Appendix A Proof of Theorem 3.1

To simplify the notation, for a 3-tensor Δ, we often write $\|\Delta\|_q = \|\Delta\|_{q,q,q}$ for all $q \in (0,\infty]$. With this notation $\|\Delta\|_2 = \|\Delta\|_F$ and we use them interchangably. Recall S^* is the set that optimizes (2) and we refer $\sigma_s(\Theta^*)$ as σ_s when unambiguous from context.

A.1 Error bounds for M-estimators

Our main result Theorem 3.1 is an application of Theorem 1 from [18]. We start by stating their result. Let $\mathcal{R}$, with dual norm $\mathcal{R}^*$ be a decomposable regularizer over $(\mathcal{M}, \mathcal{M}^\perp)$, i.e., for all $x \in \mathcal{M}$ and $y \in \mathcal{M}^\perp$, $\mathcal{R}(x+y) = \mathcal{R}(x) + \mathcal{R}(y)$. For some loss function $\mathcal{L}$, define $\theta^* := \operatorname{argmin}_{\theta\in\Omega} \mathbb{E}\,\mathcal{L}(\theta)$, and its estimator $\widehat{\theta} := \operatorname{argmin}_{\theta\in\Omega} \mathcal{L}(\theta) + \lambda_n \mathcal{R}(\theta)$, which has an error $\Delta := \widehat{\theta} - \theta$. Let $\Psi(\mathcal{M}) := \sup_{\mathcal{M}\setminus\{0\}} \frac{\mathcal{R}(u)}{\|u\|_2}$ and let

$$\mathbb{C}(\mathcal{M};\theta^*) := \{\Delta \mid \mathcal{R}(\Delta_{\mathcal{M}^\perp}) \le 3\mathcal{R}(\Delta_{\mathcal{M}}) + 4\mathcal{R}(\theta^*_{\mathcal{M}^\perp})\}.$$

Proposition A.1 *[18, Theorem 1] If $\mathcal{L}$ is convex, differentiable, $\mathcal{R}$ is a decomposable norm over $(\mathcal{M}, \mathcal{M}^\perp)$, for some choice of subspace $\mathcal{M}$ and the following assumptions hold*

$$\lambda_n \ge 2\mathcal{R}^*(\nabla\mathcal{L}(\theta^*)) \qquad \text{and} \tag{16}$$

$$\mathcal{L}(\widehat{\theta}) - \mathcal{L}(\theta^*) - \langle \nabla\mathcal{L}(\theta^*), \Delta\rangle \ge \kappa\|\Delta\|_F^2 - \tau^2, \qquad \forall\ \Delta \in \mathbb{C}(\mathcal{M};\theta^*), \tag{17}$$

$$\|\widehat{\theta} - \theta^*\|_2^2 \le 9\frac{\lambda_n^2}{\kappa^2}\Psi^2(\mathcal{M}) + \frac{\lambda_n}{\kappa}\left(2\tau^2 + 4\mathcal{R}(\theta^*_{\mathcal{M}^\perp})\right). \qquad \square$$

Equation (17) is referred to as "$\mathcal{L}$ satisfies restricted strong convexity (RSC) with curvature κ and tolerance τ^2." Notice that for the regularized MLE problem for the MBP, $\mathcal{R}(\Theta) = \|\cdot\|_{1,1,1}$ which is a decomposable norm, whereby its dual norm is given by $\mathcal{R}^*(\nabla\mathcal{L}(\Theta)) = \|\nabla\mathcal{L}(\Theta)\|_{\infty,\infty,\infty}$. Also note that $\mathcal{M}$ above can be any subspace over which $\mathcal{R}$ is decomposable. For our purpose, we choose $\mathcal{M}_{S^*} = \{A \in \mathbb{R}^{N\times N\times p} \mid A_{S^{*c}} = 0\}$,, whereby $\mathcal{M}^\perp_{S^*} = \{A \in \mathbb{R}^{N\times N\times p} \mid A_{S^*} = 0\}$, where recall S^* is the support of the best s sparse ℓ_1 approximator of Θ^* (from (2)). For this subspace, the approximation error term $\mathcal{R}(\Theta^*_{\mathcal{M}^\perp_{S^*}}) = \|\Theta_{S^{*c}}\|_{1,1,1} = \sigma_s(\Theta^*) = \sigma_s$. Further, for the corresponding subspace $\Psi(\mathcal{M}_{S^*}) = \sqrt{s}$.

The proof of the Theorem 3.1 is the application of the result above. Lemma 4.2 states that the choice $\lambda_n = 2\sqrt{2}L_f/\epsilon\sqrt{\frac{\log(2N^2p/\delta)}{n}}$ qualifies (16). Proposition 4.1 states that if the number of samples satisfy

$$n \ge c_1 \frac{G_f(\Theta^*)}{c_f^2\, c_\ell^6}\, s^3 \log(N^2 p),$$

the RSC condition (17) holds with curvature $\kappa = \min\{1, \frac{1}{4}c_f c_\ell^2\}$ and tolerance $\tau^2 = \sigma_s^2/s$. Together these two results prove the claim. ∎

A.2 Proof of Lemma 4.2

Notice that $\|\nabla\mathcal{L}(\Theta^*)\|_{\infty,\infty,\infty} = \sup_{j,k,\ell} |\nabla_{jk\ell}\mathcal{L}(\Theta^*)|$. We know that

$$\nabla_{jk\ell}\mathcal{L}(\Theta^*) = \frac{\partial\mathcal{L}(\Theta^*)}{\partial\Theta_{jk\ell}} = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{1-x_k^t}{1-z_k^t} - \frac{x_k^t}{z_k^t}\right)\frac{\partial z_k^t}{\partial\Theta_{jk\ell}} X_{lm}^{t-1}.$$

Therefore using the law of total expectation $\mathbb{E}\nabla_{jk\ell}\mathcal{L}(\Theta^*)$ equals

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left[\frac{1-x_k^t}{1-z_k^t} - \frac{x_k^t}{z_k^t}\,\middle|\, X_{i-p}^{i-1}\right]\frac{\partial z_k^t}{\partial\Theta_{jk\ell}} X_{lm}^{t-1}\right] = 0. \tag{18}$$

Observe that this means

$$\left\{D_i := \left(\frac{1-x_k^t}{1-z_k^t} - \frac{x_k^t}{z_k^t}\right)\frac{\partial z_k^t}{\partial \Theta_{jk\ell}} X_{lm}^{t-1}, \sigma(X^i)\right\}$$

is a *martingale difference sequence* since $\mathbb{E}[D_i \mid X^i] = 0$ as derived in (18), with each $|D_i| \leq \frac{L_f}{\varepsilon}$. Using the Azuma-Hoeffding inequality [28] for martingale differences, we have that

$$\mathbb{P}\left(|\nabla_{jk\ell}\mathcal{L}(\Theta^*)| > t\right) = \mathbb{P}\left(|\sum_{i=1}^{n} D_i| > nt\right) \leq 2\exp\left(\frac{-n\varepsilon^2 t^2}{2L_f^2}\right).$$

Consequently, using the union bound we have

$$\mathbb{P}\Big(\|\nabla\mathcal{L}(\Theta^*)\|_{\infty,\infty,\infty} \geq t\Big) = \mathbb{P}\Big(\sup_{j,k,\ell} |\nabla_{jk\ell}\mathcal{L}(\Theta^*)| > t\Big) \leq \sum_{j,k,\ell} \mathbb{P}\Big(|\nabla_{jk\ell}\mathcal{L}(\Theta^*)| > t\Big) \leq 2N^2 p \cdot \exp(-n\varepsilon^2 t^2/2L_f^2)$$

from which we conclude that with probability at least $1-\delta$, we have

$$\|\nabla\mathcal{L}(\Theta^*)\|_{\infty,\infty,\infty} \leq \frac{L_f}{\varepsilon}\sqrt{\frac{2}{n}\log\frac{2N^2p}{\delta}},$$

which proves the claim. ∎

A.3 Proof of Proposition 4.1

In this section we show that with high probability the likelihood function satisfies restricted strong convexity in the star-shaped set, $\mathbb{C} := \mathbb{C}(\mathcal{M}_{S^*}; \Theta^*)$ where S^* is the optimum solution from (2), with cardinality $|S^*| = s$. Recall that for an $(s, 2)$ compressible Θ^*, its best s-sparse $\ell_{1,1,1}$ approximator is supported on S^* and we have $\|\Theta^* - \Theta^*_{S^*}\|_1 = \sigma_(\Theta^*)$. Here $\mathcal{M}_{S^*}$ denotes the subspace of 3-tensors Θ supported on S^*.

Here and elsewhere, we use $\mathbb{X}$ as a shorthand for the collection of X^{t-1}, $t = 1, \ldots, n$ or equivalently for

$$\mathbb{X} := (x^t)_{-p+1}^{n-1} = (x^{n-1}, x^{n-2}, \ldots, x^{-p+1}) \in \mathcal{S}^{n+p-1}, \tag{19}$$

where $\mathcal{S} := \{0,1\}^N$.

Lemma A.2 *The remainder of the first-order Taylor expansion of the loss, around Θ^*, satisfies*

$$R\mathcal{L}(\Delta; \Theta^*) \;\geq\; \mathcal{E}(\Delta; \mathbb{X}) := \frac{c_f}{n}\sum_{t=1}^{n}\sum_{k=1}^{N}\langle \Delta_{k**}, X^{t-1}\rangle^2, \tag{20}$$

for all $\Theta \in \Omega$ and $\Delta \in \mathbb{R}^{N\times N\times p}$.

By Lemma A.2, in order to establish RSC, it is enough to show $\mathcal{E}(\Delta; \mathbb{X}) \geq \kappa^2\|\Delta\|_F^2 - \tau^2$ for all $\Delta \in \mathbb{C}(S^*; \Theta^*)$.

Before we move on, let us record an alternative form $\mathcal{E}(\Delta; \mathbb{X})$, which will be useful in subsequent analysis. Let $\mathsf{X} \in \{0,1\}^{n\times Np}$ be the design matrix with i^{th} row,

$$\mathsf{X}_{i*} := [(x^{i-1})^\top \; (x^{i-2})^\top \ldots (x^{i-p})^\top] \in \{0,1\}^{Np} \tag{21}$$

and define a *stacking operator* $\mathtt{S} : \mathbb{R}^{N\times N\times p} \to \mathbb{R}^{N\times Np}$ that reshapes a tensor as follows:

$$\mathtt{S}(\Delta) := [\Delta_{**1} \; \Delta_{**2} \; \ldots \Delta_{**p}] \in \mathbb{R}^{N\times Np}. \tag{22}$$

We have the following representation:

Lemma A.3 *For any $\Delta \in \mathbb{R}^{N\times N\times p}$, we have*

$$\mathcal{E}(\Delta; \mathbb{X}) = \frac{c_f}{n}\|\mathsf{X}\,\mathtt{S}(\Delta)^\top\|_F^2 = \frac{c_f}{n}\sum_{t=1}^{n}\sum_{j=1}^{N}\left\langle \Delta_{j**}, X^{t-1}\right\rangle^2. \tag{23}$$

We next show that at the population level $\mathbb{E}\mathcal{E}(\Delta;\mathbb{X})$ satisfies the RSC:

Lemma A.4 (Strong convexity at the population level) *Under Assumption (A1),*

$$\mathbb{E}\mathcal{E}(\Delta;\mathbb{X}) \geq c_f\, c_\ell^2\, \|\Delta\|_F^2, \qquad \text{for all } \Delta \in \mathbb{R}^{N\times N\times p}. \tag{24}$$

We then show that $\mathcal{E}(\Delta;\mathbb{X})$ concentrates around its mean for any fixed Δ:

Lemma A.5 (Concentration for fixed Δ) *For any $\Delta \in \mathbb{R}^{N\times N\times p}$,*

$$\mathbb{P}\Big(\big|\mathcal{E}(\Delta;\mathbb{X}) - \mathbb{E}\mathcal{E}(\Delta;\mathbb{X})\big| > t\|\Delta\|_{2,1,1}^2\Big) \;\leq\; 2\exp\left(\frac{-nt^2}{G_f(\Theta^*)}\right).$$

Combining the two Lemmas A.4 and A.5 and using a discretization argument, we can establish a uniform lower bound on the random function $\Delta \mapsto \mathcal{E}(\Delta;\mathbb{X})$:

Lemma A.6 (Uniform lower bound) *Proposition 4.1 holds with $R\mathcal{L}(\Delta;\Theta^*)$ replaced with $\mathcal{E}(\Delta;\mathbb{X})$.*

Putting the pieces together establishes the claimed RSC for the loss $\mathcal{L}$. ∎

A.4 Proof of Lemma A.4

Recall the notation in (21) and (22) and (27), and the statement of Lemma A.3 which gives us that

$$\mathbb{E}\mathcal{E}(\Delta;\mathbb{X}) = c_f \sum_{j=1}^{N} \mathbb{E}\left\langle \Delta_{j\cdot\cdot}, X^i\right\rangle^2,$$

for any i. Let $D := \mathsf{S}(\Delta) \in \mathbb{R}^{N\times Np}$ and the design matrix $\mathsf{X} \in \mathbb{R}^{n\times Np}$, whereby its i^{th} row is $\mathsf{X}_{i*} \in \mathbb{R}^{1\times n}$. Notice the summation above does not include i or a $\frac{1}{n}$ factor due to the expectation operator and the stationarity assumption. Hence $\mathbb{E}\mathcal{E}(\Delta;\mathbb{X}) = c_f\mathbb{E}\|D\mathsf{X}_{i*}^\top\|_2^2$ which equals

$$\mathbb{E}[\mathsf{X}_{i*}D^\top D\mathsf{X}_{i*}^\top] = \mathbb{E}(\mathrm{tr}(\mathsf{X}_{i*}D^\top D\mathsf{X}_{i*}^\top) = \mathrm{tr}(D^\top D\mathbb{E}\mathsf{X}_{i*}^\top\mathsf{X}_{i*}) = \left\langle D^\top D, \mathbf{R}\right\rangle,$$

where $\mathbf{R} \in \mathbb{R}^{Np\times Np}$ denotes the population autocorrelation matrix $\mathbb{E}\mathsf{X}_{i*}^\top\mathsf{X}_{i*}$, and hence is independent of i. Since $\mathbf{R} - \lambda_{\min}(\mathbf{R})I \succeq 0$, and $D^\top D \succeq 0$, we have that

$$\mathbb{E}\mathcal{E}(\Delta;\mathbb{X}) = c_f\left\langle D^\top D, \mathbf{R}\right\rangle \geq c_f\lambda_{\min}(\mathbf{R})\left\langle D, D\right\rangle = c_f\lambda_{\min}(\mathbf{R})\|D\|_F^2. \tag{25}$$

Now let $\mathcal{X}(\omega) \in \mathbb{C}^{N\times N}$ be the power spectrum of the process for which $m(\mathcal{X}) := \min_{\omega\in[-\pi,\pi)} \lambda_{\min}(\mathcal{X}(\omega))$, then $\lambda_{\min}(\mathbf{R}) \geq 2\pi m(\mathcal{X}(\omega))$ (see [2, Proposition 2.3]). Observe that $\mathbf{R}$ is block symmetric matrix with $\mathbf{R}_{ij} := \mathrm{Cov}(x^{t-i}, x^{t-j}) \in \mathbb{R}^{N\times N}$. Let $\mathbf{u}^\top = \begin{bmatrix} u_0^\top & u_2^\top & \dots & u_{p-1}^\top\end{bmatrix}$, where $u_i \in \mathbb{R}^N$ and let $G(\omega) = \sum_{r=0}^{p-1} u_r e^{-jr\omega}$. Consider

$$\mathbf{u}^\top\mathbf{R}\mathbf{u} = \sum_{r,s=0}^{p-1} u_r^\top \mathrm{Cov}(x^{t-r}, x^{t-s})u_s = \sum_{r,s=0}^{p-1}\int_{-\pi}^{\pi} u_r^\top\mathcal{X}(\omega)e^{j(r-s)\omega}u_s d\omega = \int_{-\pi}^{\pi} G^{\mathrm{H}}(\omega)\mathcal{X}(\omega)G(\omega)d\omega.$$

Since $\mathcal{X}(\omega)$ is a hermitian matrix, $G^{\mathrm{H}}(\omega)\mathcal{X}(\omega)G(\omega)$ is always a Real matrix. Moreover, we have that $G^{\mathrm{H}}(\omega)\mathcal{X}(\omega)G(\omega) \geq m(\mathcal{X}(\omega))G^{\mathrm{H}}(\omega)G(\omega)$, where by

$$\mathbf{u}^\top\mathbf{R}\mathbf{u} \geq m(\mathcal{X})\int_{-\pi}^{\pi} G^{\mathrm{H}}(\omega)G(\omega)d\omega = 2\pi m(\mathcal{X})\|\mathbf{u}\|_2^2,$$

by Parseval's theorem. Consequently we have that $\lambda_{\min}(\mathbf{R}) \geq 2\pi m(\mathcal{X}) = c_\ell^2$. This together with equation (25) proves Lemma A.4. ∎

A.5 Proof of Lemma A.5

We start by stating a result by Kontorovich et al. [11] for a process consisting of dependent random variables taking values in a countable space:

Proposition A.7 *[11, Theorem 1.1] Consider a $\mathcal{S}$-valued process $\{X_m\}_{m\in[n]}$ for some countable set $\mathcal{S}$. Let $\phi : \mathcal{S}^n \to \mathbb{R}$ be an L Lipschitz function of $\mathbb{X} := \{X_m\}_{m=1}^n$ with respect to the Hamming norm. Define the mixing coefficient*

$$\eta_{k\ell} \triangleq \sup_{\substack{w,w' \in \mathcal{S} \\ y \in \mathcal{S}^{k-1}}} \|\mathbb{P}\left(X_\ell^n = \cdot \mid X_k = w, X_1^{k-1} = y\right) - \mathbb{P}\left(X_\ell^n = \cdot \mid X_k = w', X_1^{k-1} = y\right)\|_{\mathrm{TV}}.$$

and let $\eta \in \mathbb{R}^{n\times n}$ be the matrix with entries $\eta_{k\ell}$ for $\ell \geq k$ and zero otherwise (i.e., an upper triangular matrix). Then, ϕ concentrates around its mean as follows:

$$\mathbb{P}\{|\phi(\mathbb{X}) - \mathbb{E}\phi(\mathbb{X})| > t\} \leq 2\exp\left(\frac{-t^2}{nL^2\|\eta\|_\infty^2}\right), \quad \forall t > 0 \tag{26}$$

where $\|\eta\|_\infty = \max_k \sum_{\ell\geq k} \eta_{k\ell}$ is the ℓ_∞ operator norm of matrix η. Further if $\mathbb{X}$ is Markov with a Dobrushin ergodicity coefficient $\tau < 1$, then $\eta_{k\ell} \leq \tau^{\ell-k}$ and $\|\eta\|_\infty \leq (1-\tau)^{-1}$. □

We apply the above result by finding the Lipschitz constant of L of the map $\mathbb{X} \to \mathcal{E}(\Delta, \mathbb{X})$, and bounding the mixing coefficients in terms of the process parameters Θ, p and the link function f. Lemma B.1 shows that

$$L^2 \leq 4c_f^2 \frac{\|\Delta\|_{2,1,1}^4}{n^2},$$

and Lemmas C.1 and C.2 together show that

$$\max_k \left(\sum_{\ell\geq k} \eta_{k\ell}\right)^2 \leq \frac{G_f(\Theta^*)}{4c_f^2}.$$

Finally, we replace t with $t\|\Delta\|_{2,1,1}^2$ in (26) which proves the claim. ∎

A.6 Proof of Lemma A.6

Recall the notation $\|\Delta\|_q := \|\Delta\|_{q,q,q}$. for the ℓ_q norm of a 3-tensor $\Delta \in \mathbb{R}^{[N]^2\times[p]}$. Also note that $\|\Delta\|_2 = \|\Delta\|_F$. We also write $\mathbb{B}_q(r)$ for the tensor ℓ_q ball of radius r, and $\partial\mathbb{B}_q(r)$ for the boundary of that ball. For example,

$$\mathbb{B}_1(r) := \{\Delta \in \mathbb{R}^{N\times N\times p} : \|\Delta\|_1 \leq r\}, \quad \partial\mathbb{B}_2(r) := \{\Delta \in \mathbb{R}^{N\times N\times p} : \|\Delta\|_2 = r\},$$

Let $G_f = G_f(\Theta^*)$ be the quantity defined in (7). Recall that S^* is the support of the best ℓ_1 approximator of Θ^* that has cardinality s, i.e., the optimal solution to (2). Let

$$\mathbb{C}^* := \mathbb{C}(S^*, \Theta^*) = \left\{\Delta \in \Omega^* : \ \|\Delta_{S^{*c}}\|_1 \leq 3\|\Delta_{S^*}\|_1 + 4\|\Theta^*_{S^{*c}}\|_1\right\}.$$

Step 1: Fixed ℓ_2 norm. We first establish the RSC (with no tolerance) for tensors in $\mathbb{C}^*$ of a given Frobenius norm, say $\|\Delta\|_2 = r_1$.

Note that for any $\Delta \in \mathbb{C}^*$, we have $\Delta = \Delta_{S^*} + \Delta_{S^{*c}}$, hence

$$\|\Delta\|_1 = \ \|\Delta_{S^*}\|_1 + \|\Delta_{S^{*c}}\|_1 \ \leq \ 4\|\Delta_{S^*}\|_1 + 4\|\Theta_{S^{*c}}\|_1 \ \leq \ 4\big(\sqrt{s}\|\Delta\|_F + \sigma_s(\Theta^*)\big)$$

using $\|\Delta_{S^*}\|_1 \leq \sqrt{s}\|\Delta_{S^*}\|_F$ and $\|\Theta_{S^{*c}}\|_1 \leq \sigma_s(\Theta^*)$. It follows that for any $r_1 > 0$,

$$\mathbb{C}^* \cap \partial\mathbb{B}_F(r_1) \ \subseteq \ \mathbb{B}_1\big(r_2\big),$$

where $r_2 := 4\big(r_1\sqrt{s} + \sigma_s(\Theta^*)\big)$. Next we consider covering $\mathbb{C}^* \cap \partial\mathbb{B}_F(r_1)$ by finding an ε-cover of $\mathbb{B}_1(r_2)$.

For a metric space (T, ρ), let $\mathcal{N}(\varepsilon, T, \rho)$ be the ε-covering number of T in ρ. The quantity $\log\mathcal{N}(\varepsilon, T, \rho)$ is called the metric entropy. The following is an adaptation of a result of [22, Lemma 3, case $q = 1$]:

Lemma A.8 *Let $\mathsf{X} \in \mathbb{R}^{n\times d}$ be a matrix with column normalization $\|\mathsf{X}_{*j}\|_2 \le \sqrt{n}$ for all j. Consider the set of matrices $\mathbb{R}^{N\times d}$ and let $\mathbb{B}_1^{N\times d}(u)$ be the (elementwise) ℓ_1 ball of radius u in that space, i.e.*

$$\mathbb{B}_1^{N\times d}(u) := \{D \in \mathbb{R}^{N\times d} : \|D\|_1 \le u\}.$$

Consider the (pseudo) metric $\rho(D_1, D_2) := \frac{1}{\sqrt{n}}\|\mathsf{X}(D_1 - D_2)^\top\|_2$ on $\mathbb{R}^{N\times d}$. Then, for a sufficiently small constant $C_1 > 0$, the metric entropy of $\mathbb{B}_1(u)$ in ρ is bounded as

$$\log \mathcal{N}\big(\varepsilon, \mathbb{B}_1^{N\times d}(u), \rho\big) \le C_2 \frac{u^2}{\varepsilon^2} \log(Nd),, \quad \forall \varepsilon \le C_1 u.$$

Now, recall the design matrix $\mathsf{X} \in \mathbb{R}^{n\times Np}$ defined in (21). Note that X satisfies the column normalization property $\|\mathsf{X}_{*j}\|_2 \le \sqrt{n}$ for all j since $\mathsf{X}_{*j} \in \{0,1\}^n$. Fix $\varepsilon \in (0, 2C_1 r_2/r_1)$ for sufficiently small $C_1 > 0$. It follows that there exists an $(r_1\varepsilon/2)$-cover $\mathcal{N}''$ of $\mathbb{B}_1^{N\times Np}(r_2)$ in the metric defined in Lemma A.8 with cardinality bounded as

$$\log|\mathcal{N}''| \;\lesssim\; \frac{r_2^2}{r_1^2\varepsilon^2}\log(N^2 p).$$

Recalling the stacking map $\Delta \mapsto \mathtt{S}(\Delta)$ from (22) (mapping tensors to matrices), we have

$$\mathtt{S}(\mathbb{C}^* \cap \partial\,\mathbb{B}_F(r_1)) \;\subseteq\; \mathtt{S}(\mathbb{B}_1(r_2)) = \mathbb{B}_1^{N\times Np}(r_2).$$

Define a (pseudo) metric on the tensor space by $\overline{\rho}(\Delta, \Delta') = \rho(\mathtt{S}(\Delta), \mathtt{S}(\Delta'))$. Since $\mathtt{S}$ is a bijection, it follows that there is an exterior $(r_1\varepsilon/2)$-covering of $\mathbb{C}^* \cap \partial\,\mathbb{B}_F(r_1)$ in metric $\overline{\rho}$ with the same cardinality as $\mathcal{N}''$; call it $\mathcal{N}'$. (Here, the exterior covering means that the elements need not belong the set they cover. Elements of $\mathcal{N}'$ are tensors in $\mathbb{B}_1(r_2)$ but not necessarily in $\mathbb{C}^* \cap \partial\,\mathbb{B}_F(r_1)$.)

We can pass from $\mathcal{N}'$ to an $(r_1\varepsilon)$-cover, say $\mathcal{N}$, of $\mathbb{C}^* \cap \partial\,\mathbb{B}_F(r_1)$ with cardinality no more than $\mathcal{N}'$, i.e. $|\mathcal{N}| \le |\mathcal{N}'|$, (see Exercise 4.2.9 in [29, p.75]). In particular, we have $\mathcal{N} \subseteq \mathbb{C}^* \cap \mathbb{B}_F(r_1)$.

Recall from Lemma A.3 that $\sqrt{\mathcal{E}(\Delta;\mathbb{X})} = c_f^{1/2}\|\mathsf{X}\,\mathtt{S}(\Delta)^\top\|_F/\sqrt{n}$. Then, by triangle inequality

$$|\sqrt{\mathcal{E}(\Delta;\mathbb{X})} - \sqrt{\mathcal{E}(\Delta';\mathbb{X})}| \le c_f^{1/2}\,\overline{\rho}(\Delta, \Delta'),$$

for any two tensors Δ and Δ'. Using $(a-b)^2 \ge \frac{1}{2}a^2 - b^2$, we have

$$\mathcal{E}(\Delta;\mathbb{X}) \ge \frac{1}{2}\mathcal{E}(\Delta';\mathbb{X}) - c_f\,\overline{\rho}^2(\Delta, \Delta').$$

If follows that

$$\inf_{\Delta\,\in\,\mathbb{C}^*\cap\partial\,\mathbb{B}_F(r_1)} \mathcal{E}(\Delta;\mathbb{X}) \;\ge\; \frac{1}{2}\inf_{\Delta\,\in\,\mathcal{N}} \mathcal{E}(\Delta;\mathbb{X}) - c_f(r_1\varepsilon)^2$$

By Lemma A.5 and the union bound, with probability at least $1 - 2|\mathcal{N}|\exp(-nt^2/G_f)$, we have

$$|\mathcal{E}(\Delta;\mathbb{X}) - \mathbb{E}\mathcal{E}(\Delta;\mathbb{X})| \le t\|\Delta\|_{2,1,1}^2, \quad \forall \Delta \in \mathcal{N}.$$

Since $\mathcal{N} \subseteq \mathbb{C}^* \cap \mathbb{B}_F(r_1)$, for any $\Delta \in \mathcal{N}$ we have $\|\Delta\|_{2,1,1}^2 \le s\|\Delta\|_2^2$ and $\|\Delta\|_2 = r_1$. It follows that with the same probability,

$$\mathcal{E}(\Delta;\mathbb{X}) \;\ge\; \mathbb{E}\mathcal{E}(\Delta;\mathbb{X}) - t\,sr_1^2 \;\ge\; (c_f c_\ell^2 - ts)\,r_1^2, \quad \forall \Delta \in \mathcal{N}$$

where we have used Lemma A.4 in the second inequality. It follows that we the same probability

$$\inf_{\Delta\,\in\,\mathbb{C}^*\cap\partial\,\mathbb{B}_F(r_1)} \mathcal{E}(\Delta;\mathbb{X}) \;\ge\; \Big(\frac{1}{2}c_f c_\ell^2 - \frac{1}{2}ts - c_f\varepsilon^2\Big)\,r_1^2.$$

To simplify, let $\sigma_s = \sigma_s(\Theta^*)$. Taking $r_1 = (\sigma_s/\sqrt{s}) + 1\{\sigma_s = 0\}$, we can balance the two terms in r_2. We obtain

$$4\sqrt{s} \;\le r_2/r_1 \;\le 8\sqrt{s}.$$

The constraint on ε is $\varepsilon \le 2C_1(r_2/r_1)$. It is enough to require $\varepsilon \le 8C_1\sqrt{s}$. Taking $\varepsilon^2 = \frac{1}{8}c_\ell^2$ and assuming that $s \ge \frac{c_\ell^2}{512C_1^2}$ satisfies the constraint. Also, taking $t = \frac{1}{4}c_f c_\ell^2/s$, we obtain

$$\mathbb{P}\Big(\inf_{\Delta \in \mathbb{C}^* \cap \partial \mathbb{B}_F(r_1)} \mathcal{E}(\Delta;\mathbb{X}) \ \ge \ (\frac{1}{4}c_f c_\ell^2)\, r_1^2 \Big) \ge 1 - 2\exp\Big(\log|\mathcal{N}| - c_f^2 c_\ell^4 \frac{n}{16 s^2 G_f} \Big) =: P_1$$

Noting that

$$\log|\mathcal{N}| \le C_3 \big(8\sqrt{s}\big)^2 \Big(\frac{8}{c_\ell^2}\Big) \log(N^2 p),$$

the probability is further bounded as

$$1 - P_1 \ \le \ 2\exp\Big(\frac{512}{c_\ell^2} C_3\, s \log(N^2 p) - c_f^2 c_\ell^4 \frac{n}{16 s^2 G_f} \Big).$$

Assuming $c_f^2 c_\ell^4 \frac{n}{16 s^2 G_f} \ge \ 1024\, c_\ell^{-2} C_3\, s \log(N^2 p)$, we have

$$1 - P_1 \ \le \ 2\exp\Big(-512\, C_3 c_\ell^{-2}\, s \log(N^2 p) \Big) \le 2(N^2 p)^{-c_1 s}.$$

where $c_1 = O(c_\ell^{-2})$ only depends on c_ℓ. Thus, we have established RSC with high probability for tensors in $\mathbb{C}^* \cap \partial \mathbb{B}_F(r_1)$ with curvature $\kappa = \frac{1}{4}c_f c_\ell^2$ and tolerance $\tau^2 = 0$.

Note that when $\sigma_s = 0$ (i.e., the case of hard sparsity), $\mathbb{C}^*$ is a cone hence the above extends immediately to all $\Delta \in \mathbb{C}^*$, since $\mathcal{E}(t\Delta;\mathbb{X}) = t^2 \mathcal{E}(\Delta;\mathbb{X})$ for all $t > 0$, thus completing the proof.

Now let us assume $\sigma_s > 0$ in the rest of the proof.

Step 2: Extending to the complement of the ℓ_2 norm ball. For $\sigma_s > 0$, since $\mathbb{C}^*$ is not a cone, we cannot use a scale-invariance argument to extend to general tensors. However, we have the following:

Lemma A.9 *Assume that RSC holds for $\mathcal{E}$ in the sense of $\mathcal{E}(\Delta;\mathbb{X}) \ge \kappa \|\Delta\|_F^2$, for all $\Delta \in \mathbb{C}^* \cap \partial\, \mathbb{B}_F(r)$. Then, RSC holds in the same sense for all $\Delta \in \mathbb{C}^* \cap \{\Delta : \ \|\Delta\|_F \ge r\}$.*

The lemma establishes the RSC of the previous step for all of $\mathbb{C}^* \cap \{\Delta : \ \|\Delta\|_F \ge r_1\}$. The proof is straightforward and follows from the observation that $\mathcal{E}(\cdot;\mathbb{X})$ satisfies $\mathcal{E}(t\Delta;\mathbb{X}) = t^2\mathcal{E}(\Delta;\mathbb{X})$, for $t \ge 1$.

Step 3: Extending to small radii. It remains to extend the result to $\Delta \in \mathbb{C}^* \cap \{\Delta : \ \|\Delta\|_F < r_1\}$. In this case, we simply take $\tau^2 := r_1^2 = \sigma_s^2/s$ (since $\sigma_s > 0$ by assumption) so that

$$\mathcal{E}(\Delta;\mathbb{X}) \ge 0 \ge \|\Delta\|_F^2 - \tau^2$$

so that the RSC holds with curvature $= 1$ and tolerance τ^2. Putting the pieces together, we have the RSC for all $\Delta \in \mathbb{C}$ with the probability given in Step 1, curvature $\kappa = \min\{\frac{1}{4}c_f c_\ell^2, 1\}$ and tolerance $\tau^2 = \sigma_s^2/s$. The proof is complete. ∎

Appendix B Auxiliary results

B.1 Proof of Lemma A.2

Recall that the loss can be written as

$$\mathcal{L}(\Theta) = -\frac{1}{n}\sum_{i=1}^{N}\sum_{t=1}^{n} \ell_{it}(\langle \Theta_{i**}, X^{t-1}\rangle)'$$

where $\ell_{it}(u) = x_i^t \log f(u) + (1 - x_i^t)\log(1 - f(u))$ is the likelihood for $x_i^t \sim \mathrm{Ber}(f(u))$. We have

$$\frac{\partial^2 \mathcal{L}}{\partial \Theta_{abc} \partial \Theta_{k\ell m}}(\Theta) = -\frac{1}{n}\sum_t \ell''_{at}(\langle \Theta_{a**}, X^{t-1}\rangle)\, X_{bc}^{t-1} X_{\ell m}^{t-1}\, 1\{k = a\},$$

where $-\ell''_{at}(u)$ equals

$$x_a^t \frac{f'^2 - ff''}{f^2}(u) + (1 - x_a^t)\frac{(1-f)f'' + f'^2}{(1-f)^2}(u) \geq c_f > 0,$$

where $c_f := \min\{\frac{f'^2 - ff''}{f^2}, \frac{f'^2 + (1-f)f''}{(1-f)^2}\}$ is assumed to be positive by assumption (A2). It follows that

$$\frac{\partial^2 \mathcal{L}}{\partial \Theta_{abc} \partial \Theta_{k\ell m}}(\Theta) \geq \frac{c_f}{n} \sum_{t=1}^{n} X_{bc}^{t-1} X_{\ell m}^{t-1} 1\{k = a\}, \quad \forall \Theta \in \Omega.$$

Thus, the Hessian quadratic form is uniformly controlled from below as:

$$\begin{aligned}\langle \Delta \nabla^2 \mathcal{L}(\Theta), \Delta \rangle &\geq \sum_{a,b,c,k,\ell,m} \Delta_{abc} \Big[\frac{c_f}{n} \sum_{t=1}^{n} X_{bc}^{t-1} X_{\ell m}^{t-1} 1\{k = a\}\Big] \Delta_{k\ell m}, = \sum_{k,b,c,\ell,m} \Delta_{kbc} \Big[\frac{c_f}{n} \sum_{t=1}^{n} X_{bc}^{t-1} X_{\ell m}^{t-1}\Big] \Delta_{k\ell m} \\ &= \frac{c_f}{n} \sum_{k=1}^{N} \sum_{t=1}^{n} \Big(\sum_{b,c} \Delta_{kbc} X_{bc}^{t-1}\Big)\Big(\sum_{\ell,m} \Delta_{k\ell m} X_{\ell m}^{t-1}\Big) = \frac{c_f}{n} \sum_{t=1}^{n} \sum_{k=1}^{N} \langle \Delta_{k**}, X^{t-1} \rangle^2,\end{aligned}$$

for all $\Theta \in \Omega$ and $\Delta \in \mathbb{R}^{N \times N \times p}$. The proof is complete.

B.2 Quadratic lower bound is Lipschitz

Lemma B.1 *The map $\mathbb{X} \mapsto \mathcal{E}(\Delta; \mathbb{X})$ is $(2c_f \|\Delta\|_{2,1,1}^2/n)$-Lipschitz w.r.t. Hamming distance on $\mathcal{S}^{n+p-1}$.* □

Proof We recall the following alternative expressions,

$$\langle \Delta_{k**}, X^{t-1} \rangle = \sum_{\alpha=1}^{p} \langle \Delta_{k*\alpha}, X_{*\alpha}^{t-1} \rangle =: \sum_{\alpha=1}^{p} \langle \Delta_{k*\alpha}, x^{t-\alpha} \rangle. \tag{27}$$

It is enough to consider two sequences $\{x^t\}$ and $\{y^t\}$ which differ in one coordinate, say $\mathbb{X} = (x^{-p+1}, x^{-p}, \ldots, x^{n-1})$ and $\mathbb{Y} = (x^{-p+1}, x^{-p}, \ldots, y^r, \ldots, x^{n-1})$, where r will be fixed. The general case follows, via triangle inequality, since any $\mathbb{Y}$ can be reached from $\mathbb{X}$ by a sequence $\mathbb{X} =: \mathbb{X}_{(0)}, \mathbb{X}_{(1)}, \ldots, \mathbb{X}_{(h)} := \mathbb{Y}$ where h is the hamming distance of $\mathbb{X}$ and $\mathbb{Y}$ in $\mathcal{S}^{n+p-1}$, such that $\mathbb{X}_{(i)}$ and $\mathbb{X}_{(i-1)}$ are Hamming distance 1 apart, for $i = 1, 2, \ldots h$.

Let X^{t-1} and Y^{t-1} be defined based on $\mathbb{X}$ and $\mathbb{Y}$ as before, i.e., the corresponding p-lag history at time $t-1$. Note that X^{t-1} and Y^{t-1} are different only for t such that $t \in \{r+1, \ldots, r+p\}$, and for such r, we have

$$|\langle \Delta_{k**}, X^{t-1} - Y^{t-1} \rangle| \leq \|\Delta_{k,*,t-r}\|_1.$$

We also have

$$|\langle \Delta_{k**}, X^{t-1} + Y^{t-1} \rangle| \leq 2\|\Delta_{k**}\|_1.$$

Combining we obtain

$$\begin{aligned}|\mathcal{E}(\Delta; \mathbb{X}) - \mathcal{E}(\Delta; \mathbb{Y})| &= \frac{c_f}{n} \Big| \sum_{t=r+1}^{r+p} \sum_{k=1}^{N} \big[\langle \Delta_{k**}, X^{t-1} \rangle^2 - \langle \Delta_{k**}, Y^{t-1} \rangle^2\big] \Big| \\ &\leq \frac{2c_f}{n} \sum_{t=r+1}^{r+p} \sum_{k=1}^{N} \|\Delta_{k,*,t-r}\|_1 \, \|\Delta_{k**}\|_1 \;\leq\; \frac{2c_f}{n} \sum_{k=1}^{N} \|\Delta_{k**}\|_1^2 = \frac{2c_f}{n} \|\Delta\|_{2,1,1}^2\end{aligned}$$

where we have used $\sum_{t=r+1}^{r+p} \|\Delta_{k,*,t-r}\|_1 = \|\Delta_{k**}\|_1$. This proves the claim. ∎

Lemma B.2 *Assume that $U \sim \mathrm{Ber}(p)$, and $V \sim \mathrm{Ber}(q)$ for $p, q \in [\varepsilon, 1-\varepsilon]$ for some $\varepsilon \in (0, \frac{1}{2})$. Then,*

$$D_{\mathrm{KL}}(U \| V) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \;\leq\; \frac{3}{4\varepsilon(1-\varepsilon)}(p-q)^2.$$

If U_i are independent Bernoulli with means p_i and V_i are independent Bernoulli with means q_i, then the vectors $\mathbf{U} = (U_1, U_2, \ldots, U_m)$ and $\mathbf{V} = (V_1, V_2, \ldots, B_m)$ satisfy

$$D_{\mathrm{KL}}(\mathbf{U}\|\mathbf{V}) = \sum_{i=1}^{m} D_{\mathrm{KL}}(U_i\|V_i) \le \frac{3}{4\varepsilon(1-\varepsilon)}\|\mathbf{p}-\mathbf{q}\|_2^2.$$

□

Proof It is enough to prove for the case $q \ge p$ (the other case follows by applying the proven case to $1-p$ and $1-q$). The second claim follows from the decomposition of the KL divergence for product distributions. Let $\delta := \varepsilon(1-\varepsilon)$. Fix p and consider the function

$$f(q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} - \frac{1}{4\delta}(p-q)^2,$$

over $q \in [p, 1-\varepsilon]$. We have

$$f'(q) = (q-p)\Big(\frac{1}{q(1-q)} - \frac{1}{2\delta}\Big).$$

We have $f(q) = f(p) + f'(\tilde{q})(q-p)$ for some $\tilde{q} \in [p,q]$. Note that $f(p) = 0$ and

$$f'(\tilde{q}) \le (\tilde{q}-p)\Big(\frac{1}{\delta} - \frac{1}{2\delta}\Big) \le \frac{1}{2\delta}(q-p)$$

using the fact that $(\tilde{q}(1-\tilde{q}))^{-1} \in [4, \delta^{-1}]$. Thus, we have $f(q) \le (q-p)^2/(2\delta)$. ∎

Appendix C Background on Markov contraction

We briefly state some properties of Markov kernels. This has been studied extensively in the literature on Markov chains and their contraction. Here we discuss properties of homogeneous Markov chains only for the sake of brevity, and since the MBP is time invariant. For a Markov chain over a discrete space $\mathcal{S}$, let $(P_{ij}) \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ be its kernel. The Kernel is a non-negative stochastic matrix, with each row as a probability distribution. Let $\mathcal{H}_1$ be the subspace $\{u \in \mathbb{R}^{|\mathcal{S}|} \mid \mathbf{1}^\top u = 0\}$. This subspace is invariant to every Markov kernel $P \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$, i.e., for all $u \in \mathcal{H}_1$, we have $u^\top P \in \mathcal{H}_1$. Consider the quantity

$$\tau_1(P) := \sup_{u\in\mathcal{H}_1} \frac{\|u^\top P\|_1}{\|u\|_1},$$

also known as the *Dobrushin ergodicity coefficient*. Hence due to the invariance of $\mathcal{H}_1$ to P, we can conclude that

$$\|u^\top P^\ell\|_1 \le \tau_1(P)^\ell \|u\|_1 \quad \forall\ u \in \mathcal{H}_1. \tag{28}$$

For all stochastic matrices $\tau_1(P) \le 1$. The inequality is strict if and only if no two rows of P are orthogonal (such Markov kernels are said to be scrambling). Some sufficient conditions for $\tau_1(P) < 1$, which are easier to interpret, are (i) P is a positive matrix, and (ii) P has a column with all entries positive.

Any process for which

$$\mathbb{P}(x^t \mid x^{t-1}, x^{t-2}, \ldots) = \mathbb{P}(x^t \mid x^{t-1}, x^{t-2}, \ldots x^p), \tag{29}$$

for some fininte p, can equivalently be represented with a Markov kernel $\mathcal{K}$ in $\mathbb{R}^{|\mathcal{S}|^p\times|\mathcal{S}|^p}$, such that

$$\mathcal{K}_{ij} = \mathbb{P}((x^t, x^{t-1}, \ldots, x^{t-p+1}) = j \mid (x^{t-1}, x^{t-2}, \ldots, x^{t-p}) = i), \qquad \forall\ i,j \in \mathcal{S}^p.$$

However, this Kernel matrix $\mathcal{K}$ is constrained since $\mathcal{K}_{ij} \neq 0$ if and only if $(i_1, i_2, \ldots, i_{p-1}) = (j_2, j_3, \ldots, j_p)$. One can then show that $\tau_1(\mathcal{K}^k) = 1$ for all $k < p$. Fortunately, under the mild assumption that $\mathbb{P}(x_t \mid x^{t-1}, x^{t-2}, \ldots, x^{t-p}) > 0$, one can show that $\tau_1(\mathcal{K}^p) < 1$. We make use of this quantity to upper bound the summation $\sum_{\ell\ge k}\eta_{k\ell}$.

One can think of $\mathcal{K}^p$ as a $|\mathcal{S}|^p \times |\mathcal{S}|^p$ Markov kernel that gives the transition probabilities for consecutive blocks of size p, i.e., for $i,j \in \mathcal{S}^p$, we have that for any t,

$$(\mathcal{K}^p)_{i,j} = \mathbb{P}(X_{t+1}^{t+p} = j \mid X_{t-p+1}^t = i).$$

Lemma C.1 *Let $\mathbb{X}$ be generated according to (1), where $f : \mathbb{R} \to [\varepsilon, 1-\varepsilon]$ is L_f-Lipschitz for some $\varepsilon \in (0, \frac{1}{2})$. Define*

$$g^2(\Theta^*) = \frac{3L_f^2}{2\varepsilon} \sum_{\ell=1}^{p} \sum_{j=1}^{N} \left(\sum_{k=1}^{N} \sum_{i=\ell}^{p} |\Theta_{jki}| \right)^2.$$

Then

$$\tau_1(\mathcal{K}^p) = \sup_{z,y \in \mathcal{S}^p} \|\mathbf{e}_z^\top \mathcal{K}^p - \mathbf{e}_y^\top \mathcal{K}^p\|_{\mathrm{TV}} \le g_f(\Theta^*).$$

□

Proof Let $\mathbb{P}_z(\cdot)$ denote $\mathbb{P}(X_t^{t+p-1} = \cdot \mid X_{t-p}^{t-1} = z)$ for any t, since the process is assumed to be time-invariant. Hence $\tau_1(\mathcal{K}^p) = \sup_{z,y \in \mathcal{S}^p} \|\mathbb{P}_z - \mathbb{P}_y\|_{\mathrm{TV}}$. By Pinsker's inequality, we have

$$\|\mathbb{P}_z - \mathbb{P}_y\|_{\mathrm{TV}}^2 \le \frac{1}{2} D_{\mathrm{KL}}(\mathbb{P}_z \| \mathbb{P}_y).$$

Now, observe that

$$\begin{aligned}
D_{\mathrm{KL}}(\mathbb{P}_z \| \mathbb{P}_y) &:= \mathop{\mathbb{E}}_{X \sim \mathbb{P}_z} \log \frac{\mathbb{P}_z(X)}{\mathbb{P}_y(X)} = \mathop{\mathbb{E}}_{X \sim \mathbb{P}_z} \log \prod_{i=1}^{p} \frac{\mathbb{P}(X_i | X_1^{i-1}; X_{1-p}^0 = z)}{\mathbb{P}(X_i | X_1^{i-1}; X_{1-p}^0 = y)} \\
&= \mathop{\mathbb{E}}_{X \sim \mathbb{P}_z} \sum_{i=1}^{p} \log \frac{\mathbb{P}(X_i | X_1^{i-1}; X_{1-p}^0 = z)}{\mathbb{P}(X_i | X_1^{i-1}; X_{1-p}^0 = y)} = \sum_{i=1}^{p} \mathop{\mathbb{E}}_{X \sim \mathbb{P}_z} \log \frac{\mathbb{P}(X_i | X_1^{i-1}; X_{1-p}^0 = z)}{\mathbb{P}(X_i | X_1^{i-1}; X_{1-p}^0 = y)} \\
&= \sum_{i=1}^{p} \mathop{\mathbb{E}}_{X \sim \mathbb{P}_z} \mathbb{E} \left[\log \left(\frac{\mathbb{P}(X_i | X_1^{i-1}; X_{1-p}^0 = z)}{\mathbb{P}(X_i | X_1^{i-1}; X_{1-p}^0 = y)} \right) \middle| X_1^{i-1}; X_{1-p}^0 = z \right] \\
&= \sum_{i=1}^{p} \mathop{\mathbb{E}}_{X \sim \mathbb{P}_z} D_{\mathrm{KL}} \left(\mathbb{P}(X_i | X_1^{i-1}; X_{1-p}^0 = z) \middle\| \mathbb{P}(X_i | X_1^{i-1}; X_{1-p}^0 = y) \right).
\end{aligned}$$

Now we know for i.i.d. bernoulli random vectors with mean vectors μ and ν, using Lemma B.2, their KL-divergence is upper bounded as $D_{\mathrm{KL}}(\mathrm{Ber}(\mu) \| \mathrm{Ber}(\nu)) \le \frac{3}{4\varepsilon(1-\varepsilon)} \|\mu - \nu\|_2^2 = \frac{3}{4\varepsilon(1-\varepsilon)} \sum_{j=1}^{N} (\mu_j - \nu_j)^2$. The Lpschitzness of f thus results in $D_{\mathrm{KL}}(\mathbb{P}_z \| \mathbb{P}_y)$ being upper bounded by

$$\sum_{\ell=1}^{p} \sum_{j=1}^{N} \frac{3}{4\varepsilon(1-\varepsilon)} L_f^2 \left| \sum_{i=\ell}^{p} \langle \Theta_{j*i}, (z_{i-\ell+1} - y_{i-\ell+1}) \rangle \right|^2 \le \sum_{i=1}^{p} \sum_{j=1}^{N} \frac{3L_f^2}{2\varepsilon} \left(\sum_{i=\ell}^{p} \|\Theta_{j*i}\|_1 \right)^2 =: g^2(p; \Theta; f),$$

since X_1^{i-1} are common to evaluation of means for both distributions. We also used $\varepsilon \le \frac{1}{2}$ above. ∎

Lemma C.2 *For a p-lag process over $\mathcal{S}$, with an equivalent kernel representation $\mathcal{K} \in \mathbb{R}^{|\mathcal{S}|^p \times |\mathcal{S}|^p}$ given by (29),*

$$\eta_{k\ell} \le \tau_1(\mathcal{K}^p)^{1 + \lfloor (\ell - k - 1)/p \rfloor}.$$

Consequently, if $\tau_1(\mathcal{K}^p) < 1$, then using Lemma C.1,

$$\left(\sum_{\ell \ge k} \eta_{k\ell} \right)^2 \le \left(\sum_{\ell \ge k} \tau_1(\mathcal{K}^p)^{1 + \lfloor (\ell - k - 1)/p \rfloor} \right)^2 \le \left(1 + \frac{p\tau_1(\mathcal{K}^p)}{1 - \tau_1(\mathcal{K}^p)} \right)^2 \le 2 + \frac{2p^2}{(\frac{1}{g(p,\Theta,f)} - 1)^2} =: \frac{1}{4c_f^2} G(p, \Theta, f).$$

□

Proof Let $z \in \mathcal{S}^{k-p}$. Observe that the $\eta_{k\ell}$-mixing coefficient is the $\sup_{w,w',y,z}$ of

$$\|\mathbb{P}(X_\ell^n | X_1^k = wyz) - \mathbb{P}(X_\ell^n | X_1^k = w'yz)\|_{\mathrm{TV}} = \frac{1}{2} \sum_{x_\ell^n} |\mathbb{P}(X_\ell^n = x_\ell^n | X_{k-p+1}^k = wy) - \mathbb{P}(X_\ell^n = x_\ell^n | X_{k-p+1}^k = w'y)|,$$

which follows from the process being p-Markov. This in turn is equal to

$$
\begin{aligned}
=&\frac{1}{2}\sum_{x_\ell^{\ell+p-1}}\sum_{x_{\ell+p}^{n}}\Big|\mathbb{P}(X_{\ell+p}^{n}=x_{\ell+p}^{n}|X_\ell^{\ell+p-1}=x_\ell^{\ell+p-1};X_{k-p+1}^{k}=wy)\mathbb{P}(X_\ell^{\ell+p-1}=x_\ell^{\ell+p-1}|X_{k-p+1}^{k}=wy)\\
&-\mathbb{P}(X_{\ell+p}^{n}=x_{\ell+p}^{n}|X_\ell^{\ell+p-1}=x_\ell^{\ell+p-1};X_{k-p+1}^{k}=w'y)\mathbb{P}(X_\ell^{\ell+p-1}=x_\ell^{\ell+p-1}|X_{k-p+1}^{k}=w'y)\Big|\\
=&\frac{1}{2}\sum_{x_\ell^{\ell+p-1}}\left(\sum_{x_{\ell+p}^{n}}\mathbb{P}(X_{\ell+p}^{n}=x_{\ell+p}^{n}|X_\ell^{\ell+p-1}=x_\ell^{\ell+p-1})\right)\Big|\mathbb{P}(X_\ell^{\ell+p-1}=x_\ell^{\ell+p-1}|X_{k-p+1}^{k}=wy)\\
&-\mathbb{P}(X_\ell^{\ell+p-1}=x_\ell^{\ell+p-1}|X_{k-p+1}^{k}=w'y)\Big|\\
=&\frac{1}{2}\sum_{x_\ell^{\ell+p-1}}\Big|\mathbb{P}(X_\ell^{\ell+p-1}=x_\ell^{\ell+p-1}|X_{k-p+1}^{k}=wy)-\mathbb{P}(X_\ell^{\ell+p-1}=x_\ell^{\ell+p-1}|X_{k-p+1}^{k}=w'y)\Big|\\
=&\frac{1}{2}\|(e_{wy}-e_{w'y})^\top\mathcal{K}^{\ell+p-1-i}\|_1.
\end{aligned}
$$

Note here $\mathbf{e}_i$ is the i^{th} row of identity in $\mathbb{R}^{|\mathcal{S}|^p\times|\mathcal{S}|^p}$, for $i\in\mathcal{S}^p$.Observe that $\ell-k-1=p\lfloor(\ell-k-1)/p\rfloor+(\ell-k-1 \bmod p)$. Applying equation (28) for stochastic matrices $\mathcal{K}^{(\ell-k-1 \bmod p)}$, and using $\frac{1}{2}(\mathbf{e}_{wy}-\mathbf{e}_{w'y})\in\mathcal{H}_1$we get

$$
\eta_{k\ell}\le\sup_{u\in\mathcal{H}_1}\|u^\top\mathcal{K}^{p+p\lfloor(\ell-k-1)/p\rfloor+(\ell-k-1 \bmod p)}\|_1\le\sup_{u\in\mathcal{H}_1}\|u^\top(\mathcal{K}^p)^{1+\lfloor(\ell-k-1)/p\rfloor}\|_1\le\tau_1(\mathcal{K}^p)^{1+\lfloor(\ell-k-1)/p\rfloor},\quad(30)
$$

where the last inequality follows again from equation (28). ∎

Appendix D Scaling of $g_f(\Theta)$ with p

Lemma D.1 *If $|\Theta_{jk\ell}|\le C_{jk}\cdot\ell^{-\alpha}$, for some $\alpha>\frac{3}{2}$. Then*

$$
g_f(\Theta)\le\sqrt{\frac{6\alpha}{\varepsilon(\alpha-3/2)}}L_f\|C\|_{2,1}.
$$

Similarly, if $|\Theta_{jk\ell}|\le(1-\beta)^\ell$, for some $\beta<1$, then

$$
g_f(\Theta)\le\frac{L_f\sqrt{3/2}}{\beta^{3/2}\varepsilon^{1/2}}\|C\|_{2,1}.
$$

□

Proof If a polynomial decay of $\Theta_{jk\ell}$ with $\frac{1}{\ell}$ is satisfied, then

$$
\sum_{k=1}^{N}\sum_{i=\ell}^{p}|\Theta_{jki}|\le\sum_{k=1}^{N}\sum_{i=\ell}^{\infty}|\Theta_{jki}|\le\sum_{k=1}^{N}\sum_{i=\ell}^{\infty}|C_{jk}|i^{-\alpha}.
$$

Now, $\sum_{i=\ell}^{\infty}i^{-\alpha}\le\frac{\alpha}{\alpha-1}\ell^{1-\alpha}$, by approximating the summation with an integral. Hence the summation in the definition of $g(\Theta;f)$ is at most

$$
\begin{aligned}
&\sum_{\ell=1}^{p}\sum_{j=1}^{N}\|C_{j*}\|_1^2\frac{\alpha^2}{(\alpha-1)^2}\ell^{2-2\alpha}=\frac{\alpha^2\|C\|_{2,1}^2}{(\alpha-1)^2}\sum_{\ell=1}^{\infty}\ell^{2-2\alpha}\\
\le&\frac{2\alpha^2}{(\alpha-1)(2\alpha-3)}\|C\|_{2,1}^2\le\frac{6\|C\|_{2,1}^2\alpha}{2\alpha-3},
\end{aligned}
$$

via another Reimann integral approximation. This shows the claim. Where we have used $\alpha>3/2$ to upper bound some terms.

Using the $(1-\beta)$ geometric decay of Θ_{*jk}, the summation in the definition of $g(\Theta; f)$ is at most

$$\begin{aligned}
&\sum_{i=1}^{p}\sum_{j=1}^{N}(\sum_{k=1}^{N}|C_{jk}|\sum_{\tau=i}^{p}(1-\beta)^{\tau})^2 \\
\leq &\sum_{i=1}^{p}\sum_{j=1}^{N}(\sum_{k=1}^{N}|C_{jk}|\frac{(1-\beta)^i}{\beta})^2 \\
\leq &\sum_{j=1}^{N}(\sum_{k=1}^{N}|C_{jk}|)^2\frac{(1-\beta)^2}{\beta^2(1-(1-\beta)^2)} \leq \frac{\|C\|_{2,1}^2}{\beta^3},
\end{aligned}$$

where we use $\beta < 1$ to upper bound some terms. ∎