# Online Supplement for "Solution Path of Time-varying Markov Random Fields with Discrete Regularization"

## Appendix A: Proofs

### A.1. Proof of Theorem 2

*Estimation error bound.* The solution of ProxGL satisfies

$$\left\|\widehat{\theta}_t - \theta_t^\star\right\|_\infty \leq \left\|\widehat{\theta}_t - \widetilde{F}^*(\widehat{\mu}_t)\right\|_\infty + \left\|\theta_t^\star - \widetilde{F}^*(\widehat{\mu}_t)\right\|_\infty \leq 2\lambda_t,$$

where in the first inequality, we used the triangle inequality, and in the second inequality, we used the assumption that the true canonical parameter $\{\theta_t^\star\}_{t=0}^T$ is feasible for ProxGL.

*Sparsistency.* Next, we show the sparsistency of the estimated parameters for any $q \in \{0\} \cup [1, \infty)$. We divide our proof into two parts. The first part establishes the sparsistency of individual parameters for smoothly-changing MRFs, while the second part extends the analysis to prove the sparsistency of individual parameters and their differences for sparsely-changing MRFs.

**Smoothly-changing MRFs.** We will establish $\operatorname{supp}(\widehat{\theta}_t) = \operatorname{supp}(\theta_t^\star)$ for any $q \geq 1$. To this goal, we show $\operatorname{supp}(\widehat{\theta}_t) \subseteq \operatorname{supp}(\theta_t^\star)$ and $\operatorname{supp}(\theta_t^\star) \subseteq \operatorname{supp}(\widehat{\theta}_t)$. For any $i \in \mathcal{S}_t$, we have

$$\begin{aligned}
\left|\widehat{\theta}_{t;i}\right| &= \left|\widehat{\theta}_{t;i} - \theta_{t;i}^\star + \theta_{t;i}^\star\right| \\
&\geq \left|\theta_{t;i}^\star\right| - \left|\widehat{\theta}_{t;i} - \theta_{t;i}^\star\right| \\
&\geq \left|\theta_{t;i}^\star\right| - 2\lambda_t \\
&> 0,
\end{aligned}$$

where in the second inequality we used the upper bound on the estimation error, and in the last inequality we used the assumption $2\lambda_t \leq \min_{i \in \mathcal{S}_t} |\theta_{t;i}^\star|$. This implies that $\operatorname{supp}(\theta_t^\star) \subseteq \operatorname{supp}(\widehat{\theta}_t)$. To prove $\operatorname{supp}(\widehat{\theta}_t) \subseteq \operatorname{supp}(\theta_t^\star)$, we first rewrite ProxGL as follows:

$$\begin{aligned}
\{\widehat{\theta}_t\}_{t=0}^T = \arg \min_{\{\theta_t\}_{t=0}^T} \; & \gamma \sum_{t=0}^{T} \sum_{i=1}^{n} \mathbb{I}[\theta_{t;i} \neq 0] + (1-\gamma) \sum_{t=1}^{T} \sum_{i=1}^{n} |\theta_{t;i} - \theta_{t-1;i}|^q \\
\text{s.t.} \; & \left|\theta_{t;i} - \left[\widetilde{F}^*(\widehat{\mu}_t)\right]_i\right| \leq \lambda_t \qquad \forall t = 0, \ldots, T, \; i = 1, \ldots, p.
\end{aligned} \tag{19}$$

We make the following observation. First, due to the use of the infinity norm for the backward mapping deviation, problem (19) decomposes into $p$ independent subproblems, one for each coordinate of $\theta_t$. On the other hand, due to

the optimality of $\{\widehat{\theta}_t\}_{t=0}^T$ and the feasibility of $\{\theta_t^\star\}_{t=0}^T$, we have for every $1 \le i \le p$

$$
\begin{aligned}
&(1-\gamma)\sum_{t=0}^{T}\mathbb{I}\{\widehat{\theta}_{t;i}\neq 0\}+\gamma\sum_{t=1}^{T}\left|\widehat{\theta}_{t;i}-\widehat{\theta}_{t-1;i}\right|^q \le (1-\gamma)\sum_{t=0}^{T}\mathbb{I}\{\theta^\star_{t;i}\neq 0\}+\gamma\sum_{t=1}^{T}\left|\theta^\star_{t;i}-\theta^\star_{t-1;i}\right|^q\\
\Longrightarrow\ &(1-\gamma)\sum_{t=0}^{T}\mathbb{I}\{\widehat{\theta}_{t;i}\neq 0\} \le (1-\gamma)\sum_{t=0}^{T}\mathbb{I}\{\theta^\star_{t;i}\neq 0\}+\gamma D\\
\Longrightarrow\ &(1-\gamma)\sum_{t=0}^{T}\mathbb{I}\{\widehat{\theta}_{t;i}\neq 0\}\mathbb{I}\{i\in\mathcal{S}_t\}+(1-\gamma)\sum_{t=0}^{T}\mathbb{I}\{\widehat{\theta}_{t;i}\neq 0\}\mathbb{I}\{i\in[n]\backslash\mathcal{S}_t\} \le (1-\gamma)\sum_{t=0}^{T}\left|\theta^\star_{t;i}\right|_0+\gamma D_q \qquad (20)\\
\Longrightarrow\ &(1-\gamma)\sum_{t=0}^{T}\mathbb{I}\{\widehat{\theta}_{t;i}\neq 0\}\mathbb{I}\{i\in[n]\backslash\mathcal{S}_t\} \le \gamma D\\
\Longrightarrow\ &\sum_{t=0}^{T}\mathbb{I}\{\widehat{\theta}_{t;i}\neq 0\}\mathbb{I}\{i\in[n]\backslash\mathcal{S}_t\} < 1
\end{aligned}
$$

where the second inequality follows from the assumption $\sum_{t=1}^T\sum_{i\in[p]}|\theta^*_{t;i}-\theta^*_{t-1;i}|^q \le D$. Moreover, the fourth inequality follows from $\mathrm{supp}(\theta_t^\star)\subseteq\mathrm{supp}(\widehat{\theta}_t)$. The last inequality follows from the assumption $0<\gamma<1/(1+D)$. This implies that $\mathbb{I}\{\widehat{\theta}_{t;i}\neq 0\}\mathbb{I}\{i\in[n]\backslash\mathcal{S}_t\}=0$ for every $t$ and $i$, and hence, we have $\mathrm{supp}(\widehat{\theta}_t)\subseteq\mathrm{supp}(\theta_t^\star)$.

**Smoothly-changing MRFs.** Finally, we show that for sparsely-changing MRFs, ProxGL with temporal $\ell_0$-regularizer satisfies $\mathrm{supp}(\widehat{\theta}_t)=\mathrm{supp}(\theta_t^\star)$ and $\mathrm{supp}(\widehat{\theta}_t-\widehat{\theta}_{t-1})=\mathrm{supp}(\theta_t^\star-\theta_{t-1}^\star)$. An argument identical to $q\ge 1$ can be invoked to show that $\mathrm{supp}(\theta_t^\star)\subseteq\mathrm{supp}(\widehat{\theta}_t)$. Similarly, given any $i\in\mathcal{D}_t$, one can write

$$
\begin{aligned}
\left|\widehat{\theta}_{t;i}-\widehat{\theta}_{t-1;i}\right| &= \left|\widehat{\theta}_{t;i}-\theta^*_{t;i}+\theta^*_{t;i}-\theta^*_{t-1;i}+\theta^*_{t-1;i}-\widehat{\theta}_{t-1;i}\right|\\
&\ge \left|\theta^*_{t;i}-\theta^*_{t-1;i}\right|-\left|\widehat{\theta}_{t;i}-\theta^*_{t;i}\right|-\left|\widehat{\theta}_{t-1;i}-\theta^*_{t-1;i}\right|\\
&\ge \left|\theta^*_{t;i}-\theta^*_{t-1;i}\right|-2\lambda_t-2\lambda_{t-1}\\
&> 0,
\end{aligned}
$$

where the last inequality is due to our assumption $2\lambda_t+2\lambda_{t-1}\le\min_{i\in\mathcal{D}_t}|\theta^\star_{t;i}-\theta^\star_{t-1;i}|$. This implies that $\mathrm{supp}(\theta_t^\star-\theta_{t-1}^\star)\subseteq\mathrm{supp}(\widehat{\theta}_t-\widehat{\theta}_{t-1})$. On the other hand, due to the optimality of $\{\widehat{\theta}_t\}_{t=0}^T$ and the feasibility of $\{\theta_t^\star\}_{t=0}^T$, we have

$$
\begin{aligned}
&(1-\gamma)\sum_{t=0}^{T}\|\widehat{\theta}_t\|_0+\gamma\sum_{t=1}^{T}\|\widehat{\theta}_t-\widehat{\theta}_{t-1}\|_0 \le (1-\gamma)\sum_{t=0}^{T}\|\theta^*_t\|_0+\gamma\sum_{t=1}^{T}\|\theta^*_t-\theta^*_{t-1}\|_0\\
\Longrightarrow\ &(1-\gamma)\sum_{t=0}^{T}\left(\sum_{i\in[n]\backslash\mathcal{S}_t}\mathbb{I}\{\widehat{\theta}_{t;i}\neq 0\}+\sum_{i\in\mathcal{S}_t}\mathbb{I}\{\widehat{\theta}_{t;i}\neq 0\}\right)\\
&+\gamma\sum_{t=1}^{T}\left(\sum_{i\in[n]\backslash\mathcal{D}_t}\mathbb{I}\{\widehat{\theta}_{t;i}-\widehat{\theta}_{t-1;i}\neq 0\}+\sum_{i\in\mathcal{D}_t}\mathbb{I}\{\widehat{\theta}_{t;i}-\widehat{\theta}_{t-1;i}\neq 0\}\right) \qquad (21)\\
&\le (1-\gamma)\sum_{t=0}^{T}\sum_{i\in\mathcal{S}_t}\mathbb{I}\{\theta^\star_{t;i}\neq 0\}+\gamma\sum_{t=1}^{T}\sum_{i\in\mathcal{D}_t}\mathbb{I}\{\theta^\star_{t;i}-\theta^\star_{t-1;i}\neq 0\}\\
\Longrightarrow\ &(1-\gamma)\sum_{t=0}^{T}\sum_{i\in[n]\backslash\mathcal{S}_t}\mathbb{I}\{\widehat{\theta}_{t;i}\neq 0\}+\gamma\sum_{t=1}^{T}\sum_{i\in[n]\backslash\mathcal{D}_t}\mathbb{I}\{\widehat{\theta}_{t;i}-\widehat{\theta}_{t-1;i}\neq 0\}\le 0
\end{aligned}
$$

where the last inequality follows from $\mathrm{supp}(\theta_t^*) \subseteq \mathrm{supp}(\widehat{\theta}_t)$ and $\mathrm{supp}(\theta_t^* - \theta_{t-1}^*) \subseteq \mathrm{supp}(\widehat{\theta}_t - \widehat{\theta}_{t-1})$. Due to $0 < \gamma < 1$, the above inequality implies that $\widehat{\theta}_{t;i} = 0$ for every $t = 0, \ldots, T$ and $i \in [n] \backslash \mathcal{S}_t$, and $\widehat{\theta}_{t;i} - \widehat{\theta}_{t-1;i} = 0$ for every $t = 1, \ldots, T$ and $i \in [n] \backslash \mathcal{D}_t$. Therefore, we have $\mathrm{supp}(\widehat{\theta}_t) \subseteq \mathrm{supp}(\theta_t^*)$ and $\mathrm{supp}(\widehat{\theta}_t - \widehat{\theta}_{t-1}) \subseteq \mathrm{supp}(\theta_t^* - \theta_{t-1}^*)$. This completes the proof. □

### A.2. Proof of Lemma 6

We only prove the first statement in the lemma, since the other two follow from identical arguments. Given any index $a < \tau < b - 1$, consider optimization (9) where all values except for $\theta_\tau$ are fixed to any feasible value (we omit terms that do not depend on $\theta_\tau$)

$$\begin{aligned} &\min_{\theta_\tau} |\theta_\tau - \theta_{\tau-1}|^q + |\theta_{\tau+1} - \theta_\tau|^q \\ &\text{s.t. } \ell_\tau \leq \theta_\tau \leq u_\tau. \end{aligned} \tag{22}$$

Observe that if an optimal solution $\theta_\tau^*$ of (22) satisfies $\theta_\tau^* > \max\{\theta_{\tau-1}, \theta_{\tau+1}\}$, then necessarily $\theta_\tau^* = \ell_\tau$, since otherwise it would be possible to decrease $\theta_\tau^*$, improving the objective value. Similarly the case $\theta_\tau^* < \min\{\theta_{\tau-1}, \theta_{\tau+1}\}$ implies that $\theta_\tau^* = u_\tau$.

To prove the first statement, assume that $\theta_\tau^* \notin \{\ell_t, u_t\}$, which can only happen if $\min\{\theta_{\tau-1}, \theta_{\tau+1}\} \leq \theta_\tau^* \leq \max\{\theta_{\tau-1}, \theta_{\tau+1}\}$. If $\theta_{\tau-1} \leq \theta_{\tau+1}$, then it follows that $\theta_\tau$ is optimal for

$$\min_{\theta_\tau \in \mathbb{R}} (\theta_\tau - \theta_{\tau-1})^q + (\theta_{\tau+1} - \theta_\tau)^q . \tag{23}$$

First, suppose that $q > 1$. Then, by convexity and differentiability, the derivative of (23) at the optimal solution should be zero. This implies

$$q(\theta_t - \theta_{t-1})^{q-1} - q(\theta_{t+1} - \theta_t)^{q-1} = 0 \Leftrightarrow \theta_t - \theta_{t-1} = \theta_{t+1} - \theta_t \Leftrightarrow \theta_\tau = (\theta_{t+1} + \theta_{t-1})/2.$$

Moreover, if $q = 1$, it can be easily seen that any feasible value $\theta_\tau \in [\theta_{\tau-1}, \theta_{\tau+1}]$ is optimal. Thus, at least one of the following five cases holds: *(i)* $\theta_\tau^* = (\theta_{t+1} + \theta_{t-1})/2 \in [\theta_{\tau-1}, \theta_{\tau+1}]$ is optimal; *(ii)* $\theta_\tau^* = \ell_t \in [\theta_{\tau-1}, \theta_{\tau+1}]$ is optimal; *(iii)* $\theta_\tau^* = u_t \in [\theta_{\tau-1}, \theta_{\tau+1}]$; *(iv)* $\theta_\tau^* = \ell_t \geq \theta_{\tau+1}$ is optimal; or *(v)* $\theta_\tau^* = u_t \leq \theta_{\tau-1}$ is optimal. Since all fives cases satisfy $\theta_\tau^* \in \{\ell_\tau, (\theta_{t+1} + \theta_{t-1})/2, u_\tau\}$, the statement of the lemma holds.

The case $\theta_{\tau-1} \geq \theta_{\tau+1}$ is handled identically, concluding the proof. □

## Appendix B: Kernel Averaging

In most realistic time-varying MRFs, the underlying graphical model changes continuously over time as the samples continue to arrive. For instance, in financial markets, the underlying stock correlation network may change quickly in response to global events. Therefore, a stock holder needs to identify the sharp changes in the market and rebalance their portfolio "on the go" (Talih and Hengartner 2005, Hallac et al. 2017). Evidently, in these applications, the sample size $N_t$ may be significantly smaller than what is required for ProxGL to provide a reliable estimation of the canonical parameters.

To address the scarcity of data in this setting, we leverage and combine the information provided by the samples over time. In particular, we propose to replace the empirical mean parameters with their weighted averages over time,

where the weights are obtained from a nonparametric kernel $K : \mathbb{R} \to \mathbb{R}_+$. Without loss of generality, suppose that $N_t = 1$[2] for every $t = 0, \ldots, T$ and consider the weighted mean parameter

$$\widehat{\mu}_t^{\text{ker}} = \sum_{s=0}^{T} w(s,t)\phi(x_s), \quad \text{where } w(s,t) = \frac{1}{Th} K\left(\frac{s-t}{Th}\right). \tag{24}$$

Here, $K(\cdot)$ is a symmetric nonnegative kernel that satisfies a set of mild conditions. Two mostly common kernels, namely the uniform kernel $K(x) = 1/2$ with domain $[-1, 1]$ and the truncated Gaussian kernel $K(x) = (\Phi(1) - \Phi(-1))^{-1} e^{-x^2/2}$ with $\Phi$ denoting the CDF of the normal distribution are valid choices. Moreover, the parameter $h$ is the *bandwidth* of the kernel, controlling the decay rate of the weights. The key insight behind kernel averaging is simple: at any given time $t$, we estimate the mean parameters by taking the weighted average of the samples over time, where the weights are obtained from a kernel that assigns smaller weights to samples that are temporally farther away from $t$.

For the theoretical analysis of kernel averaging in the context of the inference of time-varying MRFs, we refer the reader to the extended version of the paper (Fattahi and Gómez 2023).

## Appendix C: Additional experiments

### C.1. Further experiments on the effect of $\nu_0$

We test the performance of all methods by varying the number of samples per time period. Specifically, we fix $n = 50$, $T = 10$ and let $N_t = n\kappa$ for integer $1 \leq \kappa \leq 20$. Table 3 presents the results for $\kappa = \{1, 5, 10, 15, 20\}$ and varying levels of threshold in the approximate backward mapping. Note that $\nu_0 = 0$ corresponds to no thresholding, whereas $\nu_0 = 2$ sets every off-diagonal entry of the sample covariance matrix to zero. As can be seen in the table, ProxGL with $\nu_0 \in \{0, 2\}$ is clearly inferior to other choices of the parameters, so these values are excluded from our subsequent experiments.

### C.2. On using MIP solvers

An alternative to solving (7) or (11) is to simply resort to using a mixed-integer optimization solver. However, this direct approach may result in a significant computational overhead. Indeed, even if solvers are able to efficiently tackle problems for a given value of the regularization parameter $\bar{\gamma}$ or $k$, computing solutions for all possible values of the parameter would require solving $\mathcal{O}(T)$ mixed-integer problems.

[2] If $N_t > 1$, the sufficient statistics $\phi(x_s)$ in (24) can be replaced by $\frac{1}{N_t} \sum_{i=1}^{N_t} \phi(x_s^{(i)})$.

To illustrate, we consider the experiments with real data described in §6.2. Instead of tackling (11) with $q = 0$ using Algorithm 1, which solves the problem for all cardinalities $k$, we use the formulation

$$
\begin{aligned}
&\min_{\theta,z,w} \sum_{t=1}^{T} w_t \\
&\text{s.t.} \sum_{t=0}^{T} z_t \leq k \\
&\quad \ell_t z_t \leq \theta_t \leq u_t z_t && \forall t = 0, \ldots, T \\
&\quad z_t = 1 && \forall t : 0 \notin [\ell_t, u_t] \\
&\quad \theta_t - \theta_{t-1} \leq (u_t - \ell_{t-1}) w_t && \forall t = 1, \ldots, T \\
&\quad \theta_{t-1} - \theta_t \leq (u_{t-1} - \ell_t) w_t && \forall t = 1, \ldots, T \\
&\quad \theta \in \mathbb{R}^{T+1},\ z \in \{0,1\}^{T+1}, w \in \{0,1\}^{T}
\end{aligned}
\tag{25}
$$

and solve (25) for all $k \in \{0, \ldots, T+1\}$ using Gurobi 12. Table 4 reports the solution times obtained using the DP approach described in Algorithm 1, as well as the times required to solve the mixed-integer optimization problem (25) using Gurobi. Overall, the DP approach can compute complete solution paths an order-of-magnitude fastet than the time that Gurobi requires to solve the problem for a single value of the parameter.

Table 3: `F1-score` of precision matrices (F1_p), differences (F1_d), and estimation `error` for method ProxGL for different values of the shrinkage parameter $\nu$.

| $\frac{N_t}{n}$ | **ProxGL** ($\nu_0 = 0.0$) | | | **ProxGL** ($\nu_0 = 0.2$) | | | **ProxGL** ($\nu_0 = 0.5$) | | | **ProxGL** ($\nu_0 = 0.8$) | | | **ProxGL** ($\nu_0 = 2.0$) | | | **TVGL** | | | **L1E** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **F1_p** | **F1_d** | **err** | **F1_p** | **F1_d** | **err** | **F1_p** | **F1_d** | **err** | **F1_p** | **F1_d** | **err** | **F1_p** | **F1_d** | **err** | **F1_p** | **F1_d** | **err** | **F1_p** | **F1_d** | **err** |
| 2 | 0.27 | 0.04 | 177.4% | 0.28 | 0.05 | 78.8% | 0.31 | 0.09 | 34.4% | 0.40 | 0.19 | 23.7% | 0.44 | 0.35 | 24.7% | 0.41 | 0.42 | 24.1% | 0.40 | 0.11 | 28.2% |
| 10 | 0.36 | 0.14 | 23.7% | 0.42 | 0.22 | 16.7% | 0.59 | 0.36 | 12.5% | 0.76 | 0.39 | 13.4% | 0.59 | 0.41 | 21.2% | 0.85 | 0.22 | 20.6% | 0.36 | 0.12 | 24.8% |
| 20 | 0.49 | 0.39 | 11.6% | 0.61 | 0.53 | 9.1% | 0.79 | 0.60 | 8.0% | 0.89 | 0.58 | 9.5% | 0.75 | 0.44 | 19.0% | 0.97 | 0.22 | 12.6% | 0.28 | 0.12 | 23.6% |
| 30 | 0.65 | 0.63 | 7.8% | 0.78 | 0.70 | 6.5% | 0.87 | 0.73 | 6.3% | 0.92 | 0.68 | 7.8% | 0.83 | 0.46 | 17.1% | 0.91 | 0.20 | 9.2% | 0.27 | 0.12 | 22.2% |
| 40 | 0.80 | 0.76 | 6.0% | 0.89 | 0.78 | 5.3% | 0.93 | 0.77 | 5.5% | 0.94 | 0.73 | 6.8% | 0.89 | 0.46 | 15.3% | 0.85 | 0.18 | 7.4% | 0.26 | 0.12 | 21.0% |

**Table 4 Detailed computational time required to solve ProxGL (for all values of $\gamma$) on real instances with stock correlation networks, using either DP or a mixed-integer programming (MIP) solvers. Time to print the solutions to a file is not included. "Total" is the total time spent solving (11) via either DP or mixed-integer optimization; "solution path" is the average time spent per parameter solving for all values of the regularization parameter; "single param" is the average time solving for a single value of the regularization parameter.**

| N | T | # params | backwards mapping | Time DP | | | Time MIP | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | single param | solution path | total | single param | solution path | total |
| 60 | 117 | 10,716,147 | 8 | - | $1.3 \cdot 10^{-6}$ | 14 | $1.2 \cdot 10^{-5}$ | $1.5 \cdot 10^{-3}$ | 15,568 |
| 50 | 140 | 12,822,740 | 7 | - | $1.7 \cdot 10^{-6}$ | 22 | $9.5 \cdot 10^{-6}$ | $1.3 \cdot 10^{-3}$ | 17,018 |
| 40 | 176 | 16,120,016 | 8 | - | $3.5 \cdot 10^{-6}$ | 57 | $2.2 \cdot 10^{-5}$ | $3.9 \cdot 10^{-3}$ | 63,521 |
| 30 | 234 | 21,432,294 | 8 | - | $4.9 \cdot 10^{-6}$ | 105 | $4.1 \cdot 10^{-5}$ | $9.7 \cdot 10^{-3}$ | 209,619 |
| 20 | 351 | 32,148,441 | 9 | - | $2.2 \cdot 10^{-5}$ | 694 | | | † |

† Requires more than 3 days $\approx$ 260,000 seconds of computation