Link Prediction in Graphs with Autoregressive Features (supplementary material)

Emile Richard
CMLA UMR CNRS 8536,
ENS Cachan, France

Stéphane Gaïffas
CMAP - Ecole Polytechnique
& LSTA - Université Paris 6

Nicolas Vayatis
CMLA UMR CNRS 8536,
ENS Cachan, France

1 Estimation of low-rank graphs with autoregressive features

Our approach is based on the asumption that features can explain most of the information contained in the graph, and that these features are evolving with time. We make the following assumptions about the sequence $(A_t)_{t\geq 0}$ of adjacency matrices of the graphs sequence.

Low-Rank. We assume that the matrices A_t have low-rank. This reflects the presence of highly connected groups of nodes such as communities in social networks, or product categories and groups of loyal/fanatic users in a market place data, and is sometimes motivated by the small number of factors that explain nodes interactions.

Autoregressive linear features. We assume to be given a linear map $\omega : \mathbb{R}^{n\times n} \to \mathbb{R}^d$ defined by

$$\omega(A) = \Big(\langle \Omega_1, A\rangle, \cdots, \langle \Omega_d, A\rangle \Big), \tag{1}$$

where $(\Omega_i)_{1\leq i\leq d}$ is a set of $n \times n$ matrices. These matrices can be either deterministic or random in our theoretical analysis, but we take them deterministic for the sake of simplicity. The vector time series $(\omega(A_t))_{t\geq 0}$ has autoregressive dynamics, given by a VAR (Vector Auto-Regressive) model:

$$\omega(A_{t+1}) = W_0^\top \omega(A_t) + N_{t+1},$$

where $W_0 \in \mathbb{R}^{d\times d}$ is a unknown sparse matrix and $(N_t)_{t\geq 0}$ is a sequence of noise vectors in $\mathbb{R}^d$. An example of linear features is the degree (*i.e.* number of edges connected to each node, or the sum of their weights if the edges are weighted), which is a measure of popularity in social and commerce networks. Introducing

$$\mathbf{X}_{T-1} = (\omega(A_0), \ldots, \omega(A_{T-1}))^\top \quad \text{and} \quad \mathbf{X}_T = (\omega(A_1), \ldots, \omega(A_T))^\top,$$

which are both $T \times d$ matrices, we can write this model in a matrix form:

$$\mathbf{X}_T = \mathbf{X}_{T-1} W_0 + \mathbf{N}_T, \tag{2}$$

where $\mathbf{N}_T = (N_1, \ldots, N_T)^\top$.

This assumes that the noise is driven by time-series dynamics (a martingale increment), where each coordinates are independent (meaning that features are independently corrupted by noise), with a sub-gaussian tail and variance uniformly bounded by a constant σ^2. In particular, no independence assumption between the N_t is required here.

Notations. The notations $\|\cdot\|_F$, $\|\cdot\|_p$, $\|\cdot\|_\infty$, $\|\cdot\|_*$ and $\|\cdot\|_{\mathrm{op}}$ stand, respectively, for the Frobenius norm, entry-wise ℓ_p norm, entry-wise ℓ_∞ norm, trace-norm (or nuclear norm, given by the sum of the singular values) and operator norm (the largest singular value). We denote by $\langle A, B\rangle = \mathrm{tr}(A^\top B)$ the Euclidean matrix product. A vector in $\mathbb{R}^d$ is always understood as a $d \times 1$ matrix. We denote by $\|A\|_0$ the number of non-zero elements of A. The product $A \circ B$ between two matrices with matching dimensions stands for the Hadamard or entry-wise product between A and B. The matrix $|A|$ contains the absolute values of entries of A. The matrix $(M)_+$ is the componentwise positive part of the matrix M, and $\operatorname{sgn}(M)$ is the sign matrix associated to M with the convention $\operatorname{sgn}(0) = 0$

If A is a $n \times n$ matrix with rank r, we write its SVD as $A = U\Sigma V^\top = \sum_{j=1}^r \sigma_j u_j v_j^\top$ where $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ is a $r \times r$ diagonal matrix containing the non-zero singular values of A in decreasing order, and $U = [u_1, \ldots, u_r]$, $V = [v_1, \ldots, v_r]$ are $n \times r$ matrices with columns given by the left and right singular vectors of A. The projection matrix onto the space spanned by the columns (resp. rows) of A is given by $P_U = UU^\top$ (resp. $P_V = VV^\top$). The operator $\mathcal{P}_A : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ given by $\mathcal{P}_A(B) = P_U B + BP_V - P_U BP_V$ is the projector onto the linear space spanned by the matrices $u_k x^\top$ and $yv_k^\top$ for $1 \leq j, k \leq r$ and $x, y \in \mathbb{R}^n$. The projector onto the orthogonal space is given by $\mathcal{P}_A^\perp(B) = (I - P_U)B(I - P_V)$. We also use the notation $a \vee b = \max(a, b)$.

1.1 Joint prediction-estimation through penalized optimization

In order to reflect the autoregressive dynamics of the features, we use a least-squares goodness-of-fit criterion that encourages the similarity between two feature vectors at successive time steps. In order to induce sparsity in the estimator of W_0, we penalize this criterion using the ℓ_1 norm. This leads to the following penalized objective function:

$$J_1(W) = \frac{1}{dT}\|\mathbf{X}_T - \mathbf{X}_{T-1}W\|_F^2 + \kappa\|W\|_1,$$

where $\kappa > 0$ is a smoothing parameter.

Now, for the prediction of A_{T+1}, we propose to minimize a least-squares criterion penalized by the combination of an ℓ_1 norm and a trace-norm. This mixture of norms induces sparsity and a low-rank of the adjacency matrix. Such a combination of ℓ_1 and trace-norm was already studied in [3] for the matrix regression model, and in [8] for the prediction of an adjacency matrix.

The objective function defined below exploits the fact that if W is close to W_0, then the features of the next graph $\omega(A_{T+1})$ should be close to $W^\top\omega(A_T)$. Therefore, we consider

$$J_2(A, W) = \frac{1}{d}\|\omega(A) - W^\top\omega(A_T)\|_F^2 + \tau\|A\|_* + \gamma\|A\|_1,$$

where $\tau, \gamma > 0$ are smoothing parameters. The overall objective function is the sum of the two partial objectives J_1 and J_2, which is jointly convex with respect to A and W:

$$\mathcal{L}(A, W) \doteq \frac{1}{dT}\|\mathbf{X}_T - \mathbf{X}_{T-1}W\|_F^2 + \kappa\|W\|_1 + \frac{1}{d}\|\omega(A) - W^\top\omega(A_T)\|_2^2 + \tau\|A\|_* + \gamma\|A\|_1, \quad (3)$$

If we choose convex cones $\mathcal{A} \subset \mathbb{R}^{n\times n}$ and $\mathcal{W} \subset \mathbb{R}^{d\times d}$, our joint estimation-prediction procedure is defined by

$$(\hat{A}, \hat{W}) \in \operatorname*{arg\,min}_{(A,W)\in\mathcal{A}\times\mathcal{W}} \mathcal{L}(A, W). \quad (4)$$

It is natural to take $\mathcal{W} = \mathbb{R}^{d\times d}$ and $\mathcal{A} = (\mathbb{R}_+)^{n\times n}$ since there is no *a priori* on the values of the feature matrix W_0, while the entries of the matrix A_{T+1} must be positive.

In the next section we propose oracle inequalities which prove that this procedure can estimate W_0 and predict A_{T+1} at the same time.

1.2 Main result

The central contribution of our work is to bound the prediction error with high probability under the following natural hypothesis on the noise process.

Assumption 1. *We assume that $(N_t)_{t\geq 0}$ satisfies $\mathbb{E}[N_t|\mathcal{F}_{t-1}] = 0$ for any $t \geq 1$ and that there is $\sigma > 0$ such that for any $\lambda \in \mathbb{R}$ and $j = 1, \ldots, d$ and $t \geq 0$:*

$$\mathbb{E}[e^{\lambda(N_t)_j}|\mathcal{F}_{t-1}] \leq e^{\sigma^2\lambda^2/2}.$$

Moreover, we assume that for each $t \geq 0$, the coordinates $(N_t)_1, \ldots, (N_t)_d$ are independent.

The main result can be summarized as follows. The prediction error and the estimation error can be simultaneously bounded by the sum of three terms that involve homogeneously (a) the sparsity, (b) the rank of the adjacency matrix A_{T+1}, and (c) the sparsity of the VAR model matrix W_0. The tight bounds we obtain are similar to the bounds of the Lasso and are upper bounded by:

$$C_1\sqrt{\frac{\log d}{Td^2}}\|W_0\|_0 + C_2\sqrt{\frac{\log n}{d}}\|A_{T+1}\|_0 + C_3\sqrt{\frac{\log n}{d}}\operatorname{rank} A_{T+1} \ .$$

The positive constants C_1, C_2, C_3 are proportional to the noise level σ. The interplay between the rank and sparsity constraints on A_{T+1} are reflected in the observation that the values of C_2 and C_3 can be changed as long as their sum remains constant.

2 Oracle inequalities

In this section we give oracle inequalities for the mixed prediction-estimation error which is given, for any $A \in \mathbb{R}^{n\times n}$ and $W \in \mathbb{R}^{d\times d}$, by

$$\mathcal{E}(A, W)^2 \doteq \frac{1}{d}\|(W - W_0)^\top \omega(A_T) - \omega(A - A_{T+1})\|_2^2 + \frac{1}{dT}\|\mathbf{X}_{T-1}(W - W_0)\|_F^2. \qquad (5)$$

It is important to have in mind that an upper-bound on $\mathcal{E}$ implies upper-bounds on each of its two components. It entails in particular an upper-bound on the feature estimation error $\|\mathbf{X}_{T-1}(\widehat{W} - W_0)\|_F$ that makes $\|(\widehat{W} - W_0)^\top \omega(A_T)\|_2$ smaller and consequently controls the prediction error over the graph edges through $\|\omega(\widehat{A} - A_{T+1})\|_2$.

The upper bounds on $\mathcal{E}$ given below exhibit the dependence of the accuracy of estimation and prediction on the number of features d, the number of edges n and the number T of observed graphs in the sequence.

Let us recall $\mathbf{N}_T = (N_1, \ldots, N_T)^\top$ and introduce the noise processes

$$M = -\sum_{j=1}^{d}(N_{T+1})_j \Omega_j \quad \text{and} \quad \Xi = \sum_{t=1}^{T+1}\omega(A_{t-1})N_t^\top,$$

which are, respectively, $n \times n$ and $d \times d$ random matrices. The source of randomness comes from the noise sequence $(N_t)_{t\geq 0}$, see Assumption 1. If these noise processes are controlled correctly, we can prove the following oracle inequalities for procedure (4). The next result is an oracle inequality of slow type (see for instance [1]), that holds in full generality.

Theorem 1. *Let $(\hat{A}, \hat{W})$ be given by* (4) *and suppose that*

$$\tau \geq \frac{2\alpha}{d}\|M\|_{\mathrm{op}}, \quad \gamma \geq \frac{2(1-\alpha)}{d}\|M\|_\infty \quad \textit{and} \quad \kappa \geq \frac{2}{dT}\|\Xi\|_\infty \qquad (6)$$

for some $\alpha \in (0, 1)$. Then, we have

$$\mathcal{E}(\widehat{A}, \widehat{W})^2 \leq \inf_{(A,W)\in\mathcal{A}\times\mathcal{W}}\Big\{\mathcal{E}(A, W)^2 + 2\tau\|A\|_* + 2\gamma\|A\|_1 + 2\kappa\|W\|_1\Big\}.$$

For the proof of oracle inequalities of fast type, the *restricted eigenvalue* (RE) condition introduced in [1] and [4, 5] is of importance. Restricted eigenvalue conditions are implied by, and in general weaker than, the so-called *incoherence* or RIP (Restricted isometry property, [2]) assumptions, which excludes, for instance, strong correlations between covariates in a linear regression model. This condition is acknowledged to be one of the weakest to derive fast rates for the Lasso (see [10] for a comparison of conditions).

Matrix version of these assumptions are introduced in [6]. Below is a version of the RE assumption that fits in our context. First, we need to introduce the two restriction cones.

The first cone is related to the $\|W\|_1$ term used in procedure (4). If $W \in \mathbb{R}^{d\times d}$, we denote by $\Theta_W = \operatorname{sign}(W) \in \{0, \pm 1\}^{d\times d}$ the signed sparsity pattern of W and by $\Theta_W^\perp \in \{0, 1\}^{d\times d}$ the orthogonal sparsity pattern. For a fixed matrix $W \in \mathbb{R}^{d\times d}$ and $c > 0$, we introduce the cone

$$\mathcal{C}_1(W, c) \doteq \Big\{W' \in \mathcal{W} : \|\Theta_W^\perp \circ W'\|_1 \leq c\|\Theta_W \circ W'\|_1\Big\}.$$

This cone contains the matrices W' that have their largest entries in the sparsity pattern of W.

The second cone is related to mixture of the terms $\|A\|_*$ and $\|A\|_1$ in procedure (4). Before defining it, we need further notations and definitions.

For a fixed $A \in \mathbb{R}^{n \times n}$ and $c, \beta > 0$, we introduce the cone

$$\mathcal{C}_2(A, c, \beta) \doteq \Big\{ A' \in \mathcal{A} : \|\mathcal{P}_A^\perp(A')\|_* + \beta \|\Theta_A^\perp \circ A'\|_1 \leq c \Big(\|\mathcal{P}_A(A')\|_* + \beta \|\Theta_A \circ A'\|_1 \Big) \Big\}.$$

This cone consist of the matrices A' with large entries close to that of A and that are "almost aligned" with the row and column spaces of A. The parameter β quantifies the interplay between these too notions.

Definition 1 (Restricted Eigenvalue (RE))**.** *For $W \in \mathcal{W}$ and $c > 0$, we introduce*

$$\mu_1(W, c) = \inf \Big\{ \mu > 0 : \|\Theta_W \circ W'\|_F \leq \frac{\mu}{\sqrt{dT}} \|\mathbf{X}_{T+1} W'\|_F, \ \ \forall W' \in \mathcal{C}_1(W, c) \Big\}.$$

For $A \in \mathcal{A}$ and $c, \beta > 0$, we introduce

$$\begin{aligned} \mu_2(A, W, c, \beta) = \inf \Big\{ \mu > 0 : \|\mathcal{P}_A(A')\|_F \vee \|\Theta_A \circ A'\|_F \\ \leq \frac{\mu}{\sqrt{d}} \|W'^\top \omega(A_T) - \omega(A')\|_2, \ \ \forall W' \in \mathcal{C}_1(W, c), \forall A' \in \mathcal{C}_2(A, c, \beta) \Big\}. \end{aligned}$$

The RE assumption consists of assuming that the constants μ_1 and μ_2 are non-zero. Now we can state the following Theorem that gives a fast oracle inequality for our procedure using RE.

Theorem 2. *Let $(\hat{A}, \hat{W})$ be given by* (4) *and suppose that*

$$\tau \geq \frac{3\alpha}{d} \|M\|_{\mathrm{op}}, \quad \gamma \geq \frac{3(1-\alpha)}{d} \|M\|_\infty \quad \textit{and} \quad \kappa \geq \frac{3}{dT} \|\Xi\|_\infty \tag{7}$$

for some $\alpha \in (0, 1)$. Then, we have

$$\begin{aligned} \mathcal{E}(\widehat{A}, \widehat{W})^2 \leq \inf_{(A,W) \in \mathcal{A} \times \mathcal{W}} \Big\{ \mathcal{E}(A, W)^2 + \frac{25}{18} \mu_2(A, W)^2 \big(\operatorname{rank}(A) \tau^2 + \|A\|_0 \gamma^2 \big) \\ + \frac{25}{36} \mu_1(W)^2 \|W\|_0 \kappa^2 \Big\}, \end{aligned}$$

where $\mu_1(W) = \mu_1(W, 10)$ and $\mu_2(A, W) = \mu_2(A, W, 10, \gamma/\tau)$ (see Definition 1).

The proofs of Theorems 1 and 2 use tools introduced in [6] and [1].

Note that the residual term from this oracle inequality mixes the notions of sparsity of A and W via the terms $\operatorname{rank}(A)$, $\|A\|_0$ and $\|W\|_0$. It says that our mixed penalization procedure provides an optimal trade-off between fitting the data and complexity, measured by both sparsity and low-rank. This is the first result of this nature to be found in literature.

In the next Theorem 3, we obtain convergence rates for the procedure (4) by combining Theorem 2 with controls on the noise processes. We introduce

$$\begin{aligned} v_{\Omega,\mathrm{op}}^2 &= \Big\| \frac{1}{d} \sum_{j=1}^d \Omega_j^\top \Omega_j \Big\|_{\mathrm{op}} \vee \Big\| \frac{1}{d} \sum_{j=1}^d \Omega_j \Omega_j^\top \Big\|_{\mathrm{op}}, \quad v_{\Omega,\infty}^2 = \Big\| \frac{1}{d} \sum_{j=1}^d \Omega_j \circ \Omega_j \Big\|_\infty, \\ \sigma_\omega^2 &= \max_{j=1,\ldots,d} \frac{1}{T+1} \sum_{t=1}^{T+1} \omega_j(A_{t-1})^2, \end{aligned}$$

which are the (observable) variance terms that naturally appear in the controls of the noise processes. We introduce also

$$\ell_T = 2 \max_{j=1,\ldots,d} \log \log \left(\frac{\sum_{t=1}^{T+1} \omega_j(A_{t-1})^2}{T+1} \vee \frac{T+1}{\sum_{t=1}^{T+1} \omega_j(A_{t-1})^2} \vee e \right),$$

which is a small (observable) technical term that comes out of our analysis of the noise process Ξ. This term is a small price to pay for the fact that no independence assumption is required on the noise sequence $(N_t)_{t \geq 0}$, but only a martingale increment structure with sub-gaussian tails.

Theorem 3. *Consider the procedure $(\hat{A}, \hat{W})$ given by* (4) *with smoothing parameters given by*

$$\tau = 3\alpha\sigma v_{\Omega,\mathrm{op}}\sqrt{\frac{2(x+\log(2n))}{d}}, \quad \gamma = 3(1-\alpha)\sigma v_{\Omega,\infty}\sqrt{\frac{2(x+2\log n)}{d}},$$
$$\kappa = 6\sigma\sigma_\omega \frac{1}{d}\sqrt{\frac{2e(x+2\log d+\ell_T)}{T+1}}$$

for some $\alpha \in (0,1)$ and fix a confidence level $x > 0$. Then, we have

$$\begin{aligned}\mathcal{E}(\widehat{A}, \widehat{W})^2 \leq \inf_{(A,W)\in\mathcal{A}\times\mathcal{W}} \Big\{ & \mathcal{E}(A,W)^2 + 25\mu_2(A)^2 \operatorname{rank}(A)\alpha^2\sigma^2 v_{\Omega,\mathrm{op}}^2 \frac{2(x+\log(2n))}{d} \\ & + 25\mu_2(A)^2\|A\|_0(1-\alpha)^2\sigma^2 v_{\Omega,\infty}^2 \frac{2(x+2\log n)}{d} \\ & + 25\mu_1(W)^2\|W\|_0\sigma^2\sigma_\omega^2 \frac{2e(x+2\log d+\ell_T)}{d^2(T+1)} \Big\}\end{aligned}$$

with a probability larger than $1 - 17e^{-x}$, where μ_1 and μ_2 are the same as in Theorem 2.

The proof of Theorem 3 follows directly from Theorem 2 basic noise control results. In the next Theorem, we propose more explicit upper bounds for both the indivivual estimation of W_0 and the prediction of A_{T+1}.

Theorem 4. *Under the same assumptions as in Theorem 3, for any $x > 0$ the following inequalities hold with a probability larger than $1 - 17e^{-x}$:*

$$\begin{aligned}&\frac{1}{dT}\|\mathbf{X}_T(\hat{W} - W_0)\|_F^2 \\ &\quad \leq \inf_{A\in\mathcal{A}} \Big\{ \frac{1}{d}\|\omega(A) - \omega(A_{T+1})\|_F^2 + \frac{25}{18}\mu_2(A,W)^2\big(\operatorname{rank}(A)\tau^2 + \|A\|_0\gamma^2\big)\Big\} \\ &\quad\quad + \frac{25}{36}\mu_1(W_0)^2\|W_0\|_0\kappa^2\end{aligned} \tag{8}$$

$$\begin{aligned}&\|\hat{W} - W_0\|_1 \leq 5\mu_1(W_0)^2\|W_0\|_0\kappa \\ &\quad + 6\sqrt{\|W_0\|_0}\mu_1(W_0) \inf_{A\in\mathcal{A}} \sqrt{\frac{1}{d}\|\omega(A) - \omega(A_{T+1})\|_F^2 + \frac{25}{18}\mu_2(A,W)^2\big(\operatorname{rank}(A)\tau^2 + \|A\|_0\gamma^2\big)}\end{aligned} \tag{9}$$

$$\begin{aligned}\|\hat{A} - A_{T+1}\|_* \leq\ & 5\mu_1(W_0)^2\|W_0\|_0\kappa + (6\sqrt{\operatorname{rank} A_{T+1}} + 5\beta\sqrt{\|A_{T+1}\|_0})\mu_2(A_{T+1}) \\ & \times \inf_{A\in\mathcal{A}} \sqrt{\frac{1}{d}\|\omega(A) - \omega(A_{T+1})\|_F^2 + \frac{25}{18}\mu_2(A,W)^2\big(\operatorname{rank}(A)\tau^2 + \|A\|_0\gamma^2\big)}\,.\end{aligned} \tag{10}$$

[Appendix : Proof of propositions]

A Proofs of the main results

From now on, we use the notation $\|(A,a)\|_F^2 = \|A\|_F^2 + \|a\|_2^2$ and $\langle(A,a),(B,b)\rangle = \langle A,B\rangle + \langle a,b\rangle$ for any $A, B \in \mathbb{R}^{T\times d}$ and $a, b \in \mathbb{R}^d$.

Let us introduce the linear mapping $\Phi : \mathbb{R}^{n\times n} \times \mathbb{R}^{d\times d} \to \mathbb{R}^{T\times d} \times \mathbb{R}^d$ given by

$$\Phi(A,W) = \Big(\frac{1}{\sqrt{T}}\mathbf{X}_{T-1}W, \omega(A) - W^\top\omega(A_T)\Big).$$

Using this mapping, the objective (3) can be written in the following reduced way:

$$\mathcal{L}(A,W) = \frac{1}{d}\Big\|\Big(\frac{1}{\sqrt{T}}\mathbf{X}_T, 0\Big) - \Phi(A,W)\Big\|_F^2 + \gamma\|A\|_1 + \tau\|A\|_* + \kappa\|W\|_1.$$

Recalling that the error writes, for any A and W:

$$\mathcal{E}(A,W)^2 = \frac{1}{d}\|(W - W_0)^\top \omega(A_T) - \omega(A - A_{T+1})\|_F^2 + \frac{1}{dT}\|\mathbf{X}_{T-1}(W - W_0)\|_F^2,$$

we have

$$\mathcal{E}(A,W)^2 = \frac{1}{d}\big\|\Phi(A - A_{T+1}, W - W_0)\|_F^2.$$

Let us introduce also the empirical risk

$$R_n(A,W) = \frac{1}{d}\Big\|\Big(\frac{1}{\sqrt{T}}\mathbf{X}_T, 0\Big) - \Phi(A,W)\Big\|_F^2.$$

The proofs of Theorem 1 and 2 are based on tools developped in [6] and [1]. However, the context considered here is very different from the setting considered in these papers, so our proofs require a different scheme.

A.1 Proof of Theorem 1

First, note that

$$\begin{aligned} &R_n(\hat{A},\hat{W}) - R_n(A,W) \\ &\qquad = \frac{1}{d}\Big(\|\Phi(\hat{A},\hat{W})\|_F^2 - \|\Phi(A,W)\|_F^2 - 2\langle(\frac{1}{\sqrt{T}}\mathbf{X}_T, 0), \Phi(\hat{A} - A, \hat{W} - W)\rangle\Big). \end{aligned}$$

Since

$$\begin{aligned} &\frac{1}{d}\Big(\|\Phi(\hat{A},\hat{W})\|_F^2 - \|\Phi(A,W)\|_F^2\Big) \\ &\qquad = \mathcal{E}(\hat{A},\hat{W})^2 - \mathcal{E}(A,W)^2 + \frac{2}{d}\langle\Phi(\hat{A} - A, \hat{W} - W), \Phi(A_{T+1}, W_0)\rangle, \end{aligned}$$

we have

$$\begin{aligned} &R_n(\hat{A},\hat{W}) - R_n(A,W) \\ &\qquad = \mathcal{E}(\hat{A},\hat{W})^2 - \mathcal{E}(A,W)^2 + \frac{2}{d}\langle\Phi(\hat{A} - A, \hat{W} - W), \Phi(A_{T+1}, W_0) - (\frac{1}{\sqrt{T}}\mathbf{X}_T, 0)\rangle \\ &\qquad = \mathcal{E}(\hat{A},\hat{W})^2 - \mathcal{E}(A,W)^2 + \frac{2}{d}\langle\Phi(\hat{A} - A, \hat{W} - W), (-\frac{1}{\sqrt{T}}\mathbf{N}_T, N_{T+1})\rangle. \end{aligned}$$

The next Lemma will come in handy several times in the proofs.

Lemma 1. *For any $A \in \mathbb{R}^{n\times n}$ and $W \in \mathbb{R}^{d\times d}$ we have*

$$\langle(\frac{1}{\sqrt{T}}\mathbf{N}_T, -N_{T+1}), \Phi(A,W)\rangle = \langle(M, \frac{1}{T}\Xi), (A,W)\rangle = \frac{1}{T}\langle W, \Xi\rangle + \langle A, M\rangle.$$

This Lemma follows from a direct computation, and the proof is thus omitted. This Lemma entails, together with (4), that

$$\begin{aligned} \mathcal{E}(\hat{A},\hat{W})^2 \le \mathcal{E}(A,W)^2 &+ \frac{2}{dT}\langle\hat{W} - W, \Xi\rangle + \frac{2}{d}\langle\hat{A} - A, M\rangle \\ &+ \tau(\|A\|_* - \|\widehat{A}\|_*) + \gamma(\|A\|_1 - \|\widehat{A}\|_1) + \kappa(\|W\|_1 - \|\widehat{W}\|_1). \end{aligned}$$

Now, using Hölder's inequality and the triangle inequality, and introducing $\alpha \in (0,1)$, we obtain

$$\begin{aligned} \mathcal{E}(\hat{A},\hat{W})^2 \le \mathcal{E}(A,W)^2 &+ \Big(\frac{2\alpha}{d}\|M\|_{\mathrm{op}} - \tau\Big)\|\hat{A}\|_* + \Big(\frac{2\alpha}{d}\|M\|_{\mathrm{op}} + \tau\Big)\|A\|_* \\ &+ \Big(\frac{2(1-\alpha)}{d}\|M\|_\infty - \gamma\Big)\|\hat{A}\|_1 + \Big(\frac{2(1-\alpha)}{d}\|M\|_\infty + \gamma\Big)\|A\|_1 \\ &+ \Big(\frac{2}{dT}\|\Xi\|_\infty - \kappa\Big)\|\hat{W}\|_1 + \Big(\frac{2}{dT}\|\Xi\|_\infty + \kappa\Big)\|W\|_1, \end{aligned}$$

which concludes the proof of Theorem 1, using (6). □

A.2 Proof of Theorem 2

Let $A \in \mathbb{R}^{n\times n}$ and $W \in \mathbb{R}^{d\times d}$ be fixed, and let $A = U \operatorname{diag}(\sigma_1, \ldots, \sigma_r)V^\top$ be the SVD of A. Recalling that $\circ$ is the entry-wise product, we have $A = \Theta_A \circ |A| + \Theta_A^\perp \circ A$, where $\Theta_A \in \{0, \pm 1\}^{n\times n}$ is the entry-wise sign matrix of A and $\Theta_A^\perp \in \{0, 1\}^{n\times n}$ is the orthogonal sparsity pattern of A.

The definition (4) of $(\hat{A}, \hat{W})$ is equivalent to the fact that one can find $\hat{G} \in \partial\mathcal{L}(\hat{A}, \hat{W})$ (an element of the subgradient of $\mathcal{L}$ at $(\hat{A}, \hat{W})$) that belongs to the normal cone of $\mathcal{A} \times \mathcal{W}$ at $(\hat{A}, \hat{W})$. This means that for such a $\hat{G}$, and any $A \in \mathcal{A}$ and $W \in \mathcal{W}$, we have

$$\langle \hat{G}, (\hat{A} - A, \hat{W} - W)\rangle \le 0. \tag{11}$$

Any subgradient of the function $g(A) = \tau\|A\|_* + \gamma\|A\|_1$ writes

$$Z = \tau Z_* + \gamma Z_1 = \tau\Big(UV^\top + \mathcal{P}_A^\perp(G_*)\Big) + \gamma\Big(\Theta_A + G_1 \circ \Theta_A^\perp\Big)$$

for some $\|G_*\|_{\mathrm{op}} \le 1$ and $\|G_1\|_\infty \le 1$ (see for instance [7]). So, if $\hat{Z} \in \partial g(\hat{A})$, we have, by monotonicity of the sub-differential, that for any $Z \in \partial g(A)$

$$\langle \hat{Z}, \hat{A} - A\rangle = \langle \hat{Z} - Z, \hat{A} - A\rangle + \langle Z, \hat{A} - A\rangle \ge \langle Z, \hat{A} - A\rangle,$$

and, by duality, we can find Z such that

$$\langle Z, \widehat{A} - A\rangle = \tau\langle UV^\top, \widehat{A} - A\rangle + \tau\|\mathcal{P}_A^\perp(\widehat{A})\|_* + \gamma\langle\Theta_A, \widehat{A} - A\rangle + \gamma\|\Theta_A^\perp \circ \widehat{A}\|_1.$$

By using the same argument with the function $W \mapsto \|W\|_1$ and by computing the gradient of the empirical risk $(A, W) \mapsto R_n(A, W)$, Equation (11) entails that

$$\begin{aligned}
&\frac{2}{d}\langle\Phi(\widehat{A} - A_{T+1}, \widehat{W} - W_0), \Phi(\widehat{A} - A, \widehat{W} - W)\rangle \\
&\qquad \le \frac{2}{d}\langle(\frac{1}{\sqrt{T}}\mathbf{N}_T, -N_{T+1}), \Phi(\widehat{A} - A, \widehat{W} - W)\rangle - \tau\langle UV^\top, \widehat{A} - A\rangle - \tau\|\mathcal{P}_A^\perp(\widehat{A})\|_* \\
&\qquad\quad - \gamma\langle\Theta_A, \widehat{A} - A\rangle - \gamma\|\Theta_A^\perp \circ \widehat{A}\|_1 - \kappa\langle\Theta_W, \widehat{W} - W\rangle - \kappa\|\Theta_W^\perp \circ \widehat{W}\|_1.
\end{aligned} \tag{12}$$

Using Pythagora's theorem, we have

$$\begin{aligned}
&2\langle\Phi(\widehat{A} - A_{T+1}, \hat{W} - W_0), \Phi(\widehat{A} - A, \widehat{W} - W)\rangle \\
&\quad = \|\Phi(\widehat{A} - A_{T+1}, \hat{W} - W_0)\|_2^2 + \|\Phi(\widehat{A} - A, \widehat{W} - W)\|_2^2 - \|\Phi(A - A_{T+1}, W - W_0)\|_2^2.
\end{aligned} \tag{13}$$

It shows that if $\langle\Phi(\widehat{A} - A_{T+1}, W - W_0), \Phi(\widehat{A} - A, \widehat{W} - W)\rangle \le 0$, then Theorem 2 trivially holds. Let us assume that

$$\langle\Phi(\widehat{A} - A_{T+1}, W - W_0), \Phi(\widehat{A} - A, \widehat{W} - W)\rangle > 0. \tag{14}$$

Using Hölder's inequality, we obtain

$$\begin{aligned}
|\langle UV^\top, \hat{A} - A\rangle| &= |\langle UV^\top, \mathcal{P}_A(\hat{A} - A)\rangle| \le \|UV^\top\|_{\mathrm{op}}\|\mathcal{P}_A(\hat{A} - A)\|_* = \|\mathcal{P}_A(\hat{A} - A)\|_*, \\
|\langle\Theta_A, \hat{A} - A\rangle| &= |\langle\Theta_A, \Theta_A \circ (\hat{A} - A)\rangle| \le \|\Theta_A\|_\infty\|\Theta_A \circ (\hat{A} - A)\|_1 = \|\Theta_A \circ (\hat{A} - A)\|_1,
\end{aligned}$$

and the same is done for $|\langle\Theta_W, \hat{W} - W\rangle| \le \|\Theta_W \circ (\hat{W} - W)\|_1$. So, when (14) holds, we obtain by rearranging the terms of (12):

$$\begin{aligned}
\tau\|\mathcal{P}_A^\perp(\widehat{A}-A)\|_* &+ \gamma\|\Theta_A^\perp \circ (\widehat{A} - A)\|_1 + \kappa\|\Theta_W^\perp \circ (\widehat{W} - W)\|_1 \\
&\le \tau\|\mathcal{P}_A(\hat{A} - A)\|_* + \gamma\|\Theta_A \circ (\hat{A} - A)\|_1 + \kappa\|\Theta_W \circ (\hat{W} - W)\|_1 \\
&\quad + \frac{2}{d}\langle(\frac{1}{\sqrt{T}}\mathbf{N}_T, -N_{T+1}), \Phi(\widehat{A} - A, \widehat{W} - W)\rangle.
\end{aligned} \tag{15}$$

Using Lemma 1, together with Hölder's inequality, we have for any $\alpha \in (0, 1)$:

$$\begin{aligned}
\langle(\frac{1}{\sqrt{T}}\mathbf{N}_T, &- N_{T+1}), \Phi(\widehat{A} - A, \widehat{W} - W)\rangle = \langle M, \hat{A} - A\rangle + \frac{1}{T}\langle\Xi, \hat{W} - W\rangle \\
&\le \alpha\|M\|_{\mathrm{op}}\|\mathcal{P}_A(\hat{A} - A)\|_* + \alpha\|M\|_{\mathrm{op}}\|\mathcal{P}_A^\perp(\hat{A} - A)\|_* \\
&\quad + (1 - \alpha)\|M\|_\infty\|\Theta_A \circ (\hat{A} - A)\|_1 + (1 - \alpha)\|M\|_\infty\|\Theta_A^\perp \circ (\hat{A} - A)\|_1 \\
&\quad + \frac{1}{T}\|\Xi\|_\infty(\|\Theta_W \circ (\hat{W} - W)\|_1 + \|\Theta_W^\perp \circ (\hat{W} - W)\|_1)\ .
\end{aligned} \tag{16}$$

Now, using (15) together with (16), we obtain

$$\begin{aligned}&\Big(\tau-\frac{2\alpha}{d}\|M\|_{\rm op}\Big)\|\mathcal{P}_A^{\perp}(\hat{A}-A)\|_*+\Big(\gamma-\frac{2(1-\alpha)}{d}\|M\|_\infty\Big)\|\Theta_A^{\perp}\circ(\hat{A}-A)\|_1\\&\qquad+\Big(\kappa-\frac{2}{dT}\|\Xi\|_\infty\Big)\|\Theta_W^{\perp}\circ(\hat{W}-W)\|_1\\&\leq\Big(\tau+\frac{2\alpha}{d}\|M\|_{\rm op}\Big)\|\mathcal{P}_A(\hat{A}-A)\|_*+\Big(\gamma+\frac{2(1-\alpha)}{d}\|M\|_\infty\Big)\|\Theta_A\circ(\hat{A}-A)\|_1\\&\qquad+\Big(\kappa+\frac{2}{dT}\|\Xi\|_\infty\Big)\|\Theta_W\circ(\hat{W}-W)\|_1\end{aligned}$$

which proves, using (7), that

$$\tau\|\mathcal{P}_A^{\perp}(\hat{A}-A)\|_*+\gamma\|\Theta_A^{\perp}\circ(\hat{A}-A)\|_1\leq 5\tau\|\mathcal{P}_A(\hat{A}-A)\|_*+5\gamma\|\Theta_A\circ(\hat{A}-A)\|_1.$$

This proves that $\hat{A}-A\in\mathcal{C}_2(A,5,\gamma/\tau)$. In the same way, using (15) with $A=\hat{A}$ together with (16), we obtain that $\hat{W}-W\in\mathcal{C}_1(W,5)$.

Now, using together (12), (13) and (16) , and the fact that the Cauchy-Schwarz inequality entails

$$\begin{aligned}&\|\mathcal{P}_A(\hat{A}-A)\|_*\leq\sqrt{\operatorname{rank}A}\|\mathcal{P}_A(\hat{A}-A)\|_F,\quad|\langle UV^\top,\hat{A}-A\rangle|\leq\sqrt{\operatorname{rank}A}\|\mathcal{P}_A(\hat{A}-A)\|_F,\\&\|\Theta_A\circ(\hat{A}-A)\|_1\leq\sqrt{\|A\|_0}\|\Theta_A\circ(\hat{A}-A)\|_F,\quad|\langle\Theta_A,\hat{A}-A\rangle|\leq\sqrt{\|A\|_0}\|\Theta_A\circ(\hat{A}-A)\|_F\ .\end{aligned}$$

and similarly for $\hat{W}-W$, we arrive at

$$\begin{aligned}&\|\Phi(\widehat{A}-A_{T+1},\hat{W}-W_0)\|_2^2+\|\Phi(\widehat{A}-A,\widehat{W}-W)\|_2^2-\|\Phi(A-A_{T+1},W-W_0)\|_2^2\\&\quad\leq\Big(\frac{2\alpha}{d}\|M\|_{\rm op}+\tau\Big)\sqrt{\operatorname{rank}A}\|\mathcal{P}_A(\hat{A}-A)\|_F+\Big(\frac{2\alpha}{d}\|M\|_{\rm op}-\tau\Big)\|\mathcal{P}_A^{\perp}(\hat{A}-A)\|_*\\&\qquad+\Big(\frac{2\alpha}{d}\|M\|_\infty+\gamma\Big)\sqrt{\|A\|_0}\|\Theta_A\circ(\hat{A}-A)\|_F+\Big(\frac{2\alpha}{d}\|M\|_\infty-\gamma\Big)\|\Theta_A^{\perp}\circ(\hat{A}-A)\|_1\\&\qquad+\Big(\frac{2\alpha}{dT}\|\Xi\|_\infty+\kappa\Big)\sqrt{\|W\|_0}\|\Theta_W\circ(\hat{W}-W)\|_F+\Big(\frac{2\alpha}{dT}\|\Xi\|_\infty-\kappa\Big)\|\Theta_W^{\perp}\circ(\hat{W}-W)\|_1,\end{aligned}$$

which leads, using (7), to

$$\begin{aligned}&\frac{1}{d}\|\Phi(\widehat{A}-A_{T+1},\hat{W}-W_0)\|_2^2+\frac{1}{d}\|\Phi(\widehat{A}-A,\widehat{W}-W)\|_2^2-\frac{1}{d}\|\Phi(A-A_{T+1},W-W_0)\|_2^2\\&\quad\leq\frac{5\tau}{3}\sqrt{\operatorname{rank}A}\|\mathcal{P}_A(\hat{A}-A)\|_F+\frac{5\gamma}{3}\sqrt{\|A\|_0}\|\Theta_A\circ(\hat{A}-A)\|_F+\frac{5\kappa}{3}\sqrt{\|W\|_0}\|\Theta_W\circ(\hat{W}-W)\|_F.\end{aligned}$$

Since $\hat{A}-A\in\mathcal{C}_2(A,5,\gamma/\tau)$ and $\hat{W}-W\in\mathcal{C}_1(W,5)$, we obtain using Assumption 1 and $ab\leq(a^2+b^2)/2$:

$$\begin{aligned}&\frac{1}{d}\|\Phi(\widehat{A}-A_{T+1},\hat{W}-W_0)\|_2^2+\frac{1}{d}\|\Phi(\widehat{A}-A,\widehat{W}-W)\|_2^2\\&\quad\leq\frac{1}{d}\|\Phi(A-A_{T+1},W-W_0)\|_2^2+\frac{25}{18}\mu_2(A,W)^2\big(\operatorname{rank}(A)\tau^2+\|A\|_0\gamma^2\big)\\&\qquad+\frac{25}{36}\mu_1(W)^2\|W\|_0\kappa^2+\frac{1}{d}\|\Phi(\widehat{A}-A,\widehat{W}-W)\|_2^2,\end{aligned}$$

which concludes the proof of Theorem 2. □

A.3 Proof of Theorem 4

For the proof of (8), we simply use the fact that $\frac{1}{dT}\|\mathbf{X}_{T-1}(\hat{W}-W_0)\|_F^2\leq\mathcal{E}(\hat{A},\hat{W})^2$ and use Theorem 3. Then we take $W=W_0$ in the infimum over A,W.

For (9), we use the fact that since $\hat{W}-W_0\in\mathcal{C}_1(W_0,10)$, we have (see the Proof of Theorem 2),

$$\begin{aligned}\|\hat{W}-W_0\|_1&\leq 6\sqrt{\|W_0\|_0}\|\Theta_W\circ(\hat{W}-W_0)\|_F\\&\leq 6\sqrt{\|W_0\|_0}\|\mathbf{X}_{T-1}(\hat{W}-W_0)\|_F/\sqrt{dT}\\&\leq 6\sqrt{\|W_0\|_0}\mathcal{E}(\hat{A},\hat{W}),\end{aligned}$$

and then use again Theorem 3. The proof of (10) follows exactly the same scheme. □

A.4 Concentration inequalities for the noise processes

The control of the noise terms M and Ξ is based on recent developments on concentration inequalities for random matrices, see for instance [9]. Moreover, the assumption on the dynamics of the features's noise vector $(N_t)_{t \geq 0}$ is quite general, since we only assumed that this process is a martingale increment. Therefore, our control of the noise Ξ rely in particular on martingale theory.

Proposition 1. *Under Assumption 1, the following inequalities hold for any $x > 0$. We have*

$$\Big\| \frac{1}{d} \sum_{j=1}^{d} (N_{T+1})_j \Omega_j \Big\|_{\mathrm{op}} \leq \sigma v_{\Omega,\mathrm{op}} \sqrt{\frac{2(x + \log(2n))}{d}} \tag{17}$$

with a probability larger than $1 - e^{-x}$. We have

$$\Big\| \frac{1}{d} \sum_{j=1}^{d} (N_{T+1})_j \Omega_j \Big\|_{\infty} \leq \sigma v_{\Omega,\infty} \sqrt{\frac{2(x + 2\log n)}{d}} \tag{18}$$

with a probability larger than $1 - 2e^{-x}$, and finally

$$\Big\| \frac{1}{T+1} \sum_{t=1}^{T+1} \omega(A_{t-1}) N_t^\top \Big\|_{\infty} \leq \sigma \sigma_{\omega} \sqrt{\frac{2e(x + 2\log d + \ell_T)}{T+1}} \tag{19}$$

with a probability larger than $1 - 14e^{-x}$, where

$$\ell_T = 2 \max_{j=1,\ldots,d} \log\log \bigg(\frac{\sum_{t=1}^{T+1} \omega_j(A_{t-1})^2}{T+1} \vee \frac{T+1}{\sum_{t=1}^{T+1} \omega_j(A_{t-1})^2} \vee e \bigg).$$

Proof. For the proofs of Inequalities (18) and (19), we use the fact that $(N_{T+1})_1, \ldots, (N_{T+1})_d$ are independent (scalar) subgaussian random variables.

From Assumption 1, we have for any $n \times n$ deterministic self-adjoint matrices X_j that $\mathbb{E}[\exp(\lambda(N_{T+1})_j X_j)] \preceq \exp(\sigma^2 \lambda^2 X_j^2/2)$, where $\preceq$ stands for the semidefinite order on self-adjoint matrices. Using Corollary 3.7 from [9], this leads for any $x > 0$ to

$$\mathbb{P}\Big[\lambda_{\max}\Big(\sum_{j=1}^{d} (N_{T+1})_j X_j \Big) \geq x \Big] \leq n \exp\Big(-\frac{x^2}{2v^2} \Big), \quad \text{where } v^2 = \sigma^2 \Big\| \sum_{j=1}^{d} X_j^2 \Big\|_{\mathrm{op}}. \tag{20}$$

Then, following [9], we consider the dilation operator $\mathcal{L} : \mathbb{R}^{n \times n} \to \mathbb{R}^{2n \times 2n}$ given by

$$\mathcal{L}(\Omega) = \begin{pmatrix} 0 & \Omega \\ \Omega^* & 0 \end{pmatrix}.$$

We have

$$\Big\| \sum_{j=1}^{d} (N_{T+1})_j \Omega_j \Big\|_{\mathrm{op}} = \lambda_{\max}\Big(\mathcal{L}\Big(\sum_{j=1}^{d} (N_{T+1})_j \Omega_j \Big) \Big) = \lambda_{\max}\Big(\sum_{j=1}^{d} (N_{T+1})_j \mathcal{L}(\Omega_j) \Big)$$

and an easy computation gives

$$\Big\| \sum_{j=1}^{d} \mathcal{L}(\Omega_j)^2 \Big\|_{\mathrm{op}} = \Big\| \sum_{j=1}^{d} \Omega_j^\top \Omega_j \Big\|_{\mathrm{op}} \vee \Big\| \sum_{j=1}^{d} \Omega_j \Omega_j^\top \Big\|_{\mathrm{op}}.$$

So, using (21) with the self-adjoint $X_j = \mathcal{L}(\Omega_j)$ gives

$$\mathbb{P}\Big[\Big\| \sum_{j=1}^{d} (N_{T+1})_j \Omega_j \Big\|_{\mathrm{op}} \geq x \Big] \leq 2n \exp\Big(-\frac{x^2}{2v^2} \Big) \text{ where } v^2 = \sigma^2 \Big\| \sum_{j=1}^{d} \Omega_j^\top \Omega_j \Big\|_{\mathrm{op}} \vee \Big\| \sum_{j=1}^{d} \Omega_j \Omega_j^\top \Big\|_{\mathrm{op}},$$

which leads easily to (18).

Inequality (19) comes from the following standard bound on the sum of independent sub-gaussian random variables:

$$\mathbb{P}\Big[\Big|\frac{1}{d}\sum_{j=1}^{d}(N_{T+1})_j(\Omega_j)_{k,l}\Big| \geq x\Big] \leq 2\exp\Big(-\frac{x^2}{2\sigma^2(\Omega_j)_{k,l}^2}\Big)$$

together with an union bound on $1 \leq k, l \leq n$.

Inequality (20) is based on a classical martingale exponential argument together with a peeling argument. We denote by $\omega_j(A_t)$ the coordinates of $\omega(A_t) \in \mathbb{R}^d$ and by $N_{t,k}$ those of N_t, so that

$$\Big(\sum_{t=1}^{T+1}\omega(A_{t-1})N_t^\top\Big)_{j,k} = \sum_{t=1}^{T+1}\omega_j(A_{t-1})N_{t,k}.$$

We fix j, k and denote for short $\varepsilon_t = N_{t,k}$ and $x_t = \omega_j(A_t)$. Since $\mathbb{E}[\exp(\lambda\varepsilon_t)|\mathcal{F}_{t-1}] \leq e^{\sigma^2\lambda^2/2}$ for any $\lambda \in \mathbb{R}$, we obtain by a recursive conditioning with respect to $\mathcal{F}_{T-1}, \mathcal{F}_{T-2}, \dots, \mathcal{F}_0$, that

$$\mathbb{E}\Big[\exp\Big(\theta\sum_{t=1}^{T+1}\varepsilon_t x_{t-1} - \frac{\sigma^2\theta^2}{2}\sum_{t=1}^{T+1}x_{t-1}^2\Big)\Big] \leq 1.$$

Hence, using Markov's inequality, we obtain for any $v > 0$:

$$\mathbb{P}\Big[\sum_{t=1}^{T+1}\varepsilon_t x_{t-1} \geq x, \sum_{t=1}^{T+1}x_{t-1}^2 \leq v\Big] \leq \inf_{\theta>0}\exp(-\theta x + \sigma^2\theta^2 v/2) = \exp\Big(-\frac{x^2}{2\sigma^2 v}\Big),$$

that we rewrite in the following way:

$$\mathbb{P}\Big[\sum_{t=1}^{T+1}\varepsilon_t x_{t-1} \geq \sigma\sqrt{2vx}, \sum_{t=1}^{T+1}x_{t-1}^2 \leq v\Big] \leq e^{-x}.$$

Let us denote for short $V_T = \sum_{t=1}^{T+1}x_{t-1}^2$ and $S_T = \sum_{t=1}^{T+1}\varepsilon_t x_{t-1}$. We want to replace v by V_T from the previous deviation inequality, and to remove the event $\{V_T \leq v\}$. To do so, we use a peeling argument. We take $v = T+1$ and introduce $v_k = ve^k$ so that the event $\{V_T > v\}$ is decomposed into the union of the disjoint sets $\{v_k < V_T \leq v_{k+1}\}$. We introduce also $\ell_T = 2\log\log\Big(\frac{\sum_{t=1}^{T+1}x_{t-1}^2}{T+1} \vee \frac{T+1}{\sum_{t=1}^{T+1}x_{t-1}^2} \vee e\Big)$.

This leads to

$$\begin{aligned}\mathbb{P}\Big[S_T \geq \sigma\sqrt{2eV_T(x+\ell_T)}, V_T > v\Big] &= \sum_{k\geq 0}\mathbb{P}\big[S_T \geq \sigma\sqrt{2eV_T(x+\ell_T)}, v_k < V_T \leq v_{k+1}\Big] \\ &= \sum_{k\geq 0}\mathbb{P}\Big[S_T \geq \sigma\sqrt{2v_{k+1}(x + 2\log\log(e^k \vee e))}, v_k < V_T \leq v_{k+1}\Big] \\ &\leq e^{-x}(1 + \sum_{k\geq 1}k^{-2}) \leq 3.47e^{-x}.\end{aligned}$$

On $\{V_T \leq v\}$ the proof is the same: we decompose onto the disjoint sets $\{v_{k+1} < V_T \leq v_k\}$ where this time $v_k = ve^{-k}$, and we arrive at

$$\mathbb{P}\Big[S_T \geq \sigma\sqrt{2eV_T(x+\ell_T)}, V_T \leq v\Big] \leq 3.47e^{-x}.$$

This leads to

$$\mathbb{P}\Bigg[\sum_{t=1}^{T+1}\omega_j(A_{t-1})N_{t,k} \geq \sigma\Big(2e\sum_{t=1}^{T+1}\omega_j(A_{t-1})^2(x+\ell_{T,j})\Big)^{1/2}\Bigg] \leq 7e^{-x}$$

for any $1 \leq j, k \leq d$, where we introduced

$$\ell_{T,j} = 2\log\log\Big(\frac{\sum_{t=1}^{T+1}\omega_j(A_{t-1})^2}{T+1} \vee \frac{T+1}{\sum_{t=1}^{T+1}\omega_j(A_{t-1})^2} \vee e\Big).$$

The conclusion follows from an union bound on $1 \leq j, k \leq d$. This concludes the proof of Proposition 1. □

References

[1] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 37, 2009.

[2] Candès E. and Tao T. Decoding by linear programming. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2005.

[3] S. Gaiffas and G. Lecue. Sharp oracle inequalities for high-dimensional matrix prediction. *Information Theory, IEEE Transactions on*, 57(10):6942 –6957, oct. 2011.

[4] V. Koltchinskii. The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15(3):799–828, 2009.

[5] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst. Henri Poincaré Probab. Stat.*, 45(1):7–57, 2009.

[6] V. Koltchinskii, K. Lounici, and A. Tsybakov. Nuclear norm penalization and optimal rates for noisy matrix completion. *Annals of Statistics*, 2011.

[7] A. S. Lewis. The convex analysis of unitarily invariant matrix functions. *J. Convex Anal.*, 2(1-2):173–183, 1995.

[8] E. Richard, P.-A. Savalle, and N. Vayatis. Estimation of simultaneously sparse and low-rank matrices. In *Proceeding of 29th Annual International Conference on Machine Learning*, 2012.

[9] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *ArXiv e-prints*, April 2010.

[10] S. A. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3:1360–1392, 2009.