Appendices to "Sparse Gaussian Conditional Random Fields: Algorithms, Theory, and Application to Energy Forecasting"

A. Derivation of Gradient, Hessian, and Differentials

Here we derive analytic expressions for the gradient, Hessian, and various differentials of the log likelihood function. Recall that the log-likelihood is given by

$$f(\Lambda, \Theta) = -\log|\Lambda| + \mathrm{tr}\Lambda S_{yy} + 2\mathrm{tr}\Theta^T S_{xy} + \mathrm{tr}\Lambda^{-1}\Theta^T S_{xx}\Theta \tag{1}$$

We adopt the differential matrix calculus notation from (Magnus & Neudecker, 1988) where for a matrix $A \in \mathbb{R}^{m\times n}$ and a function $f : \mathbb{R}^{m\times n} \to \mathbb{R}$, $d^k f(A; U) : \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n} \to \mathbb{R}$ denotes the k-th differential of the function f evaluated at U. For example, the first differential can easily be expressed in terms of the gradient

$$df(A; U) = \mathrm{tr}\nabla_A f(A)^T U. \tag{2}$$

Other derivatives for functions of matrices (i.e., Hessians or higher order terms) are cumbersome to represent directly, but the differentials themselves can typically be expressed compactly; indeed, it is often simplest to first derive these differentials and then use them to determine analytical expressions for the Hessians and higher order derivatives. Furthermore, the Taylor expansion of a function can be represented directly in terms of its differentials; for instance the second order approximation is given by

$$f(A+\Delta) \approx f(A) + df(A;\Delta) + \frac{1}{2}d^2 f(A;\Delta) \equiv f(A) + \mathrm{vec}(\nabla_A f(A))^T \mathrm{vec}(\Delta) + \frac{1}{2}\mathrm{vec}(\Delta)^T(\nabla_A^2 f(A))\,\mathrm{vec}(\Delta) \tag{3}$$

where vec denotes the vectorization of a matrix (forming a vector by concatenating the columns), and $\nabla_A^2 f(A)$ denotes the Hessian.

Using standard rules of differential calculus, we can compute the first and second order differentials of the log-likelihood $f(\Lambda, \Theta)$,

$$df(\Lambda, \Theta; U, V) = \mathrm{tr}S_{yy}U + 2\mathrm{tr}S_{yx}V - \mathrm{tr}\Lambda^{-1}U + 2\mathrm{tr}\Lambda^{-1}\Theta^T S_{xx}V - \mathrm{tr}\Lambda^{-1}\Theta^T S_{xx}\Theta\Lambda^{-1}U \tag{4}$$

and from this expression we can easily determine the relevant gradients

$$\begin{aligned} \nabla_\Lambda f(\Lambda, \Theta) &= S_{yy} - \Lambda^{-1} - \Lambda^{-1}\Theta^T S_{xx}\Theta\Lambda^{-1} \\ \nabla_\Theta f(\Lambda, \Theta) &= 2S_{xy} + 2S_{xx}\Theta\Lambda^{-1}. \end{aligned} \tag{5}$$

Similarly, we can differentiate again to find the second differential

$$d^2 f(\Lambda, \Theta; U, V) = 2\mathrm{tr}\Lambda^{-1}U\Lambda^{-1}\Theta^T S_{xx}\Theta\Lambda^{-1}U + \mathrm{tr}\Lambda^{-1}U\Lambda^{-1}U + 2\mathrm{tr}\Lambda^{-1}V^T S_{xx}V - 4\mathrm{tr}\Lambda^{-1}U\Lambda^{-1}\Theta^T S_{xx}V. \tag{6}$$

Combining the first and second differential gives the full second order Taylor expansion shown in the paper. It also lets us determine the Hessian itself, which we use in the incoherence condition for the theoretical results

$$\nabla^2_{\Lambda,\Theta} f(\Lambda, \Theta) = \begin{bmatrix} \Lambda^{-1} \otimes (\Lambda^{-1} + 2\Lambda^{-1}\Theta^T S_{xx}\Theta\Lambda^{-1}) & -2\Lambda^{-1} \otimes \Lambda^{-1}\Theta^T S_{xx} \\ -2\Lambda^{-1} \otimes S_{xx}\Theta\Lambda^{-1} & 2\Lambda^{-1} \otimes S_{xx} \end{bmatrix} \tag{7}$$

B. Detailed Description of Newton Coordinate Descent Method

We present a detailed description and full pseudo-code for the Newton coordinate descent algorithm. The derivation mirrors that in (Hsieh et al., 2011). The complete method is shown in Algorithm 1, with the coordinate

Algorithm 1 Newton Coordinate Descent for SGCRF

Input: Input features $X \in \mathbb{R}^{m \times n}$ and outputs $Y \in \mathbb{R}^{m \times p}$; regularization parameter λ; algorithm parameters ϵ, σ, α, β.
Output: Optimized parameters Λ, Θ
Initialize: $\Lambda \leftarrow I$, $\Theta \leftarrow 0$, $\Sigma \leftarrow \Lambda^{-1}$
while (not converged) **do**
1. Compute the gradient, determine active sets S_Λ, S_Θ using (14), and check for convergence.
2. Find regularized Newton direction D_Λ, D_Θ using Algorithm 2.
3. Initialize $\alpha \leftarrow 1$ and compute

$$\mu \leftarrow (\mathrm{tr}\nabla_\Lambda f(\Lambda,\Theta)^T \Delta_\Lambda + \mathrm{tr}\nabla_\Theta f(\Lambda,\Theta)^T \Delta_\Theta + \|\Lambda + \Delta_\Lambda\|_{1\star} + \|\Theta + \Delta_\Theta\|_1).$$

while (insufficient descent) **do**
1. Compute the Cholesky decomposition $LL^T = \Lambda + \alpha D_\Lambda$, continuing if not positive definite
2. Check descent $f(\Lambda + \alpha D_\Lambda, \Theta + \alpha D_\Theta) < f(\Lambda,\Theta) + \alpha\sigma\mu$ and break if satisfied
3. $\alpha \leftarrow \beta\alpha$

end while
end while

descent inner loop for computing the approximation to the Newton direction given in Algorithm 2. This process repeats until the solution converges to a within a specified tolerance, checked using the KKT conditions.

Next, we derive the coordinatewise updates for the inner loop and highlight the key optimizations that are used in order to achieve fast performance.

B.1. Coordinate descent updates for the Newton approximation

To begin, note that for a fixed Λ and Θ, the regularized Newton direction is given by the solution to the second-order Taylor expansion which for our problem has the form

$$\begin{aligned} h(\Delta_\Lambda, \Delta_\Theta) = \; & \mathrm{tr}S_{yy}\Delta_\Lambda + 2\mathrm{tr}S_{yx}\Delta_\Theta - \mathrm{tr}\Lambda^{-1}\Delta_\Lambda \; + 2\mathrm{tr}\Lambda^{-1}\Theta^T S_{xx}\Delta_\Theta - \\ & \mathrm{tr}\Lambda^{-1}\Theta^T S_{xx}\Theta\Lambda^{-1}\Delta_\Lambda + \mathrm{tr}\Lambda^{-1}\Delta_\Lambda\Lambda^{-1}\Theta^T S_{xx}\Theta\Lambda^{-1}\Delta_\Lambda - \frac{1}{2}\mathrm{tr}\Lambda^{-1}\Delta_\Lambda\Lambda^{-1}\Delta_\Lambda + \\ & \lambda\left(\|\Lambda + \Delta_\Lambda\|_{1,\star} + \|\Theta + \Delta_\Theta\|_1\right) \end{aligned} \tag{8}$$

We split the updates into three cases. First, we consider optimizing over a diagonal element of D_Λ by finding $\mu = \arg\min_\mu h(\Delta_\Lambda + \mu e_i e_i^T, \Delta_\Theta)$ which has the explicit form

$$\begin{aligned} \underset{\mu}{\text{minimize}} \; & \frac{1}{2}\mu^2\left[\Sigma_{ii}^2 + 2\Sigma_{ii}\Psi_{ii}\right] + \mu\left[-\Sigma_{ii} + (S_{yy})_{ii} - \Psi_{ii} + (\Sigma U\Sigma)_{ii} - 2(\Sigma\Theta^T S_{xx}V\Sigma)_{ii} + 2(\Psi U\Sigma)_{ii}\right] + \\ & \lambda|\Lambda_{ii} + \mu| \end{aligned} \tag{9}$$

where $\Sigma = \Lambda^{-1}$ and $\Psi = \Sigma\Theta^T S_{xx}\Theta\Sigma$.

Next, note that for two symmetric matrices A, B, not necessarily equal, the symmetric update is given by

$$\begin{aligned} & \arg\min_\mu \; \mathrm{tr}A(U + \mu(e_i e_j^T + e_j e_i^T))B(U + \mu(e_i e_j^T + e_j e_i^T)) \\ = \; & \arg\min_\mu \; \mu^2 \mathrm{tr}A(e_i e_j^T + e_j e_i^T)B(e_i e_j^T + e_j e_i^T) + \mu\mathrm{tr}AUB(e_i e_j^T + e_j e_i^T) + \mu\mathrm{tr}A(e_i e_j^T + e_j e_i^T)BU \\ = \; & \arg\min_\mu \; \mu^2(A_{ii}B_{jj} + 2A_{ij}B_{ij} + A_{jj}B_{ii}) + 2\mu((AUB)_{ij} + (AUB)_{ji}) \end{aligned} \tag{10}$$

Applying this equivalence twice, once with $A = B = \Sigma$ and again with $A = \Sigma$, $B = \Psi$ the the symmetric update

Algorithm 2 Coordinate descent inner loop

Input: S empirical covariance, λ regularization parameter, S_Λ, S_Θ active sets and Λ, Θ current estimates
Output: Approximate regularized Newton direction D_Λ, D_Θ
Initialize: $D_\Lambda \leftarrow 0$, $D_\Theta \leftarrow 0, U \leftarrow 0, V \leftarrow 0$
while (not converged) **do**
  for coordinate (i, j) in S_Λ **do**
    1. Find μ by solving (9) or (11), using U and V for efficiency.
    2. Symmetrically update D_Λ and two rows of U

$$\begin{aligned}(D_\Lambda)_{ij}, (D_\Lambda)_{ij} &\leftarrow (D_\Lambda)_{ij} + \mu \\ U_i &\leftarrow U_i + \mu\Sigma_j \\ U_j &\leftarrow U_j + \mu\Sigma_i\end{aligned}$$

    where Σ_i denotes the ith row of Λ^{-1}.
  end for
  for coordinate (i, j) in S_Θ **do**
    1. Find μ by solving (12), using U and V for efficiency.
    2. Update D_Θ and one row of V

$$\begin{aligned}(D_\Theta)_{ij} &\leftarrow (D_\Theta)_{ij} + \mu \\ V_i &\leftarrow V_i + \mu\Sigma_j\end{aligned}$$

  end for
end while

for an off-diagonal element of matrix D_Λ, $\mu = \arg\min_\mu h(\Delta_\Lambda + \mu(e_i e_j^T + e_j e_i^T), \Delta_\Theta)$ is given by

$$\begin{aligned}\underset{\mu}{\text{minimize}}\ \ & \mu^2\left[\Sigma_{ij}^2 + \Sigma_{ii}\Sigma_{jj} + \Sigma_{ii}\Psi_{jj} + 2\Sigma_{ij}\Psi_{ij} + \Sigma_{jj}\Psi_{ii}\right] + \\ & 2\mu\left[-\Sigma_{ij} + (S_{yy})_{ij} - \Psi_{ij} + (\Sigma U\Sigma)_{ij} - \Phi_{ij} - \Phi_{ji} + (\Psi U\Sigma)_{ij} + (\Psi U\Sigma)_{ji}\right] + \\ & 2\lambda|\Lambda_{ij} + U_{ij} + \mu|\end{aligned} \tag{11}$$

where $\Phi = \Sigma\Theta^T S_{xx} V\Sigma$. Finally, we consider optimizing over an element of D_Θ

$$\begin{aligned}\underset{\mu}{\text{minimize}}\ \ & \mu^2\left[\Sigma_{jj}(S_{xx})_{ii}\right] + \mu\left[2(S_{xy})_{ij} + 2(S_{xx}\Theta\Sigma)_{ij} + 2(S_{xx}V\Sigma)_{ij} - 2(S_{xx}\Theta\Sigma U\Sigma)_{ij}\right] + \\ & \lambda|\Theta_{ij} + V_{ij} + \mu|\end{aligned} \tag{12}$$

Each equation has a quadratic form and thus can be solved in closed form. The second two have an ℓ_1 penalty and the form $\min_\mu \frac{1}{2}a\mu^2 + b\mu + \lambda|c + \mu|$ which has the solution

$$\mu = -c + S_{\lambda/a}\left(c - \frac{b}{a}\right) \tag{13}$$

B.2. Optimizations

As in the case of the MRF (Hsieh et al., 2011), there are several modifications to a naive solution that significantly reduce the running time of the algorithm.

First, consider the matrix products involved in the coordinatewise updates above. A naive implementation of the coordinate descent algorithm would require $O((n+p)^2)$ operations even though the majority of the elements are unchanged from one iteration. However, by caching products of the static matrices and maintaining a factorized form of the products involving Δ_Λ and Δ_Θ, specifically $U = \Delta_\Lambda\Sigma$, $V = \Delta_\Theta\Sigma$, we reduce this to $O((n+p))$. As a consequence, at each iteration of the loop we must update the rows of U and V corresponding to the coordinates of Δ_Λ and Δ_Θ.

Next, we describe how we drastically reduce the coordinate descent active set. At each iteration of the outer loop, we fix the active set using the current nonzero coordinates and the KKT conditions of the objective function. We include a coordinate of Λ, respectively Θ, if

$$\begin{aligned} |(\nabla_\Lambda f(\Lambda, \Theta))_{i,j}| > \lambda \text{ or } \Lambda_{ij} \neq 0 \\ |(\nabla_\Theta f(\Lambda, \Theta))_{i,j}| > \lambda \text{ or } \Theta_{ij} \neq 0. \end{aligned} \tag{14}$$

Since the size of this active set is determined by the number of nonzero elements in the parameters, for sparse solutions the speed up is very significant. Note that although we fix the active set before beginning coordinate descent, as in the MRF case (Hsieh et al., 2011), we still have convergence guarantees for the overall algorithm.

Finally, note that when taking a step we must ensure sufficient descent and that the Λ parameter remains in the semidefinite cone. We accomplish this using the Cholesky decomposition, which is also used for efficiently computing Λ^{-1}.

C. Theoretical Analysis

We will make the following assumptions about the input and output variables X and Y. These mirror similar assumptions in (Wainwright, 2009) and (Ravikumar et al., 2011), and we will discuss the precise differences.

First, the analysis here proceeds on the assumption that there is a true underlying model generating the test data, of the prescribed form (i.e., the data is generated according to a sparse Gaussian CRF). It is trivial to extend this analysis to the case of sub-Gaussian noise, but we simply assume Gaussian noise for simplicity of presentation

Assumption 1. ***Underlying model*** *The data is generated according to*

$$y|x \sim \mathcal{N}(-\Lambda^{\star\,-1}\Theta^{\star\,T}x, \Lambda^{\star\,-1}). \tag{15}$$

where each row of $[\Lambda^\star\ \Theta^{\star\,T}]$ *has at most* d *nonzero entries (i.e., the vertices corresponding to output variables in the graphical model of the CRF have maximum degree* d*).*

For simplicity, we will also denote $\Sigma^\star = \Lambda^{\star\,-1}$.

Our second assumption is a restricted convexity requirement, which ensures that the optimization problem restricted to the active set is unique. This is a common assumption for ℓ_1 approaches (the same condition appears in the least-squares analysis of (Wainwright, 2009)), and the only extension here is that we require this to hold for each output variable.

Assumption 2. ***Restricted convexity*** *For each output* i*, let* S_i *denote* $\{j : \Theta_{ji} \neq 0\}$ *(i.e.,* S_i *is the "active set" of edges directly connecting an input to* y_i*), we have that*

$$\lambda_{\min}\left(\frac{1}{m}X_{S_i}^T X_{S_i}\right) > 0. \tag{16}$$

The next assumption is more subtle (and quite strict in practice), but is again typical for exact subset selection proofs for ℓ_1 approaches. Namely, we require a mutual incoherence assumption, which effectively ensures that the connections in the CRF that correspond to the "true" edges do not correlate too much with edges that are not the support set.

Assumption 3. ***Mutual incoherence*** *Let* S *denote the active set of all variables in vector form*

$$S = \begin{bmatrix} \text{vec}(\text{supp}\{\Lambda^\star\}) \\ \text{vec}(\text{supp}\{\Theta^\star\}) \end{bmatrix} \tag{17}$$

where supp *denotes the support function (the indicator of whether an element is nonzero), and let* $\bar{S}$ *denote its complement. Then for* $H = \nabla^2_{\Lambda,\Theta} f(\Lambda, \Theta)$ *defined above*

$$\left|\left|\left| H_{\bar{S}S}(H_{SS})^{-1} \right|\right|\right|_\infty \leq 1 - \alpha \tag{18}$$

for some $\alpha > 1$*, where* $|||\cdot|||_\infty$ *denotes the matrix infinity norm, the maximum absolute row sum.*

Our first lemma shows that the gradients $\nabla_\Lambda f(\Lambda^\star, \Theta^\star)$ and $\nabla_\Theta f(\Lambda^\star, \Theta^\star)$ (the gradients evaluated at the true parameters) are small (in infinity norm) with high probability given samples on the order of $m = \Omega(\log n + \log p)$. The proof (shown below) follows from a standard bound on Gaussian random variables, and from Lemma 1 in (Ravikumar et al., 2011).

Lemma 1. *Given data generated by the model in Assumption 1 we have that*

$$P\left(\|\nabla_\Theta f(\Lambda^\star, \Theta^\star)\|_\infty > \epsilon\right) \leq 2np \exp\left\{-\frac{m\epsilon^2}{8c_{\sigma^\star}^2 c_X^2}\right\} \tag{19}$$

where $c_{\sigma^\star} = \max_i \Sigma^\star_{ii}$, *and* $c_X = \max_{j=1,\ldots,n} \|X_j\|_2/\sqrt{m}$; *the maximum normalized* ℓ_2 *norm over columns of* X. *Furthermore,*

$$P\left(\|\nabla_\Lambda f(\Lambda^\star, \Theta^\star)\|_\infty > \epsilon\right) \leq 4p^2 \exp\left\{-\frac{m\epsilon^2}{3200c_{\sigma^\star}^2}\right\} \tag{20}$$

for $0 < \delta < 40c_{\sigma^\star}$.

The next lemma is a generic primal-dual witness approach, mirroring exactly the derivation in (Wainwright, 2009), but presented in a generic form. For the presentation here, we will use a generic optimization problem minimize $f(\theta) + \lambda\|\theta\|_1$, though we will apply this specifically to our CRF problem momentarily. Intuitively, the lemma states conditions for which optimizing over the known support set is equivalent to optimizing with the ℓ_1 penalty.

Lemma 2. *Consider some sparse* $\theta^\star$ *with* $S = \mathrm{supp}(\theta^\star)$, *and consider the two optimization problems*

$$\begin{aligned} \hat{\theta} &= \arg\min_\theta f(\theta) + \lambda\|\theta\|_1 \\ \tilde{\theta} &= \arg\min_{\theta, \theta_{\bar{S}}=0} f(\theta) + \lambda\|\theta\|_1. \end{aligned} \tag{21}$$

Define $\Delta = \tilde{\theta} - \theta^\star$, *and* $R(\Delta) = -\nabla_\theta f(\tilde{\theta}) + \nabla_\theta f(\theta^\star) + \nabla^2_\theta f(\theta^\star)\Delta$. *Then if the following conditions hold*

1. *The solution* $\tilde{\theta}$ *is unique.*
2. *Mutual incoherence holds, i.e.,* $\|(\nabla^2_\theta f(\theta^\star))_{\bar{S}S}(\nabla^2_\theta f(\theta^\star))^{-1}_{SS}\|_\infty \leq 1 - \alpha$
3. $\max\{\|\nabla_\theta f(\theta^\star)\|_\infty, \|R(\Delta)\|_\infty\} \leq \lambda\alpha/8$

then the ℓ_1 *solution recovers the restricted solution,* $\tilde{\theta} = \hat{\theta}$.

In our setting, the Δ and $R(\Delta)$ are themselves matrices and thus slightly more complex. Thus, for subsequent lemmas, we define the following terms that we use to quantity the error of the second order Taylor expansions of our particular log-likelihood, evaluated at the true parameters. For any Λ, Θ, we define $\Delta_\Lambda = \Lambda - \Lambda^\star$ and $\Delta_\Theta = \Theta - \Theta^\star$ and

$$\Delta \equiv \begin{bmatrix} \Lambda \\ \Theta \end{bmatrix}. \tag{22}$$

Define

$$\begin{aligned} R_\Lambda(\Delta_\Lambda, \Delta_\Theta) &= \nabla_\Lambda f(\Lambda^\star, \Theta^\star) - \nabla_\Lambda f(\Lambda^\star + \Delta_\Lambda, \Theta^\star + \Delta_\Theta) + d(\nabla_\Lambda f(\Lambda^\star, \Theta^\star); \Delta_\Lambda, \Delta_\Theta) \\ R_\Theta(\Delta_\Lambda, \Delta_\Theta) &= \nabla_\Theta f(\Lambda^\star, \Theta^\star) - \nabla_\Theta f(\Lambda^\star + \Delta_\Lambda, \Theta^\star + \Delta_\Theta) + d(\nabla_\Theta f(\Lambda^\star, \Theta^\star); \Delta_\Lambda, \Delta_\Theta), \end{aligned} \tag{23}$$

which are the residuals of the first order Taylor expansion of the gradient (i.e., the errors in the second order Taylor expansion of the function itself), and

$$R(\Delta) = \begin{bmatrix} R_\Lambda(\Delta_\Lambda, \Delta_\Theta) \\ R_\Theta(\Delta_\Lambda, \Delta_\Theta) \end{bmatrix}. \tag{24}$$

The next lemma bounds the residual $\|R(\Delta)\|_\infty$ in terms of the distance from the true parameters, $\|\Delta\|_\infty$.

Lemma 3. *Under the definitions above, if*

$$\|\Delta\|_\infty \leq \frac{1}{d} \min\left\{ \frac{1}{3c_{\Sigma^\star}}, \frac{c_{\Theta^\star}}{2} \right\} \tag{25}$$

then

$$\|R(\Delta)\|_\infty \leq 206 c_{\Sigma^\star}^4 c_{\Theta^\star}^2 c_X^2 d^2 \|\Delta\|_\infty^2 \tag{26}$$

where $c_{\Sigma^\star} = \max_{i,j} \Sigma^\star_{ij}$ *and* $c_{\Theta^\star} = \max_{i,j} \Theta^\star_{ij}$.

Finally, we show that when the gradient evaluated at the true model are sufficiently small, then $\|\Delta\|_\infty$ itself is small.

Lemma 4. *Under the model above, suppose that*

$$\max\{\|\nabla_\Lambda f(\Lambda^\star, \Theta^\star)\|_\infty, \|\nabla_\Theta f(\Lambda^\star, \Theta^\star)\|_\infty\} \leq \frac{1}{2c_{H^\star}} \left[\min\left\{ \frac{1}{3c_{\Sigma^\star} d}, \frac{1}{412 c_{\Sigma^\star}^4 c_{\Theta^\star}^2 c_X^2 d^2} \right\} - \lambda \right]. \tag{27}$$

Then

$$\|\Delta\|_\infty \leq 2c_{H^\star}(\max\{\|\nabla_\Lambda f(\Lambda^\star, \Theta^\star)\|_\infty, \|\nabla_\Theta f(\Lambda^\star, \Theta^\star)\|_\infty\} + \lambda) \tag{28}$$

where $c_{H^\star} = \max_{i,j} H^\star_{ij}$, *the maximum element of the Hessian evaluated at the true parameters.*

These elements allow us to prove the desired theorem.

Theorem 1. *Using assumptions 1-3 above, suppose we have sample size*

$$m \geq 412^2 C^2 d^4 (1 + 8/\alpha)^2 \log(pn) \tag{29}$$

where $C = \max\{3c_{\Sigma^\star}, c_{\Theta^*}^{-1}, c_{\Sigma^\star}^4 c_{\Sigma^\star}^2 c_X^2\}$ *and choose* λ *as*

$$\lambda \geq (8/\alpha) c_{\sigma^\star} c_X \sqrt{3200} \sqrt{\frac{\log(pn) + \log 4}{m}} \tag{30}$$

then with probability greater than $1 - c_1 \exp(-c_2 m \lambda^2)$ *we have*

1. *The solution to the* ℓ_1 *regularized optimization problem,* $\tilde{\Lambda}$, $\tilde{\Theta}$, *has nonzero entries that are a strict subset of the nonzero entries of* $\Lambda^\star$, $\Theta^\star$
2. *The solution satisfies the elementwise bounds*

$$\max\{\|\tilde{\Lambda} - \Lambda^\star\|_\infty, \|\tilde{\Theta} - \Theta^\star\|_\infty\} \leq 2(1 + 8\alpha^{-1}) c_{H^\star} c_{\sigma^\star} c_X \sqrt{3200} \sqrt{\frac{\log(pn) + \log 4}{m}} \tag{31}$$

Proof. Let

$$\delta = c_{\sigma^\star} c_X \sqrt{3200} \sqrt{\frac{\log(pn) + \log 4}{m}}. \tag{32}$$

Then by Lemma 1 and the minimum bound on m we have that

$$\max\{\|\nabla_\Theta f(\Lambda^\star, \Theta^\star)\|_\infty, \|\nabla_\Lambda f(\Lambda^\star, \Theta^\star)\|_\infty\} \leq \delta \tag{33}$$

with probability greater than $1 - c_1 \exp(-c_2 m \lambda^2)$; we proceed with the proof conditioned on this event. Next, note by our choice of λ we have that $\delta \leq \alpha\lambda/8$ and thus the first half of the third condition for Lemma 2 holds. It remains to show that $R(\Delta) \leq \alpha\lambda/8$. By our minimum bound on m and our choice of λ we have that

$$\left(1 + \frac{8}{\alpha}\right) \delta \leq \frac{1}{2c_{H^\star}} \min\left\{ \frac{1}{3c_{\Sigma^\star} d}, \frac{1}{412 c_{\Sigma^\star}^4 c_{\Theta^\star}^2 c_X^2 d^2} \right\} \tag{34}$$

and thus Lemma 4 applies, which gives

$$\|\Delta\|_\infty \leq 2c_{H^\star} \left(1 + \frac{8}{a}\right) \delta. \tag{35}$$

Therefore, the assumption of $\|\Delta\|_\infty \leq \frac{1}{d}\min\{\frac{1}{3c_{\Sigma^\star}}, \frac{c_{\Theta^\star}}{2}\}$ holds and we apply Lemma 3 to establish

$$\begin{aligned}\|R(\Delta)\|_\infty &\leq 206 c_{\Sigma^\star}^4 c_{\Theta^\star}^2 c_X^2 d^2 \|\Delta\|_\infty^2 \\ &\leq 824 c_{\Sigma^\star}^4 c_{\Theta^\star}^2 c_X^2 d^2 c_{H^\star} \left(1+\frac{8}{a}\right)^2 \delta^2 \\ &\leq \left[824 c_{\Sigma^\star}^4 c_{\Theta^\star}^2 c_X^2 d^2 c_{H^\star}\left(1+\frac{8}{a}\right)^2 \delta\right]\frac{\alpha\lambda}{8} \\ &\leq \frac{\alpha\lambda}{8}.\end{aligned} \tag{36}$$

Finally, note that our Assumption 2 implies that the solution $(\tilde{\Lambda}, \tilde{\Theta})$ is unique and thus combined with the above derivation and Assumption 3, we the conditions for Lemma 2 and thus we conclude that $(\tilde{\Lambda}, \tilde{\Theta}) = (\hat{\Lambda}, \hat{\Theta})$ and the thus claim 1 and 2 are satisfied. □

C.1. Proofs of Lemmas

Proof. (of Lemma 1) Let X be given and assuming that Y is generated according to our model. We first consider $P\left(\|\nabla_\Theta f(\Lambda^\star, \Theta^\star)\|_\infty > \epsilon\right)$; as shown in Appendix A, we have

$$\nabla_\Theta f(\Lambda, \Theta) = 2S_{xy} + 2S_{xx}\Theta\Lambda^{-1}. \tag{37}$$

Writing $\beta^* = -\Theta^*\Lambda^{*-1}$ and $\Sigma^* = \Lambda^{*-1}$, we have $Y = X\beta^* + Z$ where $Z \in \mathbb{R}^{m\times p}$ has rows $Z_i \sim \mathcal{N}(0, \Sigma^*)$, and thus

$$2S_{xy} + 2S_{xx}\Theta^*\Lambda^{*-1} = \frac{2}{m}(X^TY - X^TX\beta^*) = \frac{2}{m}X^TZ. \tag{38}$$

By our assumptions that $\|X_j\|_2/\sqrt{n} < c_X$ for all columns of X and the maximum diagonal entry $\Sigma^\star$ is $c_{\sigma^\star}^2$, we have

$$\mathrm{Var}\left(\frac{2}{m}X_i^TZ_j\right) \leq \frac{4c_{\sigma^\star}^2 c_X^2}{m} \tag{39}$$

for any columns X_i and Z_j. Therefore by the union bound and Gaussian tail probability we have

$$P\left(\|\frac{1}{m}X^TZ\|_\infty > \epsilon\right) \leq 2np\exp\left\{-\frac{m\epsilon^2}{8c_{\sigma^*}^2 c_X^2}\right\} \tag{40}$$

Next, we consider $P\left(\|\nabla_\Lambda f(\Lambda^\star, \Theta^\star)\|_\infty > \epsilon\right)$ and again from Appendix A, we have

$$\nabla_\Lambda f(\Lambda, \Theta) = S_{yy} - \Lambda^{-1} - \Lambda^{-1}\Theta^T S_{xx}\Theta\Lambda^{-1} \tag{41}$$

which we can rewrite as

$$S_{yy} - \Lambda^{*-1} - \Lambda^{*-1}\Theta^{*T}S_{xx}\Theta^*\Lambda^{*-1} = \frac{1}{m}Z^TZ - \Sigma^*. \tag{42}$$

Now we can apply Lemma 1 in (Ravikumar et al., 2011) and arrive at the desired bound

$$P(\|\frac{1}{m}Z^TZ - \Sigma^*\|_\infty > \epsilon) < 4p^2\exp\left\{-\frac{m\epsilon^2}{3200c_{\sigma^\star}^2}\right\} \tag{43}$$

for $0 < \epsilon < 40c_{\sigma^\star}$. □

Proof. (of Lemma 2)

The proof here proceeds exactly as in (Wainwright, 2009) and (Ravikumar et al., 2011), so we describe it relatively quickly. The goal is to show that when solve the restricted problem, the resulting $\tilde{\theta}$ (which is zero outside the support set S) is also optimal for the full ℓ_1 problem. Defining $\Delta = \tilde{\theta} - \theta^\star$, the the full ℓ_1 optimization problem can be written as

$$\nabla_\theta^2 f(\theta^\star)\Delta + \nabla_\theta f(\theta^\star) - R(\Delta) + \lambda z = 0. \tag{44}$$

If we can show that $\|z\|_\infty < 1$, then $\tilde{\theta}$ is an optimal solution to the original ℓ_1 problem, so $\tilde{\theta} = \hat{\theta}$. Furthermore, the solution $\hat{\theta}$ cannot have support outside the support of $\theta^\star$.

We can write the above optimality condition in terms of S and $\bar{S}$, using $H = \nabla^2_\theta f(\theta^\star)$ and $g = \nabla_\theta f(\theta^\star)$ for simplicity

$$\begin{bmatrix} H_{SS} & H_{S\bar{S}} \\ H_{\bar{S}S} & H_{\bar{S}\bar{S}} \end{bmatrix} \begin{bmatrix} \Delta_S \\ 0 \end{bmatrix} + \begin{bmatrix} g_S \\ g_{\bar{S}} \end{bmatrix} + \begin{bmatrix} R(\Delta)_S \\ R(\Delta)_{\bar{S}} \end{bmatrix} + \lambda \begin{bmatrix} z_S \\ z_{\bar{S}} \end{bmatrix} \tag{45}$$

Using the fact that

$$\Delta_S = H_{SS}^{-1}(R(\Delta)_S - g_S - \lambda z_S) \tag{46}$$

we can solve for$z_{\bar{S}}$ gives

$$z_{\bar{S}} = -\frac{1}{\lambda} H_{\bar{S}S}\Delta_S + \frac{1}{\lambda}(R(\Delta)_S - g_{\bar{S}}) = \frac{1}{\lambda} H_{\bar{S}S}H_{SS}^{-1}(g_S - R(\Delta)_S) + H_{\bar{S}S}H_{SS}^{-1} z_S + \frac{1}{\lambda}(R(\Delta)_S - g_S) \tag{47}$$

Thus

$$\|z_{\bar{S}}\|_\infty \leq \frac{2-\alpha}{\lambda}(\|g\|_\infty + \|R(\Delta)\|_\infty) + 1 - \alpha \leq \frac{2-\alpha}{\lambda}\frac{\alpha\lambda}{4} + 1 - \alpha < 1. \tag{48}$$

□

Proof. (of Lemma 3) Since $R(\Delta)$ is the residual of the first order Taylor expansion of the likelihood gradient, by the mean value theorem we have that the exists $t \in [0, 1]$ such that

$$R_\Lambda(\Delta_\Lambda, \Delta_\Theta) = d^2(\nabla_\Lambda f(\Lambda^\star + t\Delta_\Lambda, \Theta^\star + t\Delta_\Lambda); \Delta_\Lambda, \Delta_\Theta) \tag{49}$$

and similarly for $R_\Theta(\Delta_\Lambda, \Delta_\Theta)$. The first and second differentials of these gradient terms are given by (note that since these are the differentials of a matrix-valued function, we cannot simplify as many of the expressions as we did for the differential of the likelihood function)

$$\begin{aligned}
d(\nabla_\Lambda f(\Lambda^\star, \Theta^\star); U, V) &= \Lambda^{-1}U\Lambda^{-1} + \Lambda^{-1}U\Lambda^{-1}\Theta^T S_{xx}\Theta\Lambda^{-1} + \Lambda^{-1}\Theta^T S_{xx}\Theta\Lambda^{-1}U\Lambda^{-1} - \\
&\quad \Lambda^{-1}V^T S_{xx}\Theta\Lambda^{-1} - \Lambda^{-1}\Theta^T S_{xx}V\Lambda^{-1} \\
d^2(\nabla_\Lambda f(\Lambda^\star, \Theta^\star); U, V) &= -2\Lambda^{-1}U\Lambda^{-1}U\Lambda^{-1} - 2\Lambda^{-1}U\Lambda^{-1}U\Lambda^{-1}\Theta^T S_{xx}\Theta\Lambda^{-1} - \\
&\quad 2\Lambda^{-1}U\Lambda^{-1}\Theta^T S_{xx}\Theta\Lambda^{-1}U\Lambda^{-1} - 2\Lambda^{-1}\Theta^T S_{xx}\Theta\Lambda^{-1}U\Lambda^{-1}U\Lambda^{-1} + \\
&\quad 2\Lambda^{-1}U\Lambda^{-1}V^T S_{xx}\Theta\Lambda^{-1} + 2\Lambda^{-1}U\Lambda^{-1}\Theta^T S_{xx}V\Lambda^{-1} + \\
&\quad 2\Lambda^{-1}V^T S_{xx}\Theta\Lambda^{-1}U\Lambda^{-1} + 2\Lambda^{-1}\Theta^T S_{xx}V\Lambda^{-1}U\Lambda^{-1} - \\
&\quad 2\Lambda^{-1}V^T S_{xx}V\Lambda^{-1} \\
d(\nabla_\Theta f(\Lambda^\star, \Theta^\star); U, V) &= -2S_{xx}\Theta\Lambda^{-1}U\Lambda^{-1} + 2S_{xx}V\Lambda^{-1} \\
d^2(\nabla_\Theta f(\Lambda^\star, \Theta^\star); U, V) &= 4S_{xx}\Theta\Lambda^{-1}U\Lambda^{-1}U\Lambda^{-1} - 4S_{xx}V\Lambda^{-1}U\Lambda^{-1}
\end{aligned} \tag{50}$$

For example, $R_\Lambda(\Delta_\Lambda, \Delta_\Theta)$ is equal to the second expression with the $\Lambda^\star$ terms replaced by $\Lambda^\star + \Delta_\Lambda$, the $\Theta^\star$ terms replaced by $\Theta^\star + \Delta_\Theta$ and U and V replaced by Δ_Λ and Δ_Θ respectively. To bound $R(\Delta)$, we bound each of these terms individually.

Although the expression is rather lengthy, note that each term in the second differentials has a quadratic expression in Δ_Λ and Δ_θ, with at most four $(\Lambda^\star + t\Delta_\Lambda)^{-1}$ terms, two $\Theta^\star + t\Delta_\Theta$ terms and one S_{xx} term. Furthermore, we use the fact that

$$\|ABC\|_\infty = \|(C^T \otimes A)\,\mathrm{vec}(B)\|_\infty \leq |||C|||_1 |||A|||_\infty \|B\|_\infty \tag{51}$$

to place the vector infinity norm around the S_{xx} term in all cases, since $\|S_{xx}\|_\infty \leq c_X^2$. Thus, each term in the second differential is bounded by

$$c_X^2 |||(\Lambda^\star + t\Delta_\Lambda)^{-1}|||_\infty^4 |||\Theta^\star + t\Delta_\Theta|||_1^2 |||\Delta|||_1 \tag{52}$$

Now, first note that since Δ as a most d entries per column

$$|||\Delta|||_1 \leq d\|\Delta\|_\infty. \tag{53}$$

Now, note that

$$(\Lambda^\star + t\Delta_\Lambda)^{-1} = (I + t\Lambda^{\star -1}\Delta_\Lambda)^{-1} \tag{54}$$

and

$$(I + t\Lambda^{\star -1}\Delta_\Lambda)^{-1} = \sum_{i=1}^{\infty}(-1)^i(t\Lambda^{\star -1}\Delta_\Lambda)^i \tag{55}$$

so that

$$\begin{aligned}|||(\Lambda^\star + t\Delta_\Lambda)^{-1}|||_\infty &\leq |||\Lambda^{\star -1}|||_\infty \sum_{i=1}^{\infty}|||\Lambda^{\star -1}|||_\infty^i |||\Delta_\Lambda|||_\infty|^i \\ &\leq \frac{c_{\Sigma^\star}}{1 - c_{\Sigma^\star} d\|\Delta\|_\infty} \leq \frac{3c_{\Sigma^\star}}{2}.\end{aligned} \tag{56}$$

Furthermore,

$$|||\Theta^\star + t\Delta_\Theta|||_1 \leq |||\Theta^\star|||_1 + |||\Delta_\Theta|||_1 \leq c_{\Theta^\star} + \frac{1}{2}d\|\Delta\|_\infty \leq \frac{3c_{\Theta^\star}}{2}. \tag{57}$$

Combining these expressions results in the bound

$$\|R(\Delta)\|_\infty \leq 206 c_{\Sigma^\star}^4 c_{\Theta^\star}^2 c_X^2 d^2 \|\Delta\|_\infty^2 \tag{58}$$

as required. □

Proof. (of Lemma 4) Let $(\Lambda^\star, \Theta^\star)$ be the true parameters with support S and $(\tilde{\Lambda}, \tilde{\Theta})$ be the solution to the optimization problem restricted to this support set. Our goal is to bound $\|\Delta\|_\infty$ where $\Delta = [\Delta_\Lambda \Delta_\Theta]$ with $\Delta_\Lambda = \tilde{\Lambda} - \Lambda^\star$ and $\Delta_\Theta = \tilde{\Theta} - \Theta^\star$.

Define

$$r := 2c_{H^\star}(\max\{\|\nabla_\Lambda f(\Lambda^\star, \Theta^\star)\|_\infty, \|\nabla_\Theta f(\Lambda^\star, \Theta^\star)\|_\infty\} + \lambda) \tag{59}$$

and note that by assumption we have

$$r \leq 2c_{H^\star}(\min\left\{\frac{1}{3c_{\Sigma^\star}d}, \frac{1}{412 c_{\Sigma^\star}^4 c_{\Theta^\star}^2 c_X d^2}\right\}) \tag{60}$$

To bound $\|\Delta\|_\infty$ observe that we have $\Delta_C = 0$ and

$$\Delta_S = H_{SS}^{\star -1}(R_S(\Delta) + G_S - \lambda Z_S) \tag{61}$$

as shown in the proof for Lemma 2. Our approach will be the same as that of (Ravikumar et al., 2011), using Brouwer's fixed point theorem. To do so, note that we can view the RHS of the above equation as a continuous function of Δ and thus by Brouwer's fixed point theorem on a compact set, it suffices to show that if show that if $\|\Delta_S\|_\infty \leq r$ then $\|H_{SS}^{*-1}(R_S - G_S - \lambda Z_S)\|_\infty \leq r$ as this implies that there is a solution to this equation such that $\|\Delta_S\| \leq r$ and by uniqueness (from Assumption 2) this solution must be $(\tilde{\Lambda}, \tilde{\Theta})$.

Taking infinity norm, we have

$$\|\Delta_S\|_\infty \leq \|H_{SS}^{*-1}\|_\infty \|R(\Delta)\|_\infty + \|H_{SS}^{*-1}\|_\infty \|G_S - \lambda Z_S\|_\infty \tag{62}$$

For the first term, through application of the bound on $R(\Delta)$ and by assumption on $\|\Delta\|_\infty$

$$\|H_{SS}^{*-1}\|_\infty \|R(\Delta)\|_\infty \leq 206 \kappa_{H^*} c_{\Sigma^\star}^4 c_{\Theta^\star}^2 c_X d^2 \|\Delta\|_\infty^2 \leq \frac{r}{2} \tag{63}$$

And for the second term

$$\|H_{SS}^{*-1}\|_\infty \|G_S - \lambda Z_S\|_\infty \leq \kappa_{H^*}(\|G\|_\infty + \lambda) \leq \frac{r}{2} \tag{64}$$

and thus the claim is proven. □

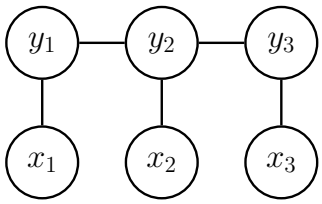

Figure 1. The chain CRF with 3 input variables and 3 output variables.

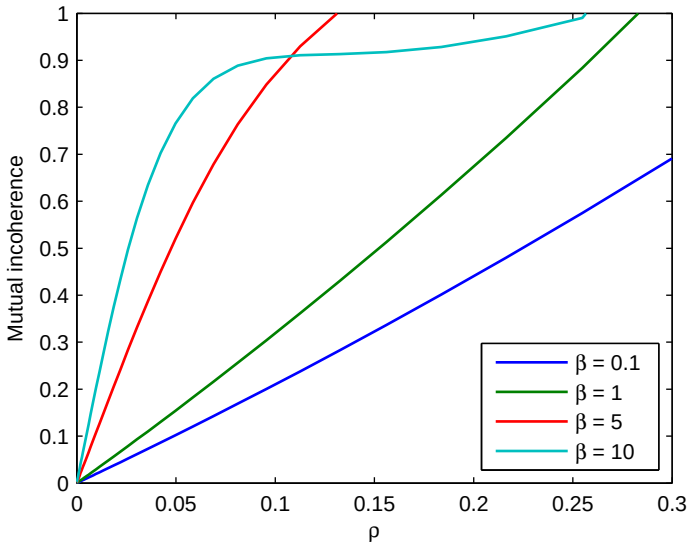

Figure 2. The mutual incoherence condition $\|\|H_{\bar{S}S}(H_{SS})^{-1}\|\|_\infty$ while varying ρ and β.

D. Mutual Incoherence for the Chain CRF

In this section we consider the mutual incoherence condition for the chain CRF, illustrated in Figure 1. For simplicity, we consider a class of models parameterized by two variables: ρ describing the conditional dependence between the output variables and β describing the relative influence of the input variables on the output variables. In particular, the class of models specified by

$$\Lambda^\star = \begin{bmatrix} 1 & \rho & 0 \\ \rho & 1 & \rho \\ 0 & \rho & 1 \end{bmatrix} \quad \Theta^\star = \begin{bmatrix} \rho\beta & 0 & 0 \\ 0 & \rho\beta & 0 \\ 0 & 0 & \rho\beta \end{bmatrix} \tag{65}$$

with positive ρ and β.

We are interested in characterizing the range over which the mutual incoherence condition

$$\|\|H_{\bar{S}S}(H_{SS})^{-1}\|\|_\infty < 1 \tag{66}$$

holds. Note that the Hessian (given in Appendix A) depends not only on these parameters, but also on the empirical covariance of the input features, S_{xx}; for the purpose of this illustration, we take S_{xx} to be the identity matrix, representing an ideal case in which the input features are perfectly uncorrelated. Under these conditions, we can see from Figure 2 that mutual incoherence indeed holds over a range of the parameters. However, as ρ increases and the output variables become more correlated, we approach the boundary at which this assumption is no longer valid.

References

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., and Ravikumar, P. Sparse inverse covariace matrix estimation using quadratic approximation. In *Neural Information Processing Systems*, 2011.

Magnus, X and Neudecker, Heinz. Matrix differential calculus. *New York*, 1988.

Ravikumar, Pradeep, Wainwright, Martin J, Raskutti, Garvesh, and Yu, Bin. High-dimensional covariance estimation by minimizing 1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Wainwright, M.J. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.