Vector-Space Markov Random Fields via Exponential Families
Appendix

A. Proof of Theorem 1

The proof follows the same lines as the proof in Yang et al. (2014). Let us denote \( Q(X) \) as \( \log \left( \frac{P(X)}{P(0)} \right) \). Note that \( X = (X_1, X_2, \ldots, X_p) \) and each \( X_s \) belongs to a vector space. Given any \( X \), let us denote \( X_s \) as \( X_s = (X_1, \ldots, X_{s-1}, 0, X_{s+1}, \ldots, X_p) \). Consider the following expansion for \( Q(X) \):

\[
Q(X) = \sum_{t \in \{1, \ldots, p\}} T[X_t \neq 0]G_t(X_t) + \cdots + \sum_{t_1, \ldots, t_k \in \{1, \ldots, p\}} T[X_{t_1} \neq 0, \ldots, X_{t_k} \neq 0]G_{t_1 \ldots t_k}(X_{t_1} \ldots X_{t_k})
\]  

(A.1)

where \( T \) is the indicator function which takes value 1 if its argument evaluates to true and 0 otherwise.

Using some simple algebra and the definition \( Q(X) = \log \left( \frac{P(X)}{P(0)} \right) \) we can show that

\[
\exp \left( Q(X) - Q(\tilde{X}_s) \right) = \frac{P(X|X_1, \ldots, X_{s-1}, X_{s+1}, \ldots, X_p)}{P(0|X_1, \ldots, X_{s-1}, X_{s+1}, \ldots, X_p)}
\]  

(A.2)

From (A.1) we have the following:

\[
(Q(X) - Q(X_s))
\]

\[
= T[X_s \neq 0] \left( G_s(X_s) + \sum_{t \in \{1, \ldots, p\}\setminus s} T[X_t \neq 0]G_{s,t}(X_s, X_t) \right)
\]  

(A.3)

Since the node conditional distribution follows the exponential family distribution defined in (6) we can show that:

\[
\log \frac{P(X|X_1, \ldots, X_{s-1}, X_{s+1}, \ldots, X_p)}{P(X_1, \ldots, X_{s-1}, X_{s+1}, \ldots, X_p)} = \langle E_s(X_{-s}), B_s(X_s) - B_s(0) \rangle + (C_s(X_s) - C_s(0))
\]  

(A.4)

Using (A.3) and (A.4) for left and right hand sides of (A.2) and setting \( X_t = 0 \) for all \( t \neq s \) we obtain:

\[
T[X_s \neq 0]G_s(X_s) = \langle E_s(0), B_s(X_s) - B_s(0) \rangle + (C_s(X_s) - C_s(0))
\]

Similarly setting \( X_r = 0 \) for all \( r \notin \{s, t\} \) we obtain:

\[
T[X_s \neq 0]G_s(X_s) + T[X_s \neq 0, X_t \neq 0]G_{s,t}(X_s, X_t) = \langle E_s(0 \cdots X_s \cdots 0), B_s(X_s) - B_s(0) \rangle + (C_s(X_s) - C_s(0))
\]

Similarly, replacing \( X_s \) with \( X_t \) in (A.2) and setting \( X_r = 0 \) for all \( r \notin \{s, t\} \) we obtain:

\[
T[X_t \neq 0]G_t(X_t) + T[X_s \neq 0, X_t \neq 0]G_{s,t}(X_s, X_t) = \langle E_t(0 \cdots X_t \cdots 0), B_t(X_t) - B_t(0) \rangle + (C_t(X_t) - C_t(0))
\]

From the above three equations we arrive at the following equality:

\[
\langle E_s(0 \cdots X_s \cdots 0) - E_s(0), B_s(X_s) - B_s(0) \rangle = \langle E_t(0 \cdots X_t \cdots 0) - E_t(0), B_t(X_t) - B_t(0) \rangle
\]  

(A.5)

The above equality should hold for the node conditional distributions to be consistent with the joint MRF distribution over \( X \) with respect to graph \( G \). So we need to find the form of \( E_t() \) that satisfies the above equation. Omitting zero vectors for clarity from (A.5), we get the following:

\[
\langle E_t(X_s), B_t(X_t) \rangle = \langle E_s(X_t), B_s(X_s) \rangle
\]

\[
\sum_j E_{sl}j(X_s)B_{slj}(X_t) = \sum_l E_{sl}l(X_t)B_{sl}(X_s)
\]  

(A.6)

We rewrite the natural parameter functions as

\[
E_{sl}j(X_s) = \sum_l \theta_{s;tlj}B_{sl}(X_s) + B_{slj}(X_s)
\]

\[
E_{sl}(X_t) = \sum_j \theta_{sl;lj}B_{lj}(X_t) + B_{sl}(X_t)
\]  

(A.7)

where \( \forall j \ B_{lj}(X_s) \) are functions in the Hilbert space \( H_s \) orthogonal to the span of functions \( B_s(X_s) \), and \( \forall j \ B_{sl}(X_s) \) are functions in the Hilbert space \( H_t \) orthogonal to the span of functions \( B_t(X_t) \); and \( \theta_{s;tlj}, \theta_{sl;lj} \) are scalars. Combining (A.6) and (A.7), we get

\[
\sum_j \sum_l \theta_{s;tlj}B_{sl}(X_s)B_{lj}(X_t) + \sum_j B_{lj}(X_s)B_{slj}(X_t)
\]

\[
= \sum_j \sum_l \theta_{sl;lj}B_{lj}(X_t)B_{lj}(X_s) + \sum_l B_{sl}(X_s)B_{sl}(X_s)
\]  

(A.8)
Rearranging terms in the above equation gives us the following equation:
\[
\sum_{j} \left( \sum_{l} (\theta_{sl;ij} - \bar{\theta}_{sl;ij}) B_{sl}(X_s) + \bar{B}_{ij}(X_s) \right) B_{ij}(X_t) = \sum_{l} B_{sl}(X_s) \bar{B}_{sl}(X_t)
\]
\[\text{(A.9)}\]

However, since \(\forall l \bar{B}_{sl}(X_t)\) is orthogonal to \(B_{ij}(X_t)\), the left and right hand sides of the above equation are equal to 0, which leads us to the following equations.
\[
\sum_{l} B_{sl}(X_s) \bar{B}_{sl}(X_t) = 0
\]
\[\text{(A.10)}\]

However, since we assumed that the sufficient statistics are minimal we get \(\forall l \bar{B}_{sl}(X_t) = 0\) from the first equality and \(\forall j, l \theta_{sl;ij} = \bar{\theta}_{sl;ij}, \bar{B}_{ij}(X_s) = 0\) from the second equality.

Hence from (A.7), we obtain \(E_s(X_t) = \theta_{sl}(B_s(X_t) - B_t(0))\) and \(E_t(X_s) = \theta_{st}^T(B_s(X_s) - B_s(0))\) where \(\theta_{sl}\) is a matrix formed by the scalars \(\theta_{sl;ij}\) such that \(\theta_{sl;ij} = \theta_{sl;ij}\) and:
\[
T[X_s \neq 0, X_t \neq 0]C_{s,t}(X_s, X_t) = (B_t(X_t) - B_t(0)) \times \theta_{sl}^T(B_s(X_s) - B_s(0))
\]
\[\text{(A.11)}\]

By extending this argument to higher order factors we can show that the natural parameters are required to be in the form specified by (7).

**B. Proof of Sparsistency**

Before proving the sparsistency result, we will show that the sufficient statistics \(B_v(X_r)\) are well behaved. Recall that \(B_v(X_r)\) indicates \(i\)th component of the vector \(B_v(X_r)\). We set the convention that whenever a variable has the subscript \(\gamma\) attached we will be referring to the set of indexes \(\{(t,j,k) : \gamma_{rj;tk} = \gamma_r, t \neq r\}\).

**Proposition 1.** Let \(X^{(j)}\) have joint distribution as in (10), then,
\[
P\left( \frac{1}{n} \sum_{j=1}^{n} B_{ri}(X^{(j)}) \geq \delta \right) \leq \exp\left(-\frac{n \delta^2}{4k^2_h}\right)
\]
\[\text{(B.12)}\]
for \(\delta \leq \min\{2 \frac{k_v}{\sqrt{n}}, k_h + k_v\} \).

**Proof.** It is clear from Taylor Series expansion and assumption 4 that
\[
\log E \left[ \exp\left(t B_{ri}(X_r)^2\right) \right] = \log \int_{\theta \in [\eta]} \exp\left(t B_{ri}(X_r)^2 + \sum_{s \in V} B_s(X_s) \times \theta^* + \sum_{t \in N(r)} \theta_{st}^r B_t(X_t) + \sum_{s \in V} C_s(X_s) - A(\theta^*) \right) v(dx)
\]
\[\text{(B.13)}\]

Using Assumption 4 we get the inequality:
\[
A(\theta^* + \bar{\nu}) - A(\theta^*) \leq k_m + \frac{k_v}{2}
\]

**Proposition 2.** Let \(X \) be a random vector with the distribution specified in (10). Then, for any positive constant \(\delta\) and some constant \(c > 0\)
\[
P\left( |B_{ri}(X_r)| \geq \delta \log(\eta) \right) \leq c\eta^{-\delta}
\]
\[\text{(B.15)}\]

**Proof.** Let \(\bar{\nu}\) be a unit vector with the same dimensions as \(\theta^*\), and exactly one non-zero entry, corresponding to the sufficient statistic \(B_{ri}(X_r)\). Then we can write
\[
\log\left( E[\exp(B_{ri}(X_r))] \right) = A(\theta^* + \bar{\nu}) - A(\theta^*)
\]

By Taylor series expansion, for some \(u \in [0, 1]\), we can rewrite last equation as
\[
A(\theta^* + \bar{\nu}) - A(\theta^*) = \nabla A(\theta^*) \cdot \bar{\nu} + \frac{1}{2} \bar{\nu}^T \nabla^2 A(\theta^* + u \bar{\nu}) \bar{\nu}
\]
\[= E[B_{ri}(X_r)] ||\bar{\nu}||_2 + \frac{1}{2} \frac{\partial^2 A(\theta^* + u \bar{\nu}) ||\bar{\nu}||_2^2}{\partial \theta_{ri}^2}
\]

Using Assumption 4 we get the inequality:
\[
A(\theta^* + \bar{\nu}) - A(\theta^*) \leq k_m + \frac{k_v}{2}
\]
Now, by using Chernoff bound, for any positive constant
a, we get \( P(B_{rt}(X_r) \geq a) \leq \exp(-a + km + \frac{1}{2}kh) \).
By setting \( a = \delta \log(\eta) \) it follows that
\[
P(B_{rt}(X_r) \geq \delta \log(\eta)) \leq \exp(-\delta \log(\eta) + km + \frac{1}{2}kh) \leq cn^{-\delta}\]
where \( c = \exp(km + \frac{1}{2}kh) \). \( \square \)

The proof of Sparsistency is based on the primal dual wit-
ness proof technique. First note that the optimality con-
dition of \((14)\), can be written as:
\[
\nabla \ell(\hat{\theta}_r; D) + \lambda_1 \sum_{t \neq r} \sqrt{\nu_{rt}} \hat{Z}_{1,rt} + \lambda_2 \hat{Z}_2 = 0 \tag{B.16}
\]
where \( \hat{Z}_{1,rt} \in \partial \| \hat{\theta}_r \|_2, \hat{Z}_2 \in \partial \| \hat{\theta}_r \|_1 \) and we de-
note \( \hat{Z} = (\hat{Z}_1, \hat{Z}_2) \), where \( \hat{Z}_1 = \{ \hat{Z}_{1,rt} \}_{t \neq r} \). And sub-
gradients \( \hat{Z}_1, \hat{Z}_2 \) should satisfy the following conditions:
\[
\forall i \ (\hat{Z}_2)_i = \text{sign} \left( (\hat{\theta}_r)_i \right) \text{ if } (\hat{\theta}_r)_i \neq 0 \\
\| (\hat{Z}_2)_i \|_1 \leq 1 \text{ otherwise}
\]
\[
\forall t \ |\hat{Z}_{1,rt}| = \frac{\hat{\theta}_{rt}}{\| \hat{\theta}_{rt} \|_2} \text{ if } \hat{\theta}_{rt} \neq 0 \\
\| \hat{Z}_{1,rt} \|_2 \leq 1 \text{ otherwise} \tag{B.17}
\]

Note that we can think of \( \hat{Z}_1 \) and \( \hat{Z}_2 \) as dual variables by
appealing to Lagrangian theory. The next lemma shows that graph structure recovery is guaranteed if the dual is
strictly feasible.

**Lemma 1.** Suppose that there exists a primal-dual pair
\((\hat{\theta}_r, \hat{Z})\) for \((14)\) such that \( \| \hat{Z}_{1,S^c} \|_{\infty,2} < 1 \) and \( \| \hat{Z}_{2,S^c} \|_{\infty} < 1 \). Then, any optimal solution \( \hat{\theta}_r \) must satisfy \( \hat{Z}_{2,S^c} = 0 \). Moreover, if the Hessian sub-matrix \( [\nabla^2 \ell(\hat{\theta}_r)]_{SS} \) is
positive definite then \( \hat{\theta}_r \) is the unique optimal solution.

**Proof.** First, note that by Cauchy–Schwarz’s and Holder’s
inequalities
\[
|\langle \hat{Z}_{1,rt}, \hat{\theta}_{rt} \rangle| \leq \| \hat{\theta}_{rt} \|_2 \| \hat{Z}_{2,rt} \|_1 \leq \| \hat{\theta}_{rt} \|_1 \cdot \| \hat{Z}_{2,rt} \|_2 \leq 1 \tag{B.18}
\]

But from \((B.16)\) and the primal optimality of \( \hat{\theta}_r \) and \( \hat{\theta}_r \) for
\((14)\),
\[
\ell(\hat{\theta}_r) + \sum_{t \neq r} \lambda_1 \sqrt{\nu_{rt}} \langle \hat{Z}_{1,rt}, \hat{\theta}_{rt} \rangle + \lambda_2 \langle \hat{Z}_2, \hat{\theta}_r \rangle \\
\geq \min_{\theta} \ell(\theta) + \sum_{t \neq r} \lambda_1 \sqrt{\nu_{rt}} \langle \hat{Z}_{1,rt}, \theta_{rt} \rangle + \lambda_2 \langle \hat{Z}_2, \theta_r \rangle \\
= \ell(\hat{\theta}_r) + \sum_{t \neq r} \lambda_1 \sqrt{\nu_{rt}} \| \hat{\theta}_{rt} \|_2 + \| \hat{\theta}_r \|_1 \\
= \ell(\hat{\theta}_r) + \sum_{t \neq r} \lambda_1 \sqrt{\nu_{rt}} \| \hat{\theta}_{rt} \|_2 + \| \hat{\theta}_r \|_1 \tag{B.19}
\]
hence, combining with \((B.18)\) with \((B.19)\) it follows that
\[
\sum_{t \neq r} \lambda_1 \sqrt{\nu_{rt}} \| \hat{\theta}_{rt} \|_2 + \| \hat{\theta}_r \|_1 = \sum_{t \neq r} \lambda_1 \sqrt{\nu_{rt}} \| \hat{\theta}_{rt} \|_2 + \| \hat{\theta}_r \|_1 = \sum_{t \neq r} \lambda_1 \sqrt{\nu_{rt}} \langle \hat{Z}_{1,rt}, \hat{\theta}_{rt} \rangle + \lambda_2 \langle \hat{Z}_2, \hat{\theta}_r \rangle. \tag{B.20}
\]

If the Hessian sub-matrix is positive definite for the
restricted problem then the problem is strictly convex and has
a unique solution. \( \square \)

Based on the above lemma, we prove sparsistency theorem
by constructing a primal-dual witness \((\hat{\theta}_r, \hat{Z})\) with the following steps:

1. Set \( \hat{\theta}_r \) = \text{argmin} \((\hat{\theta}_r)_{S^c} \) \( \ell(\hat{\theta}_r)_{S^c} \)
+ \lambda_1 \sum_{t \neq r} \sqrt{\nu_{rt}} \| \hat{\theta}_{rt} \|_2 + \lambda_2 \| \hat{\theta}_r \|_1 \)
= \lambda_1 \sum_{t \neq r} \sqrt{\nu_{rt}} \| \hat{\theta}_{rt} \|_2 + \lambda_2 \| \hat{\theta}_r \|_1

2. For \( t \in S \), we define \( \hat{Z}_{1,rt} = \frac{\hat{\theta}_{rt}}{\| \hat{\theta}_{rt} \|_2} \) and then con-
struct \( \hat{Z}_{2,S_r} \) by the stationary condition.

3. Set \( \hat{\theta}_r \) = \text{argmin} \((\hat{\theta}_r)_{S^c} \)

4. Set \( \hat{Z}_{2,S^c} \) such that \( \| \hat{Z}_{2,S^c} \|_{\infty,2} < 1 \)

5. Set \( \hat{Z}_{1,S_r} \) such that condition \((B.16)\) is satisfied.

6. The final step consists of showing, that the following
conditions are satisfied:

- \((a)\) strict dual feasibility : the condition in Lemma 1
holds with high probability
- \((b)\) correct neighbourhood recovery: the primal-dual pair
specifies the neighbourhood of \( r \), with high
probability

We begin by proving some key lemmas that are key to
our main theorem. The sub-gradient optimality condition
\((B.16)\) can be rewritten as:
\[
\nabla \ell(\hat{\theta}_r; D) - \nabla \ell(\hat{\theta}_r^*; D) = W^n - \lambda_1 \sum_{t \neq r} \sqrt{\nu_{rt}} \hat{Z}_{1,rt} - \lambda_2 \hat{Z}_2 \tag{B.21}
\]
where \( W^n = -\nabla \ell(\hat{\theta}_r^*; D) \) and \( \hat{\theta}_r^* \) is the true model
parameter. By applying mean-value theorem coordinate wise to \((B.20)\), we get:
\[
\nabla^2 \ell(\hat{\theta}_r^*; D)[\hat{\theta}_r^* \cdot \theta_r^*] = W^n - \lambda_1 \sum_{t \neq r} \sqrt{\nu_{rt}} \hat{Z}_{1,rt} - \lambda_2 \hat{Z}_2 + R^n \tag{B.21}
\]
where \( R^n \) is the remainder term after applying mean-value theorem: \( R^n_j = [\nabla^2 \ell(\hat{\theta}_r^*; D) - \nabla^2 \ell(\hat{\theta}_r^*; D)]^T (\hat{\theta}_r - \theta_r^*) \)
for some \( \hat{\theta}_r^* \) on the line between \( \hat{\theta}_r \) and \( \theta_r^* \), and with \( [\cdot]^T \)
denoting the j-th row of matrix. The following lemma
controls the score term \( W^n \)
Lemma 2. Recall $\nu_{r, \text{max}} = \max_{t} \nu_{r, t}, \nu_{r, \text{min}} = \min_{t} \nu_{r, t}, p' = \max(n, p)$, Assume that

$$\frac{2(2 - \alpha)}{\alpha} \sqrt{k_1(n, p) k_4^n \nu_{r, \text{max}} \log(\frac{p r_{\text{max}}}{\nu_{r, \text{max}}})} \leq \lambda_1 + \lambda_2 \leq \frac{4(2 - \alpha) \sqrt{\nu_{r, \text{min}}}}{\alpha \sqrt{r_{\text{max}}}} k_1(n, p) k_2(n, p) k_4$$

(B.22)

for some constant $k_4 \leq \min \{ \frac{2}{\nu_{r, t}}, k_h + k_v \}$ and suppose also that $n \geq \frac{8k^2}{k_4} \log(\sum t m_t)$ then,

$$P \left( \| W^n_r \|_{\infty, 2} > \frac{\alpha}{2 \alpha} \sqrt{\frac{nu_{r, \text{max}} (\lambda_1 + \lambda_2)}{4}} \right) \leq 1 - c_1 p^{-3} (\sum t m_t) - \exp(-c_2 n) - \exp(-c_3 n)$$

(B.23)

Proof. Define $W^n_r = -\nabla_{\theta_{r, t}} l(\theta_{r, t} ; \mathcal{D})$. Let $W^n_{i,j,k}$ be the element in $W^n_r$ corresponding to parameter $\theta_{r, j,k}$. Note that $W^n_{i,j,k} = \frac{1}{n} \sum_{t=1}^{n} V_{i,j,k}$ where

$$V_{i,j,k} = B_{ij} (X_{i}^{(t)}) B_{ik} (X_{i}^{(t)}) - \nabla_{\theta_{r, j,k}} A_r (\theta_{r}^{*} + \sum_{s \in V \setminus r} \theta_{r,s} B_s (X_{s}^{(t)})) B_{ik} (X_{i}^{(t)})$$

so for $t' \in \mathbb{R}$

$$E \left[ \exp \left( t' V_{i,j,k} \right) | X_{V \setminus r}^{(t)} \right] = \int_{X_{V \setminus r}^{(t)}} \exp \left[ t' B_{ij} (X_{i}^{(t)}) B_{ik} (X_{i}^{(t)}) - \nabla_{\theta_{r, j,k}} A_r (\theta_{r}^{*} + \sum_{s \in V \setminus r} \theta_{r,s} B_s (X_{s}^{(t)})) B_{ik} (X_{i}^{(t)}) \right]$$

$$+ C (X_{i}^{(t)}) + \theta_{r}^{*} B_{i} (X_{i}^{(t)}) + \sum_{s \in V \setminus r} B_{r} (X_{i}^{(t)}) \theta_{r,s} B_{s} (X_{s}^{(t)})$$

$$- A_r (\theta_{r}^{*} + \sum_{s \in V \setminus r} \theta_{r,s} B_{s} (X_{s}^{(t)})) \right] dX_{r}$$

$$= \exp \left[ A_r (\theta_{r}^{*} + t' B_{ik} (X_{i}^{(t)}) + \sum_{s \in V \setminus r} \theta_{r,s} B_{s} (X_{s}^{(t)})) \right]$$

$$- A_r (\theta_{r}^{*} + \sum_{s \in V \setminus r} \theta_{r,s} B_{s} (X_{s}^{(t)}))$$

$$- \nabla_{\theta_{r, j,k}} A_r \left( \theta_{r}^{*} + \sum_{s \in V \setminus r} \theta_{r,s} B_{s} (X_{s}^{(t)}) \right) t' B_{ik} (X_{i}^{(t)})$$

$$= \exp \left[ \nabla_{\theta_{r, j,k}, \theta_{r, j,k}} A_r (c \frac{c}{2} B_{ik} (X_{i}^{(t)})) \right]^{2} t'^{2}$$

where $c = \theta_{r}^{*} + \sum_{s \in V \setminus r} \theta_{r,s} B_{s} (X_{s}^{(t)}) + v_{t} t' B_{ik} (X_{i}^{(t)})$ for some $v_{t} \in [0, 1]$. Therefore,

$$\frac{1}{n} \sum_{i=1}^{n} \log \left[ \exp (t' V_{i,j,k}^{(i)}) | X_{V \setminus r}^{(i)} \right] = \frac{1}{2} \sum_{i=1}^{n} \nabla_{\theta_{r, j,k}, \theta_{r, j,k}} A_r (c \frac{c}{2} B_{ik} (X_{i}^{(t)}))^{2} t'^{2}$$

Next lets define event $\varepsilon_1 = \{ \max_{i=1}^{n} B_{i} (X_{i}^{(t)}) \leq 4 \log(p') \}$. Then, from Proposition 2 we get $P(\varepsilon_1) \leq c_1 n p^{-3} (\sum t m_t) \leq c_1 p^{-3} (\sum t m_t)$. If $t' \leq k_2(n, p)$, Assumption 5 implies that

$$\frac{1}{n} \sum_{i=1}^{n} \log \left[ \exp (t' V_{i,j,k}^{(i)}) | X_{V \setminus r}^{(i)} \right] \leq k_1(n, p) \frac{1}{n} \sum_{i=1}^{n} B_{ik} (X_{i}^{(t)})^{2} t'^{2}$$

Now, lets define event $\varepsilon_2 = \{ \max_{i=1}^{n} \sum_{t=1}^{n} (B_{i,j} (X_{i}^{(t)}))^{2} \leq k_4 \}$ where $k_4 \leq \min \{2k_v/3, k_h + k_v \}$. Then, by proposition (1) we obtain that if $n \geq \frac{8k^2}{k_4} \log(\sum t m_t)$:

$$P(\varepsilon_2) \leq \exp \left( -\frac{n k_2^2}{4k_h} + \log \left( \sum t m_t \right) \right) \leq \exp \left( -n c_2 \right)$$

(B.24)

Therefore, for $t' \leq k_2(n, p)$,

$$\frac{1}{n} \sum_{i=1}^{n} \log \left[ \exp (t' V_{i,j,k}^{(i)}) | X_{V \setminus r}^{(i)} \right] \leq \frac{k_1(n, p) k_2 t'^{2}}{2}$$

(B.25)

Hence, by the standard Chernoff bound technique, for $t' \leq k_2(n, p)$

$$P \left( \frac{1}{n} \sum_{i=1}^{n} |V_{i,j,k}^{(i)}| > \delta | \varepsilon_1, \varepsilon_2 \right) \leq 2 \exp \left( \frac{\frac{2}{k_1(n, p) k_2 t'^{2}} - t' \delta}{2} \right)$$

(B.26)

Setting $t' = \frac{\delta}{k_1(n, p) k_4}$, for $\delta \leq k_1(n, p) k_2(n, p) k_4$, we arrive to:

$$P \left( \frac{1}{n} \sum_{i=1}^{n} |V_{i,j,k}^{(i)}| > \delta | \varepsilon_1, \varepsilon_2 \right) \leq 2 \exp \left( \frac{\frac{\delta}{k_1(n, p) k_2 t'^{2}} - t' \delta}{2} \right)$$

(B.27)

Supposing that $\alpha \sqrt{\nu_{r, \text{min}}} \frac{4 \lambda_1 + \lambda_2}{4 \sqrt{\nu_{r, \text{max}}} m_r} \leq k_1(n, p) k_2(n, p) k_4$. It then follows that $\delta = \frac{\alpha \sqrt{\nu_{r, \text{min}}} \frac{4 \lambda_1 + \lambda_2}{4 \sqrt{\nu_{r, \text{max}}} m_r}}{2 - \alpha}$ satisfies

$$P \left( \frac{1}{n} \sum_{i=1}^{n} |V_{i,j,k}^{(i)}| > \frac{\alpha \sqrt{\nu_{r, \text{min}}} \frac{4 \lambda_1 + \lambda_2}{4 \sqrt{\nu_{r, \text{max}}} m_r} | \varepsilon_1, \varepsilon_2 \right) \leq 2 \exp \left( \frac{\frac{\delta}{2} \frac{\nu_{r, \text{min}} \alpha \sqrt{\nu_{r, \text{min}}} \frac{4 \lambda_1 + \lambda_2}{4 \sqrt{\nu_{r, \text{max}}} m_r} | \varepsilon_1, \varepsilon_2 \right)$$

(B.28)
Form which, we obtain the following using union bound

\[
P\left(\|W^n_t\|_2 > \frac{\alpha}{2-\alpha} \sqrt{\nu_{\min}(\lambda_1+\lambda_2)} \mid \varepsilon_1, \varepsilon_2\right) \leq 2 \exp\left(-\frac{\nu_{\min} n (\lambda_1+\lambda_2)^2}{2-\alpha} \log \nu_{\max} + \log \nu_{\max} + \log \nu_{\max}\right) \tag{B.29}
\]

and hence,

\[
P\left(\|W^n_t\|_\infty > \frac{\alpha}{2-\alpha} \sqrt{\nu_{\min}(\lambda_1+\lambda_2)} \mid \varepsilon_1, \varepsilon_2\right) \leq 2 \exp\left(-\frac{\nu_{\min} n (\lambda_1+\lambda_2)^2}{2-\alpha} \log \nu_{\max} + \log \nu_{\max} + \log \nu_{\max}\right) \tag{B.30}
\]

Finally for \(\lambda_1 + \lambda_2 \geq \frac{8(2-\alpha)}{\alpha} \sqrt{k_1(n,p)} k_4 \frac{\nu_{\min} \log(\nu_{\max})}{\text{max} \nu_{\max}}\), we obtain

\[
P\left(\|W^n_t\|_\infty > \frac{\alpha}{2-\alpha} \sqrt{\nu_{\min}(\lambda_1+\lambda_2)} \right) \leq c_1\nu^{-3} \left(\sum_i m_i\right) + \exp(-c_2n) + \exp(-c_3n) \tag{B.31}
\]

Lemma 3. Suppose that \(\lambda_1 + \lambda_2 \leq \frac{40\log p}{\text{max} d, k_3(n,p) \nu_{\min}^2} \) and \(\|W^n_t\|_\infty \leq \frac{5\sqrt{\nu_{\min}} \text{C}_{\min}}{4(2-\alpha)} \lambda_1 + \lambda_2\), then,

\[
P\left(\|\theta^{*}_t\|_S - (\hat{\theta}_t)_S \right) \leq 1 - \nu^{-3} \left(\sum_i m_i\right) \tag{B.32}
\]

for some constant \(c > 0\).

Proof. We define \(F(u_S)\) as:

\[
F(u_S) = \ell((\theta^{*}_t)_S + u_S; D) - \ell((\theta^{*}_t)_S; D) + \lambda_1 \sum_{t \in N(r)} \sqrt{\nu_{\min}} (\|\theta^{*}_t + u_{rt}\|_2 - \|\theta^{*}_t\|_2) + \lambda_2 (\|\theta^{*}_t\|_S + u_{S1} - \|\theta^{*}_t\|_S) \tag{B.33}
\]

From the construction of \(\hat{\theta}_t\), it is clear that \(\hat{u}_s = (\hat{\theta}_s)_S - (\theta^{*}_s)_S\) minimizes \(F\). And since \(F(0) = 0\), we have \(F(\hat{u}_s) = 0\). We now show that for some \(B > 0\) with \(\|u_S\|_\infty \leq B\), we have \(F(u_S) > 0\). Using this and the fact that \(F\) is convex we can then show that \(\|u_S\|_\infty \leq B\).

Let \(u_S\) an arbitrary vector with \(\|u_S\|_\infty \leq \frac{5\sqrt{\nu_{\min}} \text{C}_{\min}}{4(2-\alpha)} \lambda_1 + \lambda_2\). Then, from the Taylor Series expansion of log likelihood function in \(F\), we have:

\[
F(u_S) = \nabla \ell ((\theta^{*}_t)_S; D)^T u_S + u_S \nabla^2 \ell ((\theta^{*}_t)_S + v u_S) u_S + \lambda_1 \sum_{t \in N(r)} \sqrt{\nu_{rt}} (\|\theta^{*}_t + u_{rt}\|_2 - \|\theta^{*}_t\|_2) + \lambda_2 (\|\theta^{*}_t\|_S + u_{S1} - \|\theta^{*}_t\|_S) \tag{B.34}
\]

for some \(v \in [0, 1]\)

We now bound each of the terms in the right hand side of (B.34). From (B.29) and using Cauchy-Schwarz inequality we obtain:

\[
\left| \nabla \ell ((\theta^{*}_t)_S; D)^T u_S \right| \leq \|\nabla \ell ((\theta^{*}_t)_S; D)\|_\infty \|u_S\|_1 \leq \left|\nabla \ell ((\theta^{*}_t)_S; D)\|_\infty \|u_S\|_\infty \tag{B.35}
\]

where the last inequality holds because \(\alpha \in (0, 1]\). Moreover, from triangle inequality we have:

\[
\lambda_1 \sum_{t \in S} \sqrt{\nu_{rt}} (\|\theta^{*}_t + u_{rt}\|_2 - \|\theta^{*}_t\|_2) 
\geq \lambda_1 \sum_{t \in N(r)} \sqrt{\nu_{rt}} \|u_{rt}\|_2 
\geq \frac{\lambda_1}{\text{max} \nu_{\min}} d_r \nu_{\max} \lambda_1 (\lambda_1 + \lambda_2) \tag{B.36}
\]

Also,

\[
\lambda_2 (\|\theta^{*}_t\|_S + u_{S1} - \|\theta^{*}_t\|_S) 
\geq \lambda_2 \|u_{S1}\|_1 
\geq \lambda_2 \|u_{S1}\|_1 \tag{B.37}
\]

On the other hand, by Taylor’s approximation of \(\nabla^2 \ell\), there exists \(\alpha_{ik} \in [0, 1]\) and \(\tilde{u}_{ik}\) between \(\theta_{ik}\) and \(\theta_{ik} + v u_{ik}\) such that

\[
\Lambda_{\min} \left(\nabla^2 \ell ((\theta^{*}_t)_S + v u_S)\right) \geq \min_{\beta \in [0, 1]} \Lambda_{\min} \left(\nabla^2 \ell ((\theta^{*}_t)_S + \beta u_S)\right) \geq \Lambda_{\min} (Q_{\beta S}^2) - \max_{\nu \in [0, 1]} \sum_{m,t} \alpha_{ik} v \left(\nabla^3 A \left(\tilde{u}_{ik}\right)\right)_{klt} \tag{B.38}
\]
Consider the event $\varepsilon_1$ as defined in the previous proof. We know that $P(\varepsilon_1) \geq 1 - c_1 p^{-3} (\sum t m_t)$. Conditioned on $\varepsilon_1$ and using Assumption 5 we arrive to the following:
\[
\begin{align*}
\lambda_{\min} (\nabla^2 \ell ((\theta^*_r)_S + v u_S)) \\
\geq C_{\min} - 4 \log p' \|u_S\|_1 D_{\max} \nu_{r \max} k_3 (n, p) \\
\geq C_{\min} - 4 \log p' d_r \frac{\nu_{r \max}}{r_{\max}} \|u_S\|_{\infty, 2} D_{\max} \nu_{r \max} k_3 (n, p) \\
\geq C_{\min} - \frac{4 \log p' d_r}{2} \frac{\nu_{r \max}}{r_{\max}}.
\end{align*}
\]
where the last inequality holds for $\lambda_1 + \lambda_2 \leq \frac{4 \log p' d_r}{2} \frac{\nu_{r \max}}{r_{\max}}$.
Finally using the above bounds we arrive at the following:
\[
F(u_S) \geq d_r \nu_{r \max} \frac{5}{C_{\min}} (\lambda_1 + \lambda_2)^2 \left( -1 - \frac{1}{4} + \frac{5}{2} \right) > 0
\]
(40)
Therefore
\[
\|((\theta^*_r)_S - (\theta_r)_S)\|_{\infty, 2} \leq \frac{5}{C_{\min}} (\lambda_1 + \lambda_2)
\]
(41)

\textbf{Lemma 4.} Suppose that $\lambda_1 + \lambda_2 \leq \frac{\alpha}{2 - \alpha} \frac{\sqrt{\nu_{r \min}} c^2_{\lambda \alpha}}{\sqrt{\nu_{r \min}} d_r}$ and $\|W^*\|_{\infty, 2} \leq \frac{\alpha (\lambda_1 + \lambda_2) \nu_{r \max}}{4 (2 - \alpha)}$, then,
\[
P \left( \frac{\|R^n\|_{\infty, 2}}{\lambda_1 + \lambda_2} \leq \alpha \frac{\sqrt{\nu_{r \min}}}{4 (2 - \alpha)} \right) \geq 1 - c p^{-3} \left( \sum t m_t \right)
\]
(41)
for some constant $c > 0$.

\textbf{Proof.} Recall that $R^n_{t,j} = [\nabla^2 \ell ((\theta^*_r)_S + D) - \nabla^2 \ell ((\theta^*_r)_S + D)]^T_j (\theta_r - \theta^*_r)$ where $[\cdot]^T_j$ denotes the $j$-th row of a matrix. Let us also rewrite $R^n_{t,j,k}$ to the coordinate of $R^n$ corresponding to $\theta_{r,j,k}$. Then,
\[
R^n_{t,j,k} = \frac{1}{n} \sum_{i=1}^n B_{t,k}(X^n_i) \left[ \nabla^2 A_r \left( \begin{array}{c} \theta^*_r + \sum_{s \neq r} \theta_{r,s} B_s(X^n_i) \\ \theta^*_r + \sum_{s \neq r} \theta_{r,s} B_s(X^n_i) \end{array} \right) \right] - \\
\nabla^2 A_r \left( \begin{array}{c} \theta^*_r + \sum_{s \neq r} \theta_{r,s} B_s(X^n_i) \\ \theta^*_r + \sum_{s \neq r} \theta_{r,s} B_s(X^n_i) \end{array} \right) \] \otimes B^n_{i,j,k} \\
(42)
with $B^n_{i,j,k}$ the vector of sufficient statistics evolute at the $i$-th sample. Introducing the notation $(\theta_r, B^i_r) := (\theta_r + \sum_{s \neq r} \theta_{r,s} B_s(X^n_i))$, from the mean value theorem we obtain
\[
\nabla^2 A_r(\theta^*_r + \sum_{s \neq r} \theta_{r,s} B_s(X^n_i)) - \nabla^2 A_r(\theta^*_r + \sum_{s \neq r} \theta_{r,s} B_s(X^n_i)) = (\theta_r - \theta^*_r)
\]
(43)
where $\theta_r$ and $\theta^*_r$ are the parameters to the $\theta$-vector, respectively. Therefore, combining (B.42) with (B.43) and using basic properties of krocker product we obtain
\[
R^n_{t,j,k} \leq \frac{1}{n} \sum_{i=1}^n \left[ B_{t,k}(X^n_i) \right] \|B_{r,k}\|_{\infty, 2} D_{\max} \|\theta_r - \theta^*_r\|_2^2 \\
\leq 4 \log p' \nu_{r \max} k_3 (n, p) D_{\max} \|\theta_r - \theta^*_r\|_2^2
\]
(44)
which implies
\[
\|R^n\|_{\infty, 2} \leq \frac{4 \sqrt{\nu_{r \max}} \log p' \nu_{r \max} k_3 (n, p) D_{\max}}{\nu_{r \max}} \|\theta_r - \theta^*_r\|_2^2 \\
\leq \frac{\lambda_1 + \lambda_2}{\nu_{r \max}} \frac{4 \sqrt{\nu_{r \max}}}{2 - \alpha}
\]
(45)
with probability at least $1 - c p^{-3} (\sum t m_t)$. □

We now prove theorem 2 using lemmas 2-4. Recalling that $Q^n = \nabla^2 \ell ((\theta^*_r)_S + D)$ and the fact that we have set $(\hat{\theta}_r)_S^{sc} = 0$ in our primal-dual construction, we can rewrite condition (B.21) as the following equations:
\[
Q^n_{S,S}(Q^n_{S,S})^{-1} [W^n_S - 1 \nu_{r \max} \sqrt{\nu_{r \max}} Z_{1,r,l} - \lambda_2 \hat{Z}_{2,S}^{sc} + R^n_S] = 0
\]
(46)
Since the matrix $Q^n_{S,S}$ is invertible, the conditions (B.46) and (B.47) can be rewritten as :
\[
Q^n_{S,S}(Q^n_{S,S})^{-1} [W^n_S - 1 \nu_{r \max} \sqrt{\nu_{r \max}} Z_{1,r,l} - \lambda_2 \hat{Z}_{2,S}^{sc} + R^n_S] = 0
\]
(46)
\[
W^n_S - 1 \nu_{r \max} \sqrt{\nu_{r \max}} Z_{1,r,l} - \lambda_2 \hat{Z}_{2,S}^{sc} + R^n_S
\]
(47)
Rearranging yields the following condition:
\[
\lambda_1 \sum_{t \in N(r)} \sqrt{\nu_{r \max}} Z_{1,r,l} = \\
W^n_S + R^n_S - Q^n_{S,S}(Q^n_{S,S})^{-1} [W^n_S + R^n_S] - \lambda_2 \hat{Z}_{2,S}^{sc} + Q^n_{S,S}(Q^n_{S,S})^{-1} [1 \nu_{r \max} \sqrt{\nu_{r \max}} Z_{1,r,l} + \lambda_2 \hat{Z}_{2,S}^{sc}]
\]
(48)

\textbf{Strict Dual Feasibility:} we now show that for the dual sub-vector $\hat{Z}_{1,S}^{sc}$, we have $\|\hat{Z}_{1,S}^{sc}\|_{\infty, 2} < 1$. We get the following equation from (B.49), by applying triangle inequality:
\[ \lambda_1 \sqrt{\nu_{\min}} \left\| \hat{Z}_{1,S^c} \right\|_{\infty,2} \leq \]
\[ \lambda_1 \left\| \sum_{t \notin N(r)} \sqrt{\nu_t} \hat{Z}_{1,r} \right\|_{\infty,2} \leq \]
\[ \left( W^n \right)_{\infty,2} + \left( R^n \right)_{\infty,2} + \left( Q^n S \right)_{\infty,2} \left( Q^n S \right)^{-1} \left( \lambda_1 + \lambda_2 \right) \]

where \( \nu_{\min} = \min \nu_t \), \( \nu_{\max} = \max \nu_t \) and \( d_r = |N(r)| \)

Using mutual incoherence bound 2 on the above equation gives us:

\[ \left\| \hat{Z}_{1,S^c} \right\|_{\infty,2} \leq \]
\[ \frac{1}{\lambda_1 \sqrt{\nu_{\max}}} \left( W^n \right)_{\infty,2} + \left( R^n \right)_{\infty,2} + \left( Q^n S \right)_{\infty,2} \left( Q^n S \right)^{-1} \left( \lambda_1 + \lambda_2 \right) \]

Using the previous lemmas we obtain the following:

\[ \left\| \hat{Z}_{1,S^c} \right\|_{\infty,2} \leq \]
\[ \frac{1}{\lambda_1 \sqrt{\nu_{\max}}} \left[ \alpha \left( \lambda_1 + \lambda_2 \right) \right] + \lambda_2 \sqrt{\nu_{\max}} \left( 1 + \frac{m_{\min}}{m_{\max}} (1 - \alpha) \left( \lambda_1 + \lambda_2 \right) \right) \]

If \( \lambda_2 < \left( \frac{\alpha}{2 - \alpha + \frac{m_{\min}}{m_{\max}}} \right) \lambda_1 \), then,

\[ \left\| \hat{Z}_{1,S^c} \right\|_{\infty,2} < 1 \]

We have shown that the dual is strictly feasible with high probability and also the solution is unique. And hence based on Lemma 1 the method correctly excludes all edges not in the set of edges.

**Correct Neighbourhood Recovery:** To show that all correct neighbours are recovered, it suffices to show that

\[ \left\| (\theta^*_r)_S - (\hat{\theta}_r)_S \right\|_{\infty,2} \leq \frac{\theta_{\min}}{2} \]

where \( \theta_{\min} = \min_{t \notin V \setminus r} \left\| \theta_t \right\|_2 \).

Using Lemma 3 we can show the above inequality holds if

\[ \theta_{\min} \geq \frac{10}{c_{\max}} \left( \lambda_1 + \lambda_2 \right) \]

### C. Full MyFitnessPal Graph

Figure 1 shows a high-level view of the entire VS-MRF learned from the MyFitnessPal food database. The three macro-nutrients (fat, carbs, and protein) correspond to the three largest hubs with the remaining nine micro-nutrients representing smaller hubs.

---

**References**

Figure 1. Full MRF learned from the MyFitnessPal food database. The hubs correspond to point-inflated gamma nutrient nodes, with the three largest hubs being the macro-nutrients (fat, carbs, and protein).