The Benefits of Learning with Strongly Convex Approximate Inference
*** Supplemental ***

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

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 $s, s' \in \mathcal{S}$,*

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

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, $\frac{1}{m}\sum_{j=1}^{m} \hat{\boldsymbol{y}}^{(j)}$. 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,

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

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

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

The second line follows from the fact that $\hat{\boldsymbol{\theta}}_m$ is the minimizer of $\mathcal{L}^{\text{R}}_m$, which is differentiable, so $\nabla\mathcal{L}^{\text{R}}_m(\hat{\boldsymbol{\theta}}_m) = \mathbf{0}$. 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 $1-\delta$,

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

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

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

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

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

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

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

Thus, using the union bound, we have that

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

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

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

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 $1-\delta$,

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

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 $1-\delta$,

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

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 $\Omega(1/|G|)$-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

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

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

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

Applying this identity recursively, we have that

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

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

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

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,

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

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 $-H^{\text{TR}}$ Strong Convexity (Proposition 2)

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 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, $-H^{\text{TR}}$, 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 1. 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.

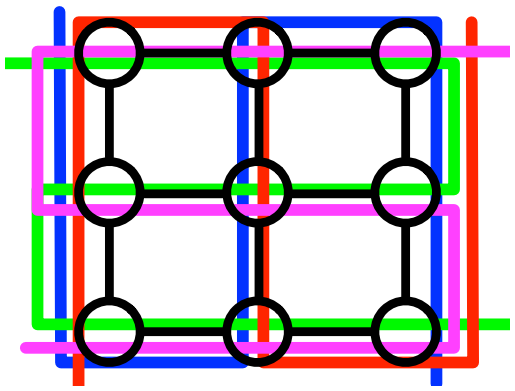


Figure 1. Covering the edges of a grid graph with 4 chains.

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.

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

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

F. Figures for Experimental Results

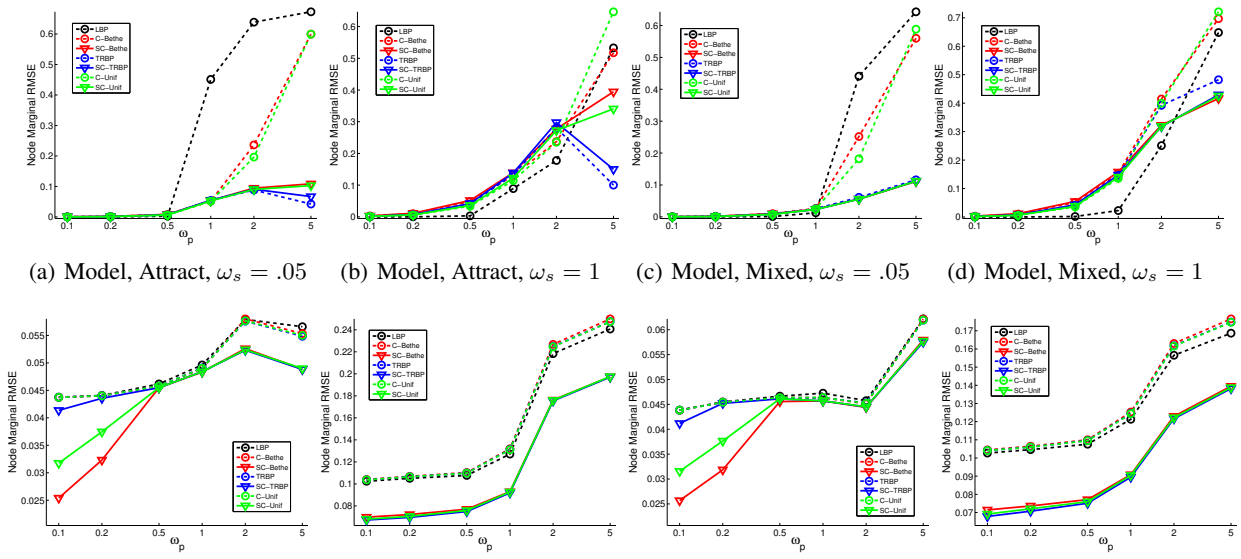
In all plots, results are averaged over 20 trials and the y-axis has been rescaled to fit the data. See Section 5.3 for discussion.

References

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

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. PAC-Bayesian collective stability. In *Artificial Intelligence and Statistics*, 2014.



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

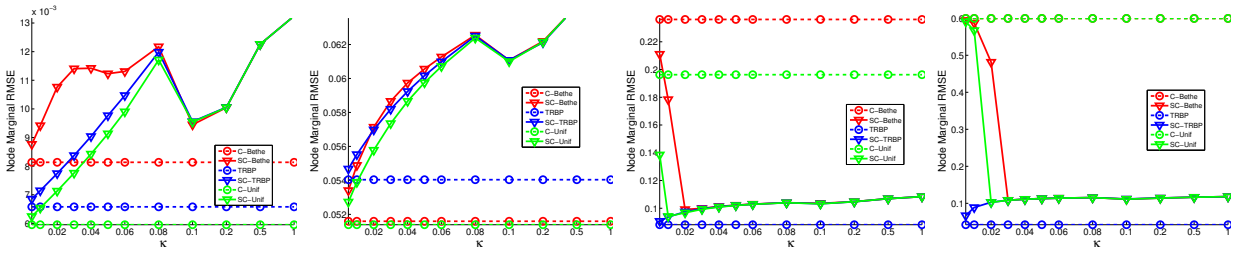
(e) Learned, Attract, $\omega_s = .05$ (f) Learned, Attract, $\omega_s = 1$ (g) Learned, Mixed, $\omega_s = .05$ (h) Learned, Mixed, $\omega_s = 1$

Figure 2. 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).

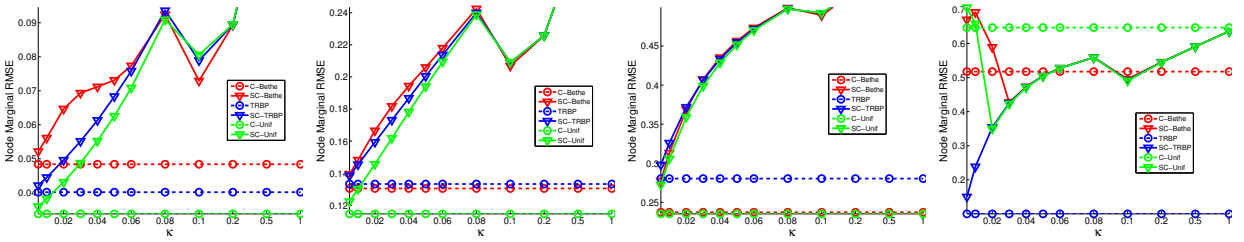
Shalev-Schwartz, S. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University of Jerusalem, 2007.

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

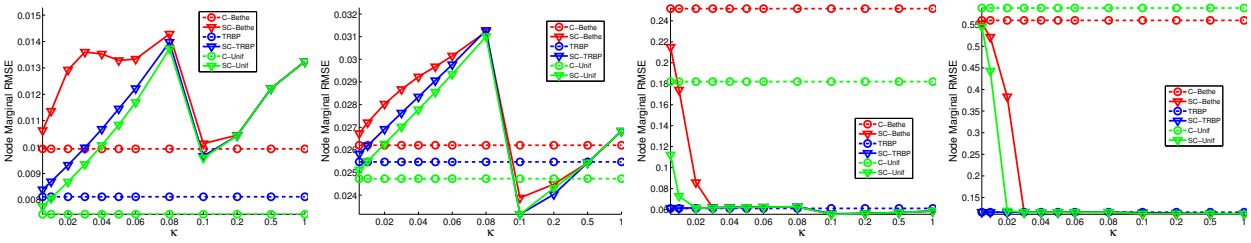
Wainwright, M. and Jordan, M. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., 2008.



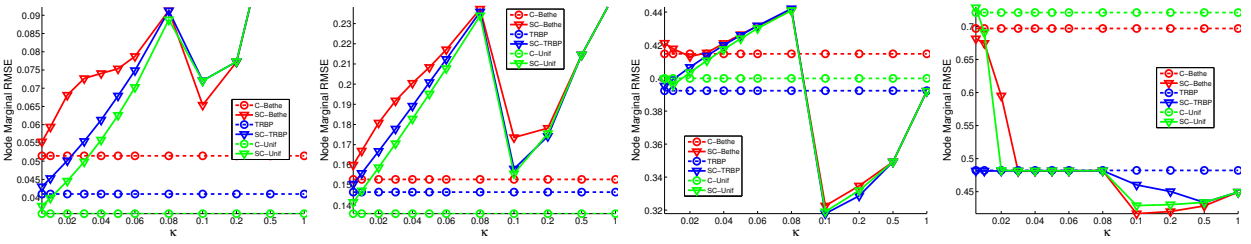
(a) Attract, $\omega_s = .05, \omega_p = .5$ (b) Attract, $\omega_s = .05, \omega_p = 1$ (c) Attract, $\omega_s = .05, \omega_p = 2$ (d) Attract, $\omega_s = .05, \omega_p = 5$



(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, $\omega_s = 1, \omega_p = 5$

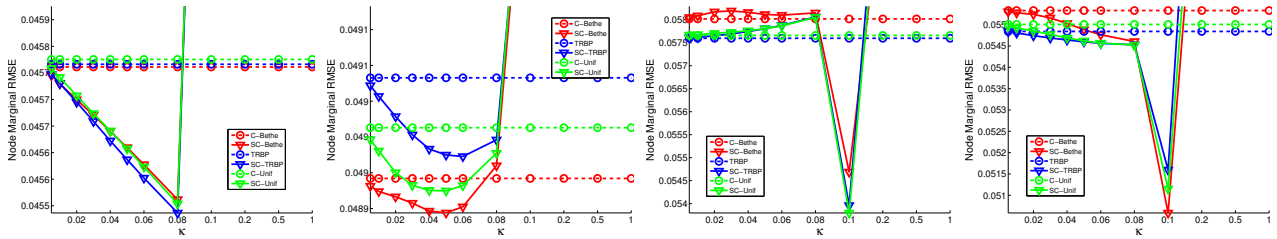


(i) Mixed, $\omega_s = .05, \omega_p = .5$ (j) Mixed, $\omega_s = .05, \omega_p = 1$ (k) Mixed, $\omega_s = .05, \omega_p = 2$ (l) Mixed, $\omega_s = .05, \omega_p = 5$

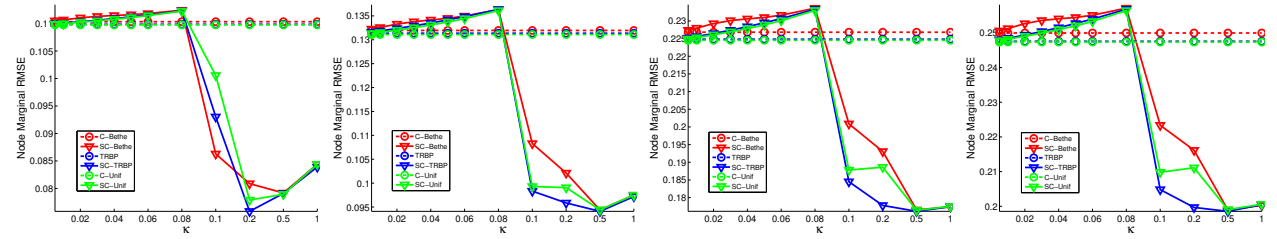


(m) Mixed, $\omega_s = 1, \omega_p = .5$ (n) Mixed, $\omega_s = 1, \omega_p = 1$ (o) Mixed, $\omega_s = 1, \omega_p = 2$ (p) Mixed, $\omega_s = 1, \omega_p = 5$

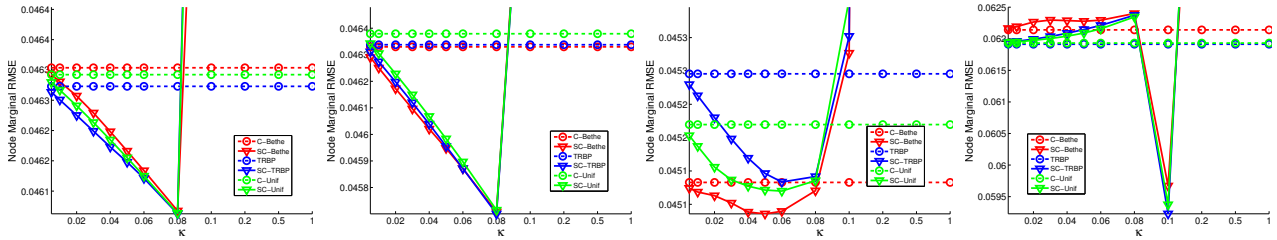
Figure 3. Plots of RMSE of the node marginals as a function of the convexity parameter, κ, which determines the minimum modulus of convexity used in the counting number QP. For $\kappa < .1$, we use Eq. 13; for $\kappa \geq .1$, we use Eq. 23 and report the score for the post hoc optimal C. SC algorithms 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.



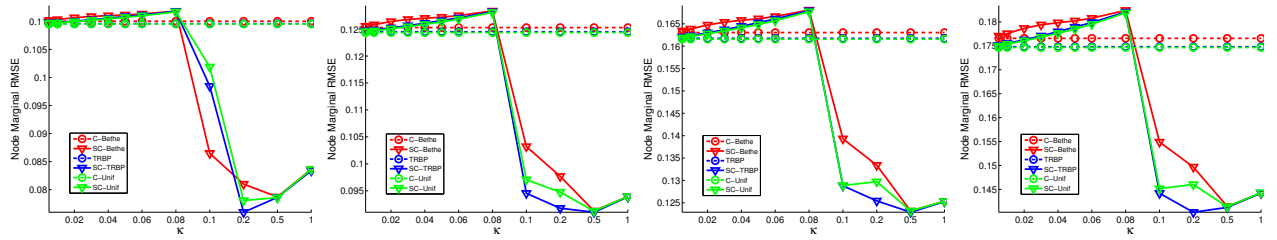
(a) Attract, $\omega_s = .05, \omega_p = .5$ (b) Attract, $\omega_s = .05, \omega_p = 1$ (c) Attract, $\omega_s = .05, \omega_p = 2$ (d) Attract, $\omega_s = .05, \omega_p = 5$



(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, $\omega_s = 1, \omega_p = 5$

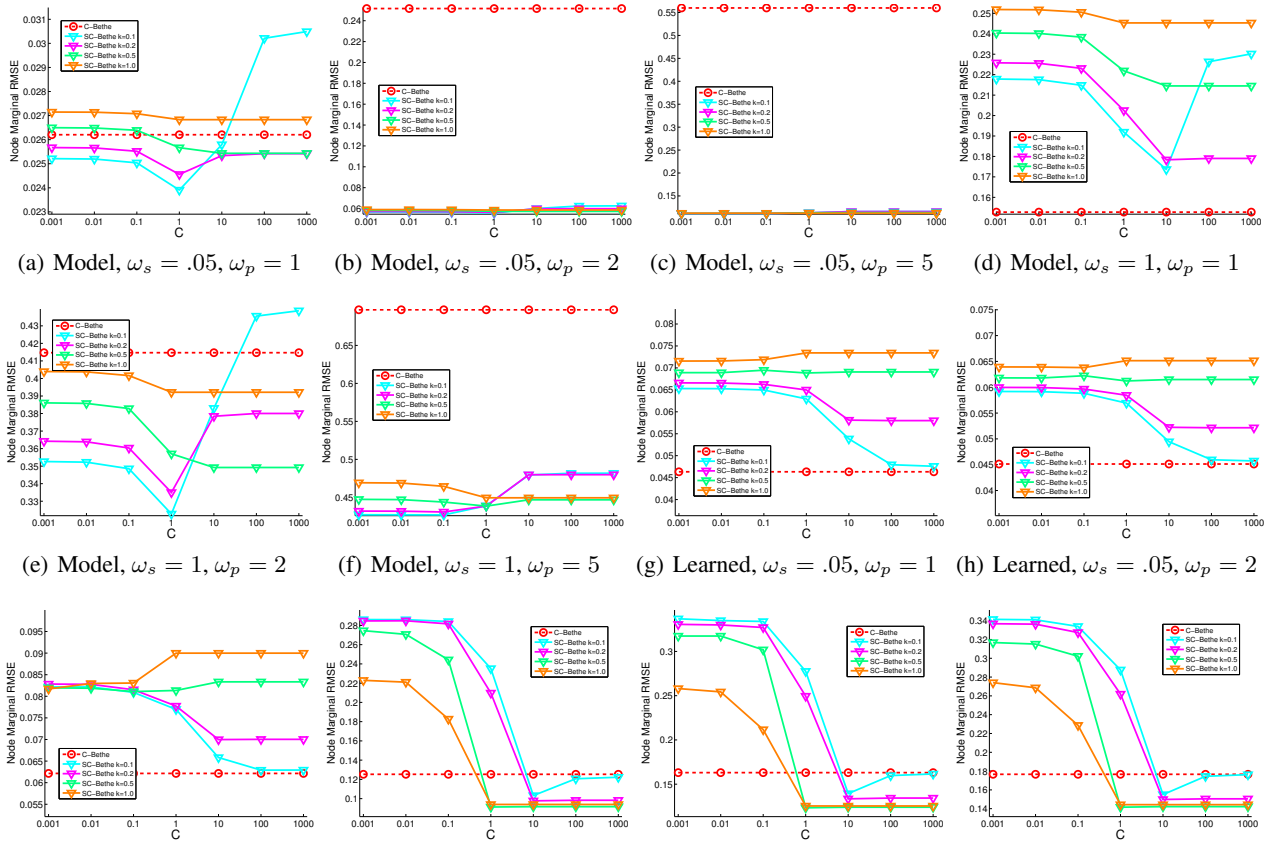


(i) Mixed, $\omega_s = .05, \omega_p = .5$ (j) Mixed, $\omega_s = .05, \omega_p = 1$ (k) Mixed, $\omega_s = .05, \omega_p = 2$ (l) Mixed, $\omega_s = .05, \omega_p = 5$



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

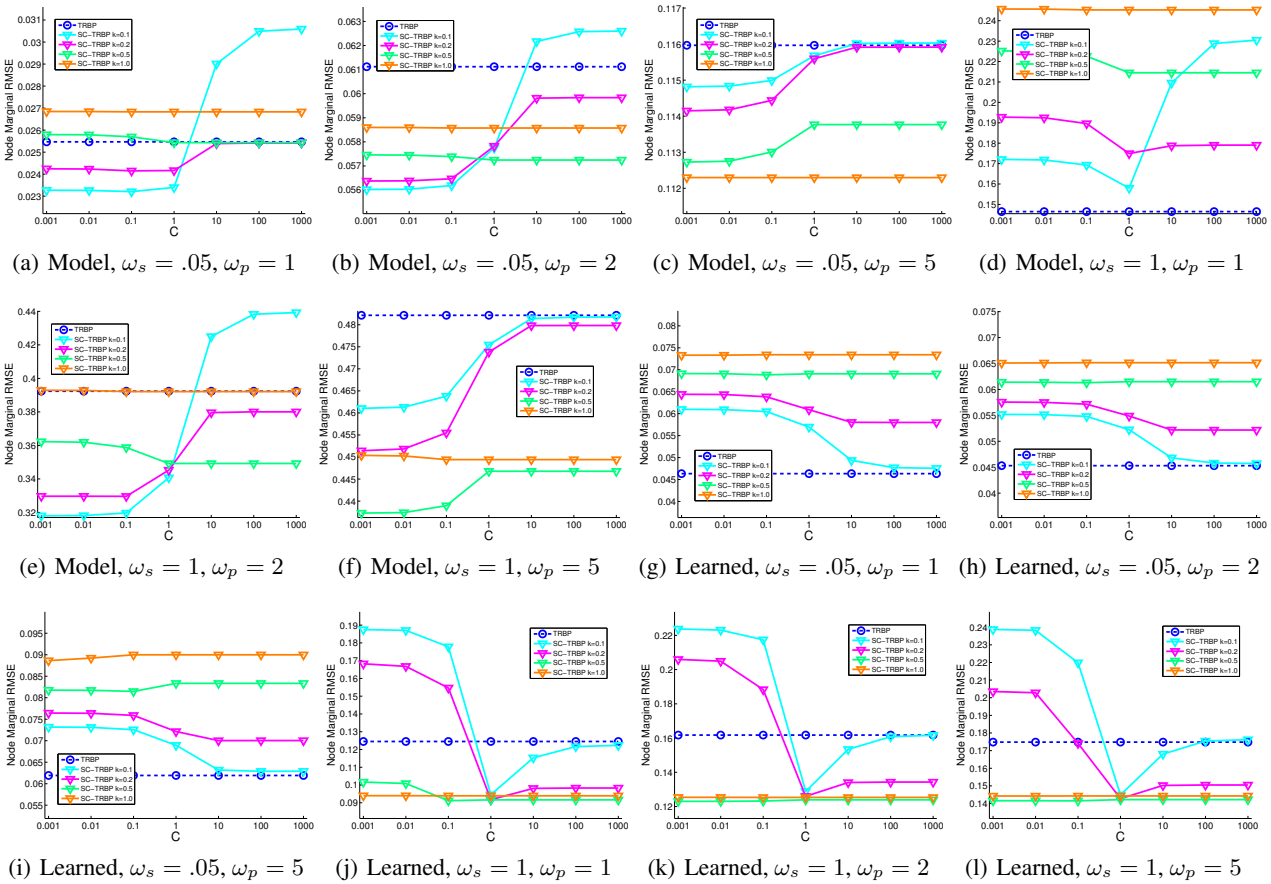


(a) Model, $\omega_s = .05, \omega_p = 1$ (b) Model, $\omega_s = .05, \omega_p = 2$ (c) Model, $\omega_s = .05, \omega_p = 5$ (d) Model, $\omega_s = 1, \omega_p = 1$

(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 5. 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 6. 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.