b) replacing H with a tractable surrogate, Ψ. $\tilde{\mathcal{M}}$ is usually called the local marginal polytope, and each $\tilde{\boldsymbol{\mu}} \in \tilde{\mathcal{M}}$ is a set of pseudomarginals. An important property of the free energy is that, when $-\Psi$ is strongly convex, the free energy is strongly convex. For the reasons discussed in Section 1, we are interested in identifying cases in which $-\Psi$ is strongly convex with $\kappa = \Omega(1)$.

3 Tree-Reweighting

The *tree-reweighted* entropy approximation [17] 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, independently of the number of factors.

Fix a model and a graph G, and assume we are given a distribution ρ over the spanning trees of G, denoted $\mathcal{T}(G)$. Further, assume that each edge e has positive marginal probability, $\rho(e) > 0$ (i.e., appears in at least one tree T with $\rho(T) > 0$). For a spanning tree $T \triangleq (\mathcal{V}, \mathcal{E}_T)$, let H_T denote its entropy, which can be computed efficiently from a vector of marginals $\boldsymbol{\mu}$ via the Bethe entropy formula,

$$H_T(\boldsymbol{\mu}) \triangleq \sum_{v\in\mathcal{V}} (1 - \deg(v)) H_v(\mu_v) \;+\; \sum_{e\in\mathcal{E}_T} H_e(\mu_e); \tag{2}$$

H_v, and H_e are the node- and edge-wise local entropies, and $\deg(v)$ is the degree of node v. The tree-reweighted entropy is then $H^{\text{TR}}(\tilde{\boldsymbol{\mu}}) \triangleq \sum_{T\in\mathcal{T}(G)} \rho(T) H_T(\tilde{\boldsymbol{\mu}})$.

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

Lemma 1. *[15, Appendix C] Fix a model, 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 $\rho_e^\star \triangleq \min_{e\in\mathcal{E}} \rho(e)$ denote the minimum edge probability. Let $\kappa_T^\star$ 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, $-H^{\text{TR}}$, is $(\rho_e^\star \kappa_T^\star)$-strongly convex w.r.t. the 2-norm.*

It is well known that the negative entropy, $-H$, is convex [16]. Wainwright [15] showed that the negative entropy is in fact *strongly* convex, lower-bounding the modulus by $\Omega(1/N)$, which decreases as a function of the number of factors, N. This is a pessimistic lower bound, since it considers all graphical models in the exponential family. Indeed, we can show that the class of tree-structured models with finite weights and bounded degree induce a negative entropy function that is $\Omega(1)$-strongly convex. A key component of our analysis is the idea of Markov *contraction*.

Definition 2. Fix a model with finite parameters, $\mathbf{w}$, which induce a probability density $p_{\mathbf{w}}$. Fix a graph $G \triangleq (\mathcal{V}, \mathcal{E})$. For $\{u, v\} \in \mathcal{E}$, define the *contraction coefficient* between u and v as

$$\vartheta_{\mathbf{w}}(u, v) \triangleq \sup_{\mathbf{x} \in \mathcal{X},\, y, y' \in \mathcal{Y}} \left\| p_{\mathbf{w}}(Y_u \mid \mathbf{X} = \mathbf{x}, Y_v = y) - p_{\mathbf{w}}(Y_u \mid \mathbf{X} = \mathbf{x}, Y_v = y') \right\|_{\text{TV}}. \tag{3}$$

Denote the maximum of the contraction coefficients by $\vartheta^{\star}_{\mathbf{w}} \triangleq \sup_{\{u,v\} \in \mathcal{E}} \vartheta_{\mathbf{w}}(u, v)$.

The contraction coefficients measure the dependence between adjacent variables in a graphical model. A contraction coefficient of 1 implies determinism, and 0 implies independence. Observe that determinism can only be induced by infinite weight or feature magnitude. Thus, a model with finite weights and features always has $\vartheta^{\star}_{\mathbf{w}} < 1$. As we show in Appendix A, this contraction causes dependence to decay with graph distance, and implies the following bound on strong convexity, which is constant with respect to N.

Proposition 1. *Fix a model with weights* $\mathbf{w}$ *and tree structure* T*. Suppose its node degrees are uniformly upper-bounded by* Δ_T*, and that the maximum contraction coefficient* $\vartheta^{\star}_{\mathbf{w}} < 1/\Delta_T$*. It then follows that the negative tree entropy,* $-H_T$*, is* $\Omega(1)$*-strongly convex w.r.t. the 2-norm.*

In general, the contraction coefficients are intractable to compute, since they involve a supremum over $\mathcal{X}$. It is therefore impossible to verify the conditions of Proposition 1 in certain cases. However, there may be special cases in which it is feasible to compute Equation 3; for instance, if $\mathcal{X}$ exhibits internal structure that can be exploited. We further conjecture that templating the model (e.g., as a homogeneous Markov chain) may further reduce the time complexity, and may even reduce the contraction coefficients. We leave these as open problems for future work.

4 Counting Numbers

Various authors [e.g., 1, 2, 9, 10, 18] have proposed convex approximations to the Bethe entropy based on the concept of *counting numbers*. This technique generalizes the Bethe entropy with

$$H^{\mathbf{c}}(\tilde{\boldsymbol{\mu}}) \triangleq \sum_{v \in \mathcal{V}} c_v H_v(\tilde{\mu}_v) + \sum_{e \in \mathcal{E}} c_e H_e(\tilde{\mu}_e), \tag{4}$$

where $c_v \geq 0$ and $c_e \geq 0$ are the counting numbers associated with node v and edge e. (Note that $H^{\mathbf{c}}$ generalizes H^{TR}, since we can recreate H^{TR} with $c_v = 1 - \sum_{e: v \in e} \rho(e)$ and $c_e = \rho(e)$.) While existing work focuses on finding counting numbers that preserve convexity, we show how to find counting numbers that preserve *strong* convexity, with a bounded modulus.

Since $-H_v$ and $-H_e$ are convex, it is clear from Equation 4 that $-H^{\mathbf{c}}$ is convex for nonnegative counting numbers. Heskes [2] derived a more sophisticated set of 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 $\{\alpha_{v,e} \geq 0\}_{v, e: v \in e}$, such that

$$\forall v \in \mathcal{V},\;\; c_v = \alpha_v - \sum_{e: v \in e} \alpha_{v,e}, \quad \text{and} \quad \forall e \in \mathcal{E},\;\; c_e = \alpha_e + \sum_{v: v \in e} \alpha_{v,e}. \tag{5}$$

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_v and α_e are uniformly lower-bounded; the $\alpha_{v,e}$ variables, however, are only required to be nonnegative.

Proposition 2. *If* $H^{\mathbf{c}}$ *satisfies Equation 5, then for any* $\kappa > 0$ *such that* $\forall v, e$*,* $\alpha_v \geq \kappa$*,* $\alpha_e \geq \kappa$ *and* $\alpha_{v,e} \geq 0$*, it follows that* $-H^{\mathbf{c}}$ *is* κ*-strongly convex with respect to the 2-norm.*

Proposition 2 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. [10] proposed a "convexified" Bethe approximation. Their algorithm finds a set of counting numbers that best approximates the Bethe counting numbers, $c^{\text{B}}_v = 1 - \deg(v)$ and $c^{\text{B}}_e = 1$, while satisfying Heskes' convexity conditions (Equation 5). Via Proposition 2, incorporating the constraint that $\alpha_v \geq \kappa$ and $\alpha_e \geq \kappa$ ensures that the resulting approximation is κ-strongly convex. This yields the following constrained quadratic program:

$$\min_{\mathbf{c},\, \{\alpha_{v,e} \geq 0\}} \;\; \|\mathbf{c} - \mathbf{c}^{\text{B}}\|_2^2 \quad \text{s.t. } \forall v \in \mathcal{V},\, c_v + \sum_{e: v \in e} \alpha_{v,e} \geq \kappa\,;\; \forall e \in \mathcal{E},\, c_e - \sum_{v: v \in e} \alpha_{v,e} \geq \kappa.$$

One can also add constraints (or terms in the objective) to encourage uniform solutions for c_e [1], or to enforce *variable-valid* counting numbers [10]. The space of counting number optimizations is a rich area of research.

There is a trade-off between the modulus of convexity (and its associated stability and convergence benefits) and the accuracy of the marginals. Higher values of κ lead to more convex free energies, but possibly at the cost of increased approximation error. Clearly, an empirical study of this trade-off is the next step, which we plan to explore in future work.

5 Conclusion

In this paper, we analyzed the strong convexity of two variational methods for marginal inference in undirected graphical models. We provided conditions under which the tree-reweighted and counting number entropy approximations are strongly convex, with moduli that are constant with respect to the size of the model, thus improving prior guarantees. The scope of this work was to provide theoretical guarantees; thus, no empirical studies are presented. We plan to address this in future work.

Acknowledgements

This work was supported by the National Science Foundation (NSF) under grant no. IIS1218488, and 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

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

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

[3] S. Kakade, S. Shalev-Shwartz, and A. Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization, 2009.

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

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

[6] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Intl. Conference on Machine Learning*, 2001.

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

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

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

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

[11] A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Intl. Conference on Machine Learning*, 2012.

[12] M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62(1-2):107–136, 2006.

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

[14] B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *Uncertainty in Artificial Intelligence*, 2002.

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

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

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

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

A Proof of Proposition 1

Fix any finite weight vector $\mathbf{w} : \|\mathbf{w}\| < \infty$. Observe that the maximizers of the variational energy (Equation 1) are limited to some subset $\mathcal{M}$ that depends on $\mathbf{w}$ (and $\mathbf{x}$) via $\boldsymbol{\theta} = \boldsymbol{\theta}(\mathbf{x}; \mathbf{w})$. To highlight the dependence on $\mathbf{w}$, we let $\mathcal{M}_{\mathbf{w}}$ denote the set of realizable marginal vectors under $\mathbf{w}$; i.e.,

$$\mathcal{M}_{\mathbf{w}} \triangleq \{\boldsymbol{\mu} \in \mathcal{M} : \exists \mathbf{x} \in \mathcal{X},\, \Phi(\boldsymbol{\theta}(\mathbf{x}; \mathbf{w})) = \boldsymbol{\theta}(\mathbf{x}; \mathbf{w}) \cdot \boldsymbol{\mu} + H_T(\boldsymbol{\mu})\}\,. \tag{6}$$

It is easy to see that $\mathcal{M}_{\mathbf{w}}$, like $\mathcal{M}$, is a convex set, and that a marginal vector $\boldsymbol{\mu}$ maximizes the variational energy for weights $\mathbf{w}$ if, and only if, $\boldsymbol{\mu} \in \mathcal{M}_{\mathbf{w}}$. Therefore, though H_T is not a function $\mathbf{w}$, we can define an associated entropy function,

$$H_{\mathbf{w}}(\boldsymbol{\mu}) \triangleq \begin{cases} H_T(\boldsymbol{\mu}) & \text{if } \boldsymbol{\mu} \in \mathcal{M}_{\mathbf{w}}, \\ -\infty & \text{otherwise,} \end{cases}$$

which preserves the equivalence

$$\max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\theta} \cdot \boldsymbol{\mu} + H_T(\boldsymbol{\mu}) = \max_{\boldsymbol{\mu} \in \mathcal{M}_{\mathbf{w}}} \boldsymbol{\theta} \cdot \boldsymbol{\mu} + H_T(\boldsymbol{\mu}) = \max_{\boldsymbol{\mu} \in \mathcal{M}_{\mathbf{w}}} \boldsymbol{\theta} \cdot \boldsymbol{\mu} + H_{\mathbf{w}}(\boldsymbol{\mu}). \tag{7}$$

For an input $\mathbf{x}$, denote by $\Sigma_{\mathbf{w}}(\mathbf{Y} \mid \mathbf{x})$ the (grounded) *covariance matrix* of $\mathbf{Y}$ conditioned on $\mathbf{X} = \mathbf{x}$,

$$\Sigma_{\mathbf{w}}(\mathbf{Y} \mid \mathbf{x}) \triangleq \mathop{\mathbb{E}}_{\mathbf{w}}\left[\hat{\mathbf{y}}\hat{\mathbf{y}}^\top \mid \mathbf{x}\right] - \mathop{\mathbb{E}}_{\mathbf{w}}\left[\hat{\mathbf{y}} \mid \mathbf{x}\right] \mathop{\mathbb{E}}_{\mathbf{w}}\left[\hat{\mathbf{y}}^\top \mid \mathbf{x}\right], \tag{8}$$

where $\mathbb{E}_{\mathbf{w}}[\cdot \mid \mathbf{x}]$ denotes an expectation over the distribution $p_{\mathbf{w}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$. Let $\Sigma_{\mathbf{w}}^{-1}(\mathbf{Y} \mid \mathbf{x})$ denote its inverse (i.e., the *precision matrix*). The (inverse) covariance matrix has the following relationship to the convexity of $-H_{\mathbf{w}}$.

Lemma 2. *For a tree-structured model with weights* $\mathbf{w}$*, the negative entropy,* $-H_{\mathbf{w}}$*, is* $(1/\lambda_{\mathbf{w}}^{max})$*-strongly convex in* $\mathcal{M}_{\mathbf{w}}$ *with respect to the 2-norm, where*

$$\lambda_{\mathbf{w}}^{max} \triangleq \sup_{\mathbf{x} \in \mathcal{X}} \|\Sigma_{\mathbf{w}}(\mathbf{Y} \mid \mathbf{x})\|_2$$

is the maximum eigenvalue of the covariance matrix of $\mathbf{Y}$*, conditioned on* $\mathbf{w}$ *and any input.*

The proof is given in Appendix B.

By Lemma 2, to lower-bound the convexity of $-H_{\mathbf{w}}$, it suffices to upper-bound the spectral norm of $\Sigma_{\mathbf{w}}(\mathbf{Y} \mid \cdot)$. A simple way to do this (used by Wainwright [15]) 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 features are in $[0, 1]$; this yields a (loose) upper bound of $N/4$. For our purposes, this bound is too loose, since it grows with the size of the network.

A better approach is to analyze the 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 (Definition 2). 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 requires a technical lemma that is often credited to Dobrushin. We use a version of this given by Kontorovich [5].

Lemma 3 (5, 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

$$\|K\nu\|_{\text{TV}} = \sum_{\omega} \left| \sum_{\omega'} K(\omega \,|\, \omega')\nu(\omega') \right| \leq \vartheta \sum_{\omega'} |\nu(\omega')| = \vartheta \, \|\nu\|_{\text{TV}} \,,$$

where

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

is the contraction coefficient of K.

For the following, we use the shorthand $p_{\boldsymbol{\theta}}(y)$ to denote $p_{\mathbf{w}}(Y = y \,|\, \mathbf{X} = \mathbf{x})$, and similar probabilities. Using this notation, the maximum contraction coefficient for a fixed $\mathbf{x}$ is

$$\vartheta^{\star}_{\boldsymbol{\theta}} \triangleq \sup_{\substack{\{u,v\}\in\mathcal{E} \\ y_v, y'_v \in \mathcal{Y}}} \|p_{\boldsymbol{\theta}}(Y_u|y_v) - p_{\boldsymbol{\theta}}(Y_u|y'_v)\|_{\text{TV}} \leq \vartheta^{\star}_{\mathbf{w}}.$$

The inequality follows from $\vartheta^{\star}_{\mathbf{w}}$ being a uniform upper bound over all $\mathbf{x}$ (see Definition 2). Similarly, we let $\sigma_{\boldsymbol{\theta}}(y_u, y_v)$ denote the entry of the covariance matrix corresponding to $Y_u = y_u$ and $Y_v = Y_v$ (given $\mathbf{X} = \mathbf{x}$).

Fix any $\mathbf{x} \in \mathcal{X}$. Let $\pi(1), \ldots, \pi(\ell)$ 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 $Y_{\pi(\ell)} = y_{\pi(\ell)}$ can be written recursively as

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

Note that the second equality follows from the Markov property; since $Y_{\pi(\ell)}$ is conditionally independent of $Y_{\pi(1)}$ given $Y_{\pi(\ell-1)}$, we have that $p_{\boldsymbol{\theta}}(y_{\pi(\ell)} \,|\, y_{\pi(\ell-1)}, y_{\pi(\ell)}) = p_{\boldsymbol{\theta}}(y_{\pi(\ell)} \,|\, y_{\pi(\ell-1)})$. The absolute-value sum of covariances between a node assignment $y_{\pi(1)}$ and the states of $Y_{\pi(\ell)}$ can be

bounded via the contraction lemma as

$$\begin{aligned}
&\sum_{y_{\pi(\ell)}} \left|\sigma_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(\ell)})\right| \\
&\quad= \sum_{y_{\pi(\ell)}} \left|p_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(\ell)}) - p_{\boldsymbol{\theta}}(y_{\pi(1)})p_{\boldsymbol{\theta}}(y_{\pi(\ell)})\right| \\
&\quad= \sum_{y_{\pi(\ell)}} \left| \sum_{y_{\pi(\ell-1)}} p_{\boldsymbol{\theta}}(y_{\pi(\ell)} \,|\, y_{\pi(\ell-1)}) \left(p_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(\ell-1)}) - p_{\boldsymbol{\theta}}(y_{\pi(1)})p_{\boldsymbol{\theta}}(y_{\pi(\ell-1)})\right)\right| \\
&\quad\leq \vartheta^{\star}_{\mathbf{w}} \sum_{y_{\pi(\ell-1)}} \left|p_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(\ell-1)}) - p_{\boldsymbol{\theta}}(y_{\pi(1)})p_{\boldsymbol{\theta}}(y_{\pi(\ell-1)})\right| \\
&\quad\vdots \\
&\quad\leq (\vartheta^{\star}_{\mathbf{w}})^{\ell-2} \sum_{y_{\pi(2)}} \left|p_{\boldsymbol{\theta}}(y_{\pi(1)}, y_{\pi(2)}) - p_{\boldsymbol{\theta}}(y_{\pi(1)})p_{\boldsymbol{\theta}}(y_{\pi(2)})\right| \\
&\quad\leq (\vartheta^{\star}_{\mathbf{w}})^{\ell-2} \sum_{y_{\pi(2)}} \left| \sum_{y'_{\pi(1)}} p_{\boldsymbol{\theta}}(y_{\pi(2)}|y'_{\pi(1)}) \left(p_{\boldsymbol{\theta}}(y_{\pi(1)}, y'_{\pi(1)}) - p_{\boldsymbol{\theta}}(y_{\pi(1)})p_{\boldsymbol{\theta}}(y'_{\pi(1)})\right)\right| \\
&\quad\leq \frac{k}{4} (\vartheta^{\star}_{\mathbf{w}})^{\ell-1}.
\end{aligned}$$

This follows from recursive applications of Lemma 3, and the fact that the covariance of any variable assignment is at most $1/4$ in magnitude; similarly, 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

$$\begin{aligned}
&\sum_{y_c, y_d} \left|\sigma_{\boldsymbol{\theta}}((y_a, y_b), (y_c, y_d))\right| \\
&\quad= \sum_{y_c, y_d} \left|p_{\boldsymbol{\theta}}(y_a, y_b, y_c, y_d) - p_{\boldsymbol{\theta}}(y_a, y_b)p_{\boldsymbol{\theta}}(y_c, y_d)\right| \\
&\quad= \sum_{y_c, y_d} \left|p_{\boldsymbol{\theta}}(y_a, y_d \,|\, y_b, y_c)p_{\boldsymbol{\theta}}(y_b, y_c) - p_{\boldsymbol{\theta}}(y_a \,|\, y_b)p_{\boldsymbol{\theta}}(y_b)p_{\boldsymbol{\theta}}(y_d \,|\, y_c)p_{\boldsymbol{\theta}}(y_c)\right| \\
&\quad= \sum_{y_c, y_d} \left|p_{\boldsymbol{\theta}}(y_a \,|\, y_b)p_{\boldsymbol{\theta}}(y_d \,|\, y_c)p_{\boldsymbol{\theta}}(y_b, y_c) - p_{\boldsymbol{\theta}}(y_a \,|\, y_b)p_{\boldsymbol{\theta}}(y_b)p_{\boldsymbol{\theta}}(y_d \,|\, y_c)p_{\boldsymbol{\theta}}(y_c)\right| \\
&\quad= \sum_{y_c, y_d} p_{\boldsymbol{\theta}}(y_a \,|\, y_b)p_{\boldsymbol{\theta}}(y_d \,|\, y_c) \left|p_{\boldsymbol{\theta}}(y_b, y_c) - p_{\boldsymbol{\theta}}(y_b)p_{\boldsymbol{\theta}}(y_c)\right| \\
&\quad= p_{\boldsymbol{\theta}}(y_a \,|\, y_b) \sum_{y_c} \left|p_{\boldsymbol{\theta}}(y_b, y_c) - p_{\boldsymbol{\theta}}(y_b)p_{\boldsymbol{\theta}}(y_c)\right| \sum_{y_d} p_{\boldsymbol{\theta}}(y_d \,|\, y_c) \\
&\quad= p_{\boldsymbol{\theta}}(y_a \,|\, y_b) \sum_{y_c} \left|\sigma_{\boldsymbol{\theta}}(y_b, y_c)\right| \\
&\quad\leq \frac{k}{4} (\vartheta^{\star}_{\mathbf{w}})^{\ell-1}.
\end{aligned}$$

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 $1/4$.

The preceding discussion yields bounds for 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

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

where $\ell(u, v)$ is the length of the path from node u to v. The second inequality is 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 equality applies the geometric series identity, since $\Delta_T \, \vartheta^\star_{\mathbf{w}} < \Delta_T / \Delta_T = 1$. An analogous argument bounds the absolute-value sum of covariances involving any edge variable assignment. It therefore follows that the 1-norm of the covariance matrix is independent of N; that is,

$$\left\| \Sigma_{\mathbf{w}}(\mathbf{Y} \,|\, \mathbf{x}) \right\|_1 = \mathrm{O}(1).$$

Recall that the 1-norm of the covariance matrix upper-bounds the spectral norm, since the covariance matrix is symmetric. Thus, via Lemma 2, the negative entropy, $-H_{\mathbf{w}}$, of the model with weights $\mathbf{w}$ is $\Omega(1)$-strongly convex with respect to the 2-norm.

B Proof of Lemma 2

We begin with two facts regarding the log-partition and its duality with the negative entropy.

Fact 1. *For any* $\mathbf{w}$ *and* $\mathbf{x}$*, the covariance matrix of* $\mathbf{Y}$ *conditioned on* $\mathbf{X} = \mathbf{x}$ *is the* Hessian *(i.e., matrix of second derivatives) of the log-partition; i.e.,* $\nabla^2 \Phi(\boldsymbol{\theta}) = \Sigma_{\mathbf{w}}(\mathbf{Y} \,|\, \mathbf{x})$.

Fact 2. *The log-partition,* Φ*, is the* convex conjugate *of the negative entropy,* $-H$*; meaning,*

$$-H(\boldsymbol{\mu}) = \sup_{\boldsymbol{\theta}} \; \boldsymbol{\theta} \cdot \boldsymbol{\mu} - \Phi(\boldsymbol{\theta}). \tag{9}$$

Fact 1 is well known for graphical models; for a derivation, see Wainwright and Jordan [16]. Fact 2 follows from Equation 1. A direct consequence of these facts is the following.

Lemma 4. *Fix a tree-structured model with weights* $\mathbf{w}$*. For any marginals* $\boldsymbol{\mu} \in \mathcal{M}_{\mathbf{w}}$ *in the marginal polytope of* $\mathbf{w}$*, with*

$$\mathbf{x}^\star \triangleq \operatorname*{arg\,max}_{\mathbf{x} \in \mathcal{X}} \; \boldsymbol{\theta} \cdot \boldsymbol{\mu} - \Phi(\boldsymbol{\theta}),$$

we have that,

$$\nabla^2(-H_{\mathbf{w}}(\boldsymbol{\mu})) = \Sigma_{\mathbf{w}}^{-1}(\mathbf{Y} \,|\, \mathbf{x}^\star).$$

Proof For $\boldsymbol{\mu} \in \mathcal{M}_{\mathbf{w}}$, $-H_{\mathbf{w}}(\boldsymbol{\mu})$ has an explicit form (Equation 2) that is convex and clearly twice differentiable. From Fact 2, $-H_{\mathbf{w}}(\boldsymbol{\mu})$ has a variational form (Equation 9) that is maximized by a set of potentials, $\boldsymbol{\theta}^\star = \boldsymbol{\theta}(\mathbf{x}^\star; \mathbf{w})$, based on $\mathbf{w}$ (by definition of $\mathcal{M}_{\mathbf{w}}$) and $\mathbf{x}^\star$. In other words, Φ is the Legendre transform of $-H_{\mathbf{w}}$. Thus, via Legendre duality, the Hessians of $-H_{\mathbf{w}}$ and Φ have an inverse relationship, where

$$\nabla^2(-H_{\mathbf{w}}(\boldsymbol{\mu})) = \left(\nabla^2 \Phi(\boldsymbol{\theta}) \right)^{-1}.$$

The proof is completed by Fact 1. ■

We can now show that $-H_{\mathbf{w}}$ is strongly convex in $\mathcal{M}_{\mathbf{w}}$. For twice-differentiable functions, Definition 1 is equivalent to the following.

Fact 3. *Let $\varphi : \mathcal{S} \to \mathbb{R}$ denote a twice-differentiable function of a convex set $\mathcal{S}$. If, for all $s, s' \in \mathcal{S}$,*

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

then φ is κ-strongly convex with respect to $\|\cdot\|$.

Therefore, if $-H_{\mathbf{w}}$ satisfies

$$\kappa \|\boldsymbol{\mu}\|_2^2 \leq \langle \boldsymbol{\mu}, \nabla^2(-H_{\mathbf{w}}(\boldsymbol{\mu}'))\, \boldsymbol{\mu} \rangle .$$

for all $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathcal{M}_{\mathbf{w}}$, then $-H_{\mathbf{w}}$ is κ-strongly convex with respect to the 2-norm. By Lemma 4, there exists an $\mathbf{x}^\star \in \mathcal{X}$ such that

$$\kappa \|\boldsymbol{\mu}\|_2^2 \leq \langle \boldsymbol{\mu}, \Sigma_{\mathbf{w}}^{-1}(\mathbf{Y} \mid \mathbf{x}^\star)\, \boldsymbol{\mu} \rangle . \tag{10}$$

This means that $-H_{\mathbf{w}}$ is κ-strongly convex if the minimum eigenvalue of $\Sigma_{\mathbf{w}}^{-1}(\mathbf{Y} \mid \mathbf{x}^\star)$ is lower-bounded by κ; or, equivalently, that the maximum eigenvalue of $\Sigma_{\mathbf{w}}(\mathbf{Y} \mid \mathbf{x}^\star)$ is upper-bounded by $1/\kappa$. Since the eigenspectrum of $\Sigma_{\mathbf{w}}(\mathbf{Y} \mid \mathbf{x}^\star)$ is uniformly upper-bounded by $\lambda_{\mathbf{w}}^{\max}$ over all $\hat{\mathbf{x}} \in \hat{\mathcal{X}}$, it follows that Equation 10 holds for $\kappa = 1/\lambda_{\mathbf{w}}^{\max}$.

C Proof of Proposition 2

The proof of Proposition 2 requires several technical lemmas.

Fact 4. *A differentiable function, $\varphi : \mathcal{S} \to \mathbb{R}$, of a convex set $\mathcal{S}$, is κ-strongly convex with respect to a norm $\|\cdot\|$ iff, for all $s, s' \in \mathcal{S}$,*

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

Lemma 5 (13, Lemma 16)**.** *The function $\varphi(\mathbf{z}) \triangleq \sum_i^d z_i \log z_i$ is 1-strongly convex in the probability simplex, $\{\mathbf{z} \in [0,1]^d : \|\mathbf{z}\|_1 = 1\}$, with respect to the 1-norm.*

Lemma 6 (2, 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 2.

Proof [Proposition 2] Substituting Equation 5 into Equation 4 and rearranging the terms, we obtain

$$\begin{aligned} -H^{\mathsf{c}}(\tilde{\boldsymbol{\mu}}) &= -\sum_{v \in \mathcal{V}} \alpha_v H_v(\tilde{\mu}_v) - \sum_{e \in \mathcal{E}} \alpha_e H_e(\tilde{\mu}_e) + \sum_{e \in \mathcal{E}} \sum_{v : v \in e} \alpha_{v,e} (H_v(\tilde{\mu}_v) - H_e(\tilde{\mu}_e)) \\ &= -\sum_{v \in \mathcal{V}} \alpha_v H_v(\tilde{\mu}_v) - \sum_{e \in \mathcal{E}} \alpha_e H_e(\tilde{\mu}_e) - \sum_{e \in \mathcal{E}} \sum_{v : v \in e} \alpha_{v,e} H_{e|v}(\tilde{\mu}_e). \end{aligned}$$

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 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 5, $-H_v$ and $-H_e$ are 1-strongly convex in the probability simplex with respect to the 1-norm. By Fact 4, this means that every node v satisfies,

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

Therefore,

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

The same holds for every edge e. Further, by Lemma 6, $H_{e|v}(\tilde{\mu}_e) = H_v(\tilde{\mu}_v) - H_e(\tilde{\mu}_e)$ is convex, meaning

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

Thus, decomposing the gradient of $-H^{\mathbf{c}}$, we have that

$$\begin{aligned}
&\left\langle \nabla(-H^{\mathbf{c}}(\tilde{\boldsymbol{\mu}})) - \nabla(-H^{\mathbf{c}}(\tilde{\boldsymbol{\mu}}')),\, \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}' \right\rangle \\
&\quad = \quad \sum_{v\in\mathcal{V}} \alpha_v \left\langle \nabla(-H_v(\tilde{\mu}_v)) - \nabla(-H_v(\tilde{\mu}'_v)),\, \tilde{\mu}_v - \tilde{\mu}'_v \right\rangle \\
&\qquad + \sum_{e\in\mathcal{E}} \alpha_e \left\langle \nabla(-H_e(\tilde{\mu}_e)) - \nabla(-H_e(\tilde{\mu}'_e)),\, \tilde{\mu}_e - \tilde{\mu}'_e \right\rangle \\
&\qquad + \sum_{e\in\mathcal{E}} \sum_{v:v\in e} \alpha_{v,e} \left\langle \nabla(-H_{e|v}(\tilde{\mu}_e)) - \nabla(-H_{e|v}(\tilde{\mu}'_e)),\, \tilde{\mu}_e - \tilde{\mu}'_e \right\rangle \\
&\quad \geq \quad \kappa \sum_{v\in\mathcal{V}} \|\tilde{\mu}_v - \tilde{\mu}'_v\|_2^2 + \kappa \sum_{e\in\mathcal{E}} \|\tilde{\mu}_e - \tilde{\mu}'_e\|_2^2 + 0 = \kappa \left\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}'\right\|_2^2 .
\end{aligned}$$

which gives completes the proof. ■