Supplementary Materials:
Online Debiasing for Adaptively Collected High-dimensional Data with Applications to Time Series Analysis

Yash Deshpande

Institute for Data, Systems and Society, Massachusetts Institute of Technology

Adel Javanmard

Data Sciences and Operations Department, University of Southern California

Mohammad Mehrabi

Data Sciences and Operations Department, University of Southern California

A Sparse inverse covariance

In Section 3.1 (Figure 2) we provided a numerical example wherein the offline debiasing does not admit an asymptotically normal distribution. As we see from the heat map in Figure 4b, the precision matrix Ω has $\sim 20\%$ non-negligible entries per row. The goal of this section is to show that when Ω is sufficiently sparse, the offline debiased estimator has an asymptotically normal distribution and can be used for valid inference on model parameters.

The idea is to show that the decorrelating matrix M is sufficiently close to the precision matrix Ω. Since Ω is deterministic, this helps with controlling the statistical dependence between M and ε. Formally, starting from the decomposition (4) we write

$$\begin{aligned}\widehat{\theta}^{\mathsf{off}} &= \theta_0 + (I - M\widehat{\Sigma})(\widehat{\theta}^{\mathsf{L}} - \theta_0) + \frac{1}{n} M X^{\mathsf{T}} \varepsilon \\ &= \theta_0 + (I - M\widehat{\Sigma})(\widehat{\theta}^{\mathsf{L}} - \theta_0) + \frac{1}{n}(M - \Omega) X^{\mathsf{T}} \varepsilon + \frac{1}{n} \Omega X^{\mathsf{T}} \varepsilon \,, \qquad (47)\end{aligned}$$

where we recall that $\widehat{\Sigma}$ is the empirical covariance of all the covariate vectors (episodes

$E_0, \dots, E_{K-1}$). Therefore, we can write

$$\begin{aligned} \sqrt{n}(\widehat{\theta}^{\mathsf{off}} - \theta_0) &= \Delta_1 + \Delta_2 + \frac{1}{\sqrt{n}}\Omega X^{\mathsf{T}}\varepsilon\,, \\ \Delta_1 &= \sqrt{n}(I - M\widehat{\Sigma})(\widehat{\theta}^{\mathsf{L}} - \theta_0)\,, \\ \Delta_2 &= \frac{1}{\sqrt{n}}(M - \Omega)X^{\mathsf{T}}\varepsilon\,. \end{aligned} \tag{48}$$

The term $\Omega X^{\mathsf{T}}\varepsilon/\sqrt{n}$ is gaussian with $O(1)$ variance at each coordinate. For bias term Δ_1, we show that $\Delta_1 = O(s_0(\log p)/\sqrt{n})$ by controlling $|I - M\widehat{\Sigma}|$. To bound the bias term Δ_2 we write

$$\|\Delta_2\|_\infty \le \frac{1}{\sqrt{n}}\|M - \Omega\|_1 \|X^{\mathsf{T}}\varepsilon\|_\infty\,, \tag{49}$$

where $\|M - \Omega\|_1$ denotes the $\ell_1 - \ell_1$ norm of $M - \Omega$ (the maximum ℓ_1 norm of its columns). By using [2, Proposition 3.2], we have $\|X^{\mathsf{T}}\varepsilon\|_\infty/\sqrt{n} = O_P(\sqrt{\log(dp)})$. Therefore, to bound Δ_2 we need to control $\|M - \Omega\|_1$. We provide such bound in our next lemma, under the sparsity assumption on the rows of Ω.

Define

$$s_\Omega \equiv \max_{i \in [dp]} \Big| j \in [dp] : \quad \Omega_{i,j} \neq 0 \Big|\,,$$

the maximum sparsity of rows of Ω. In addition, let the (offline) decorrelating vectors m_a be defined as follows, for $a \in [dp]$:

$$m_a \in \arg\min_{m \in \mathbb{R}^{dp}} \quad \frac{1}{2} m^{\mathsf{T}}\widehat{\Sigma}m - \langle m, e_a \rangle + \mu\|m\|_1\,. \tag{50}$$

Lemma A.1. *Consider the decorrelating vectors m_a, $a \in [dp]$, given by optimization (50) with $\mu = 2\tau\sqrt{\frac{\log(dp)}{n}}$. Then, for some proper constant $c > 0$ and the sample size condition $n \ge 32\alpha(\omega^2 \vee 1)s_\Omega \log(dp)$, the following happens with probability at least $1 - \exp(-c\log(dp^2)) - \exp(-cn(1 \wedge \omega^{-2}))$:*

$$\max_{i \in [dp]} \|m_a - \Omega e_a\|_1 \le \frac{192\tau}{\alpha} s_\Omega \sqrt{\frac{\log(dp)}{n}},$$

where $\alpha$and ω are defined in Proposition F.4.

The proof of Lemma A.1 is deferred to Section G.2.

By employing this lemma, if Ω is sufficiently sparse, that is $s_\Omega = o(\sqrt{n}/\log(dp))$, then the bias term $\|\Delta_2\|_\infty$ also vanishes asymptotically and the (offline) debiased estimator $\widehat{\theta}^{\mathsf{off}}$ admits an unbiased normal distribution. We formalize such distributional characterization in the next theorem.

Theorem A.2. *Consider the* $\mathsf{VAR}(d)$ *model* (5) *for time series and let* $\widehat{\theta}^{\mathsf{off}}$ *be the (offline) debiased estimator* (3)*, with the decorrelating matrix* $M = (m_1, \dots, m_{dp})^{\mathsf{T}} \in \mathbb{R}^{dp\times dp}$ *constructed as in* (50)*, with* $\mu = 2\tau\sqrt{\log(dp)/n}$*. Also, let* $\lambda = \lambda_0\sqrt{\log(dp)/n}$ *be the regularization parameter in the Lasso estimator* $\widehat{\theta}^{\mathsf{L}}$*, with* τ, λ_0 *large enough constants.*

Suppose that $s_0 = o(\sqrt{n}/\log(dp))$ *and* $s_\Omega = o(\sqrt{n}/\log(dp))$*, then the following holds true for any fixed sequence of integers* $a(n) \in [dp]$*: For all* $x \in \mathbb{R}$*, we have*

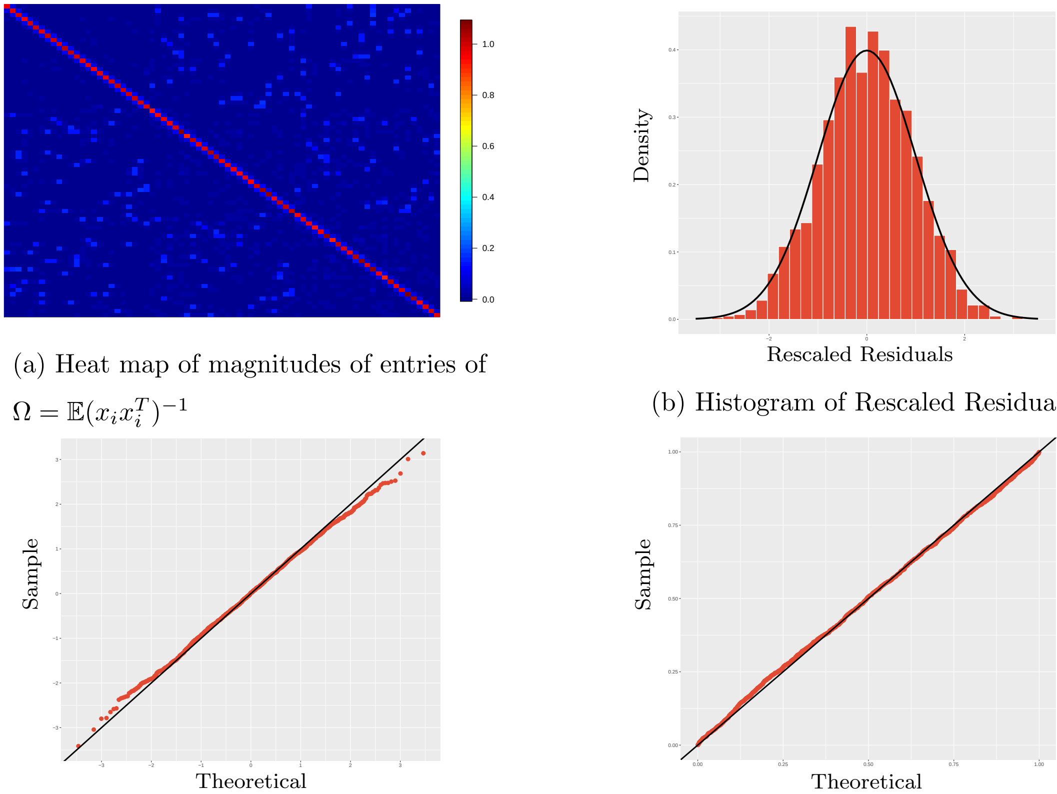
$$\lim_{n\to\infty} \sup_{\|\theta_0\|_0 \le s_0} \left| \mathbb{P}\left\{ \frac{\sqrt{n}(\widehat{\theta}^{\mathsf{off}}_a - \theta_{0,a})}{\sqrt{V_{n,a}}} \le x \right\} - \Phi(x) \right| = 0\,, \tag{51}$$

where $V_{n,a} \equiv \sigma^2(M\widehat{\Sigma}M^{\mathsf{T}})_{a,a}$*.*

We refer to Section G.3 for the proof of Theorem A.2.

Numerical example. Consider a $\mathsf{VAR}(d)$ model with parameters $p = 25, d = 3, T = 70$, and Gaussian noise terms with covariance matrix Σ_ζ satisfying $\Sigma_\zeta(i,j) = \rho^{|i-j|}$ for $\rho = 0.1$. Let A_i matrices have entries generated independently from $b \cdot \text{Bern}(q) \cdot \text{Unif}(\{+1,-1\})$ formula with parameters $b = 0.15$, $q = 0.05$. Figure 7a shows the magnitudes of the entries of the precision matrix $\Omega = \mathbb{E}(x_i x_i^T)^{-1}$; as we see Ω is sparse. Figures 7b, 7c, and 7d demonstrate normality of the rescaled residuals of the offline debiased estimator built by decorrelating matrix M with rows coming from optimization described in (50).

After this paper was posted, we learned of simultaneous work (an updated version of [1]) that also studies the performance of the (offline) debiased estimator for time series with *sparse* precision matrix. We would like to highlight some of the differences between our discussion in Section A and that paper: 1) [1] considers decorrelating matrix M



(a) Heat map of magnitudes of entries of $\Omega = \mathbb{E}(x_i x_i^T)^{-1}$

(b) Histogram of Rescaled Residuals

(c) QQ plot of Rescaled Residuals

(d) PP plot of Rescaled Residuals

Figure 7: A Simple example of a $\mathsf{VAR}(d)$ process with parameters $p = 25, d = 3, T = 70$, and noise term covariance matrix Σ_ζ s.t $\Sigma_\zeta(i,j) = \rho^{|i-j|}$ with $\rho = 0.1$. A_i matrices have independent elements coming from $b \cdot \text{Bern}(q).\text{Unif}(\{+1,-1\})$ formula with $b = 0.15, q = 0.05$. Normality of rescaled residuals (figures 7b, 7c, and 7d) validates the successful performance of offline debiasing estimator under sparsity of precision matrix Ω (figure 7a) as we discussed in theorem A.2.

constructed by an optimization of form (14), using the entire sample covariance $\widehat{\Sigma}^{(K)}$, while we work with the Lagrangian equivalent (50). 2) [1] considers $\mathsf{VAR}(1)$ model, while we work with $\mathsf{VAR}(d)$ models. 3) [1] assumes a stronger notion of sparsity, viz. the sparsity of the entire precision matrix as well as the transition matrix to scale as $o(\sqrt{n}/\log p)$. Our results only require the *row-wise sparsity* of the precision matrix to scale as $o(\sqrt{n}/\log p)$, cf. Theorem A.2.

We would like to also discuss the related work [42] which studies the problem of statistical inference on the coefficients of autoregressive models. This work follows the same idea of debiasing, but uses an offline debiasing, in our terminology. Specifically, they adopt the framework of [29] and propose a high dimensional test statistic based on score function, called the decorrelated score function. It shows that under proper sparsity assumptions on the model coefficients and the precision matrix, their bias-corrected estimator achieves asymptotical Gaussianity. The work [42] considers the simultaneous (group) inference on a fixed number of coefficients and in the univariate case (testing an individual coefficient) their sparsity assumption becomes equivalent to the assumptions of Theorem A.2 on s_0 and s_Ω. However, the decorrelated score matrix in [42] is constructed by Lasso or Dantzig selector which is different from our proposal in (50). Let us reiterate that the method of (50) is an offline debiasing approach and hence, as discussed in Section 3.1 (Figure 2), may in general fail in providing valid statistical tests for time series. Apart from this point, in order to further delineate the differences of online debiasing with offline bias-correction methods, such as [42] , we consider a set of numerical examples with sufficiently sparse precision matrix, where the test of [42] is also guaranteed to have valid statistical performance. We use the software package of [42] for an implementation of their method, and consider configurations that are inspired by the package built-in example. We consider $\mathsf{VAR}(1)$ model where time-series samples are generated from $X_t = AX_t + \varepsilon_t$, with $\varepsilon_t \sim \text{Unif}[0,1]$. The generative matrix A is of form $A = \text{diag}(A_0, A_0, \dots, A_0)$ with A_0 a symmetric 2×2 matrix:

$$A_0 = \begin{bmatrix} q_2 & q_1 \\ q_1 & q_2 \end{bmatrix} . \tag{52}$$

Each configuration is determined by the number of samples T, the primary generative matrix coefficients q_1, q_2, and the matrix dimension p. For each configuration, we test all coordinates of A and report the two measures true positive rate (TPR) and false positive rate (FPR), along with the running time of the algorithms. Note that the [42] method outputs two test statistics U, R, which are constructed in almost similar ways (except the last step) and so has the same running time. Table 1 demonstrates the statistical performance of our online debiasing and the bias-correction method of [42]. The reported values are averaged out over 10 independent experiments, and the running times are in seconds.

The first interesting observation is about the statistical power, where it can be observed that both online debiasing and [42] have comparable performance. Note that on the one hand, the approach of [42] uses the Lasso or Dantzig selector to construct decorrelated score and hence searches over the space of sparse matrices. However, the online debiasing searches over the larger space of approximately sparse matrices (cf. optimization (14)). This factor plays in favor of online debiasing to potentially have higher power. On the other hand, the online debiasing framework uses samples gradually and the decorrelating matrices $M^{(\ell)}$ are constructed from subsets of samples. This factor plays in favor of offline debiasing methods that use the entire samples in constructing the decorrelated score matrix. The interplay between these two factors leads the two methods to have comparative statistical power.

The other interesting observation is about the running time of the two methods, where it can be observed that online debiasing enjoys a significantly lower running time. In fact, the online debiasing method with T samples, has $\log T$ number of episodes, and for each one a $dp \times dp$ debiasing matrix is constructed by solving dp optimization problems– because of row by row construction of each matrix. A delicate point we would like to make is that in online debiasing, at each round, we focus on one row– say i– of the generative matrices $A^{(1)}, \dots, A^{(d)}$ (stacked together as in equation (11)), and construct the decorrelating matrices $M^{(\ell)}$. However, these decorrelating matrices only depend on the covariate matrix (X in (11)) and so do not change across different

Configuration	**Online Debiasing**			**U test**			**R test**		
(p, T, q_1, q_2)	TPR	FPR	Time	TPR	FPR	Time	TPR	FPR	Time
$(6, 400, 1/15, 1/15)$	0.35	0.046	12.52	0.325	0.067	241.90	0.325	0.067	241.90
$(6, 600, 1/15, 1/15)$	0.375	0.07	14.53	0.36	0.05	376.27	0.358	0.05	376.27
$(6, 1000, 1/15, 1/15)$	0.65	0.062	16.54	0.63	0.064	660.32	0.63	0.064	660.32
$(8, 300, 1/2, 1/4)$	0.993	0.02	5.63	1	0.037	348.71	1	0.031	348.71
$(8, 300, 1/4, 1/8)$	0.762	0.025	17.60	0.793	0.031	335	0.793	0.031	335
$(8, 300, 1/8, 1/16)$	0.35	0.035	5.62	0.493	0.043	361.49	0.493	0.043	361.49

Table 1: Overall performance of online debiasing (test (43)) and the U-test and R-test proposed by [42] for testing the entries of the generative matrix A for a VAR(1) model. We consider $A = \text{diag}(A_0, \dots, A_0)$ with A_0 given by (52). For each configuration, we report the true positive rate (TPR), false positive rate (FPR) and the running time for each test. The reported numbers are averaged out over10 independent realizations of each configuration. The running times are in seconds.

rounds. That said, one needs to compute them once for all rows $i \in [p]$. The approach of [42], on the other hand, constructs a separate score vector for each entry of the generative matrices $A^{(1)}, \dots, A^{(d)}$ which is computationally much more demanding.

B Estimating noise variance for VAR model

Define $\widetilde{V}_{n,a} \equiv \frac{1}{n}\sum_{\ell=1}^{K-1}\sum_{t\in E_\ell}\langle m_a^\ell, x_t\rangle^2$. Note that $V_{n,a} = \sigma^2 \widetilde{V}_{n,a}$ and calculating $\widetilde{V}_{n,a}$ does not require the knowledge of σ^2. We define

$$z_a \equiv \sqrt{\frac{n}{\widetilde{V}_{n,a}}}\widehat{\theta}^{\mathsf{on}}_a \,.$$

Using the distributional characterization of the online debiased estimator $\widehat{\theta}^{\mathsf{on}}$, and by a very similar argument in Theorem 3.8, we know that for $a \notin \text{supp}(\theta_0)$, $\theta_{0,a} = 0$ and so z_a is asymptotically zero mean Gaussian of variance σ^2. This suggests to use the following MAD (median absolute deviation) to estimate σ^2.

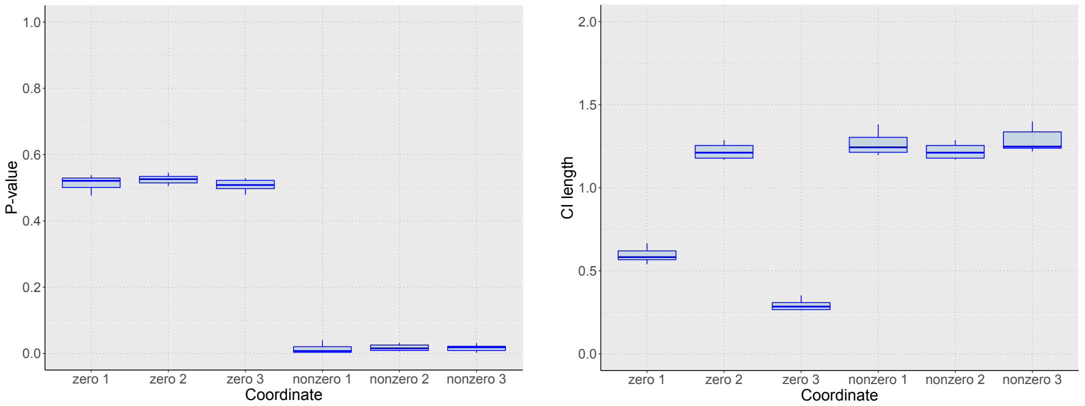
We let $|z|$ be the vector of absolute values of z, i.e., $|z| = (|z_1|, |z_2|, \dots, |z_{dp}|)$. Denote by $|z|_{(a)}$ its a-th smallest entry, i.e., $|z|_{(1)} \le |z|_{(2)} \le \dots \le |z|_{dp}$. We then estimate σ using the MAD estimator

$$\widehat{\sigma} = \frac{|z|_{(dp/2)}}{\Phi^{-1}(3/4)} \,. \tag{53}$$

A similar variance estimator has been proposed by [27] in the context of approximate message passing and in [20] for (offline) debiased estimator. The main idea here is that the MAD estimator is robust to outliers and hence including entries z_a with $a \in \text{supp}(\theta_0)$ have negligible asymptotic contribution to the estimate $\widehat{\sigma}$, given that $s_0 = o(p)$.

C Robustness to the episode growth rate

We follow the guideline in Section 3.1 to choose the batch sizes where the length of episodes grow exponentially; namely $r_\ell = \beta^\ell$, for a constant $\beta > 1$, and $\ell \ge 1$. As it was proved in Theorem 3.4, for any constant $\beta > 1$, the online debiased estimator is asymptotically unbiased assuming $s_0 \log(dp)/\sqrt{n} = o(1)$, and results in valid statistical inference (controlling type I error in the context of hypothesis testing and producing confidence intervals with the target coverage). In this section, we investigate the robustness of these outputs (p-values and confidence intervals) with respect to the choice of tuning parameter β. To this end, we consider the $\mathsf{VAR}(1)$ time-series data setup (5) with problem dimension $p = 20$, and the noise covariance $\Sigma_\zeta(i,j) = 0.1^{|i-j|}$. The entries of the time series generative matrix A are chosen i.i.d. from a Bernoulli distribution with success probability 0.01, and then multiplied by a number chosen uniformly from the set $\{-2, +2\}$, i.e., $A_{ij} \sim \text{Bern}(0.01) \cdot \text{Unif}(\{-2, +2\})$. Fixing the matrix A, we generate $T = 200$ samples $X_{1:200}$ and run the online debiasing with the tuning parameter β picked from a grid of equidistant 11 elements over the interval $[2, 4]$, i.e., $\beta \in \{2, 2.2, 2.4, ..., 3.8, 4\}$. We compute the average length of confidence intervals and p-values across 100 experiments (independent realizations of $X_{1:200}$). For each coordinate, we will end up with 11×2 numbers. Plots 8a and 8b respectively



(a) Boxplot of p-values

(b) Boxplot of confidence interval lengths

Figure 8: Boxplots of p-values and CI lengths for 3 zero and 3 nonzero coordinates of autoregressive matrix A, for 11 choices of episode growth rate $\beta \in \{2, 2.2, ..., 3.8, 4\}$. It can be seen that the variation in each box is small. This implies the robustness of online debiasing framework with respect to the choice of episode growth rate β.

	zero1	zero 2	zero 3	nonzero1	nonzero 2	nonzero 3
CV of p-value	0.0406	0.0248	0.0319	0.928	0.545	0.610
CV of CI length	0.0741	0.0366	0.107	0.047	0.036	0.049

Table 2: Coefficient of variation (CV) for the p-values and the confidence interval lengths for the six selected entries of the A matrix.

denote the boxplots for the average length of 80%-confidence intervals and the average p-values corresponding to the six selected entries of A, three of which are truly zero and the other three are nonzero. As we see the outputs of the online debiasing approach is relatively robust against the choice of the episode growth rate β.

Recalling the coefficient of variation (CV) as a measure of variability, defined as the ratio of the standard deviation to the mean, in Table 2 we give the coefficient of variation for the p-values and length of 80%-confidence intervals for the six coordinates across the 11 choices of β. The small CVs indicate the robustness of p-values and confidence intervals to the specific choice of β.

D Implementation and extensions

D.1 Iterative schemes to implement online debiasing

The online debiased estimator (15) involves the decorrelating matrices $M^{(\ell)}$, whose rows $(m_a^\ell)_{a\in[dp]}$ are constructed by the optimization (14). For the sake of computational efficiently, it is useful to work with a Lagrangian equivalent version of this optimization. Consider the following optimization

$$\text{minimize}_{\|m\|_1\le L} \quad \frac{1}{2} m^\mathsf{T} \widehat{\Sigma}^{(\ell)} m - \langle m, e_a \rangle + \mu_\ell \|m\|_1 \,, \tag{54}$$

with μ_ℓ and L taking the same values as in Optimization (14).

The next result, from [17, Chapter 5] is on the connection between the solutions of the unconstrained problem (54) and (14). For the reader's convenience, the proof is also given in Appendix G.1.

Lemma D.1. *A solution of optimization* (54) *is also a solution of the optimization problem* (14)*. Also, if problem* (14) *is feasible then problem* (54) *has bounded solution.*

Using the above lemma, we can instead work with the Lagrangian version (54) for constructing the decorrelating vector m_a^ℓ.

Here, we propose to solve optimization problem (54) using iterative method. Note the objective function evolves slightly at each episode and hence we expect the solutions m_a^ℓ and $m_a^{\ell+1}$ to be close to each other. An appealing property of iterative methods is that we can leverage this observation by setting m_a^ℓ as the initialization for the iterations that compute $m_a^{\ell+1}$, yielding shorter convergence time. In the sequel we discuss two of such iterative schemes.

D.1.1 Coordinate descent algorithms

In this method, at each iteration we update one of the coordinates of m, say m_j, while fixing the other coordinates. We write the objective function of (54) by separating m_j

from the other coordinates:

$$\frac{1}{2}\widehat{\Sigma}^{(\ell)}_{j,j} m_j^2 + \sum_{r,s\neq j} \widehat{\Sigma}^{(\ell)}_{r,s}\, m_r m_s - m_a + \mu_\ell \|m_{\sim j}\|_1 + \mu_\ell |m_j|\,, \tag{55}$$

where $\widehat{\Sigma}^{(\ell)}_{j,\sim j}$ denotes the j^{th} row (column) of $\widehat{\Sigma}^{(\ell)}$ with $\widehat{\Sigma}^{(\ell)}_{j,j}$ removed. Likewise, $m_{\sim j}$ represents the restriction of m to coordinates other than j. Minimizing (55) with respect to m_j gives

$$m_j + \frac{1}{\widehat{\Sigma}^{(\ell)}_{j,j}} \Big(\widehat{\Sigma}^{(\ell)}_{j,\sim j} m_{\sim j} - \mathbb{I}(a=j) + \mu_\ell \,\mathrm{sign}(m_j)\Big) = 0\,.$$

It is easy to verify that the solution of the above is given by

$$m_j = \frac{1}{\widehat{\Sigma}^{(\ell)}_{j,j}} \eta\Big(-\widehat{\Sigma}^{(\ell)}_{j,\sim j} m_{\sim j} + \mathbb{I}(a=j); \mu_\ell\Big)\,, \tag{56}$$

with $\eta(\cdot;\,\cdot) : \mathbb{R}\times\mathbb{R}_+ \to \mathbb{R}$ denoting the soft-thresholding function defined as

$$\eta(z,\mu) = \begin{cases} z-\mu & \text{if } z > \mu\,, \\ 0 & \text{if } -\mu \le z \le \mu\,, \\ z+\mu & \text{otherwise}\,. \end{cases} \tag{57}$$

For a vector u, $\eta(u;\mu)$ is perceived entry-wise.

This brings us to the following update rule to compute $m_a^\ell \in \mathbb{R}^{dp}$ (solution of (54)). Th notation Π_L, in line 5 below, denotes the Euclidean projection onto the ℓ_1 ball of radius L and can be computed in $O(dp)$ times using the procedure of [9].

```
1: (initialization): $m(0) \leftarrow m_a^{(\ell-1)}$
2: for iteration $h = 1, \dots, H$ do
3:     for $j = 1, 2, \dots, dp$ do
4:         $m_j(h) \leftarrow \frac{1}{\widehat{\Sigma}^{(\ell)}_{j,j}} \eta\Big( -\widehat{\Sigma}^{(\ell)}_{j,\sim j} m_{\sim j}(h-1) + \mathbb{I}(a=j); \mu_\ell\Big)$
5:     $m(h) \leftarrow \Pi_L(m(h))$
6: return $m_a^\ell \leftarrow m(H)$
```

In our experiments we implemented the same coordinate descent iterations explained above to solve for the decorrelating vectors m_a^ℓ.

D.1.2 Gradient descent algorithms

Letting $\mathcal{L}(m) = (1/2)m^{\mathsf{T}}\widehat{\Sigma}^{(\ell)}m - \langle m, e_a\rangle$, we can write the objective of (54) as $\mathcal{L}(m) + \mu_\ell\|m\|_1$. Projected gradient descent, applied to this constrained objective, results in a sequence of iterates $m(h)$, with $h = 0, 1, 2, \dots$ the iteration number, as follows:

$$\begin{aligned} m(h+1) = \arg\min_{\|m\|_1\le L} \Big\{ & \mathcal{L}(m(h)) + \langle \nabla\mathcal{L}(m(h)), m - m(h)\rangle \\ & + \frac{\eta}{2}\|m - m(h)\|_2^2 + \mu_\ell\|m\|_1 \Big\}. \end{aligned} \tag{58}$$

In words, the next iterate $m(h+1)$ is obtained by constrained minimization of a first order approximation to $\mathcal{L}(m)$, combined with a smoothing term that keeps the next iterate close to the current one. Since the objective function is convex ($\widehat{\Sigma}^{(\ell)} \succeq 0$), iterates (58) are guaranteed to converge to the global minimum of (54).

Plugging for $\mathcal{L}(m)$ and dropping the constant term $\mathcal{L}(m(h))$, update (58) reads as

$$\begin{aligned} m(h+1) &= \arg\min_{\|m\|_1\le L} \Big\{ \langle \widehat{\Sigma}^{(\ell)}m(h) - e_a, m - m(h)\rangle + \frac{\eta}{2}\|m - m(h)\|_2^2 + \mu_\ell\|m\|_1 \Big\} \\ &= \arg\min_{\|m\|_1\le L} \Big\{ \frac{\eta}{2}\Big(m - m(h) + \frac{1}{\eta}(\widehat{\Sigma}^{(\ell)}m(h) - e_a)\Big)^2 + \mu_\ell\|m\|_1 \Big\}. \end{aligned} \tag{59}$$

To compute the update (59), we first solve the unconstrained problem which has a closed form solution given by $\eta\Big(m(h) - \frac{1}{\eta}(\widehat{\Sigma}^{(\ell)}m(h) - e_a); \frac{\mu_\ell}{\eta}\Big)$, with η the soft thresholding function given by (57). The solution is then projected onto the ball of radius L.

In the following box, we summarize the projected gradient descent update rule for constructing the decorrelating vectors m_a^ℓ.

```
(initialization): $m(0) \leftarrow m_a^{(\ell-1)}$
for iteration $h = 1, \dots, H$ do
    $m(h) \leftarrow \eta\Big(m(h) - \frac{1}{\eta}(\widehat{\Sigma}^{(\ell)}m(h) - e_a); \frac{\mu_\ell}{\eta}\Big)$
    $m(h) \leftarrow \Pi_L(m(h))$
return $m_a^\ell \leftarrow m(H)$
```

E Numerical experiments

In this section, we evaluate the performance of online debiasing framework on synthetic data. In the interest of reproducibility, an R implementation of our algorithm is publicly available[6].

Consider the $\mathsf{VAR}(d)$ time series model (5). In the first setting, we let $p = 20$, $d = 3$, $T = 50$ and construct the covariance matrix of noise terms Σ_ζ by putting 1 on its diagonal and $\rho = 0.3$ on its off-diagonal. To make it closer to the practice, instead of considering sparse coefficient matrices, we work with *approximately* sparse matrices. Specifically, the entries of $A^{(i)}$ are generated independently from a Bernoulli distribution with success probability $q = 0.1$, multiplied by $b \cdot \text{Unif}(\{+1, -1\})$ with $b = 0.1$, and then added to a Gaussian matrix with mean 0 and standard error $1/p$. In formula, each entry is generated independently from

$$b \cdot \text{Bern}(q) \cdot \text{Unif}(\{+1, -1\}) + \mathcal{N}(0, 1/p^2)\,.$$

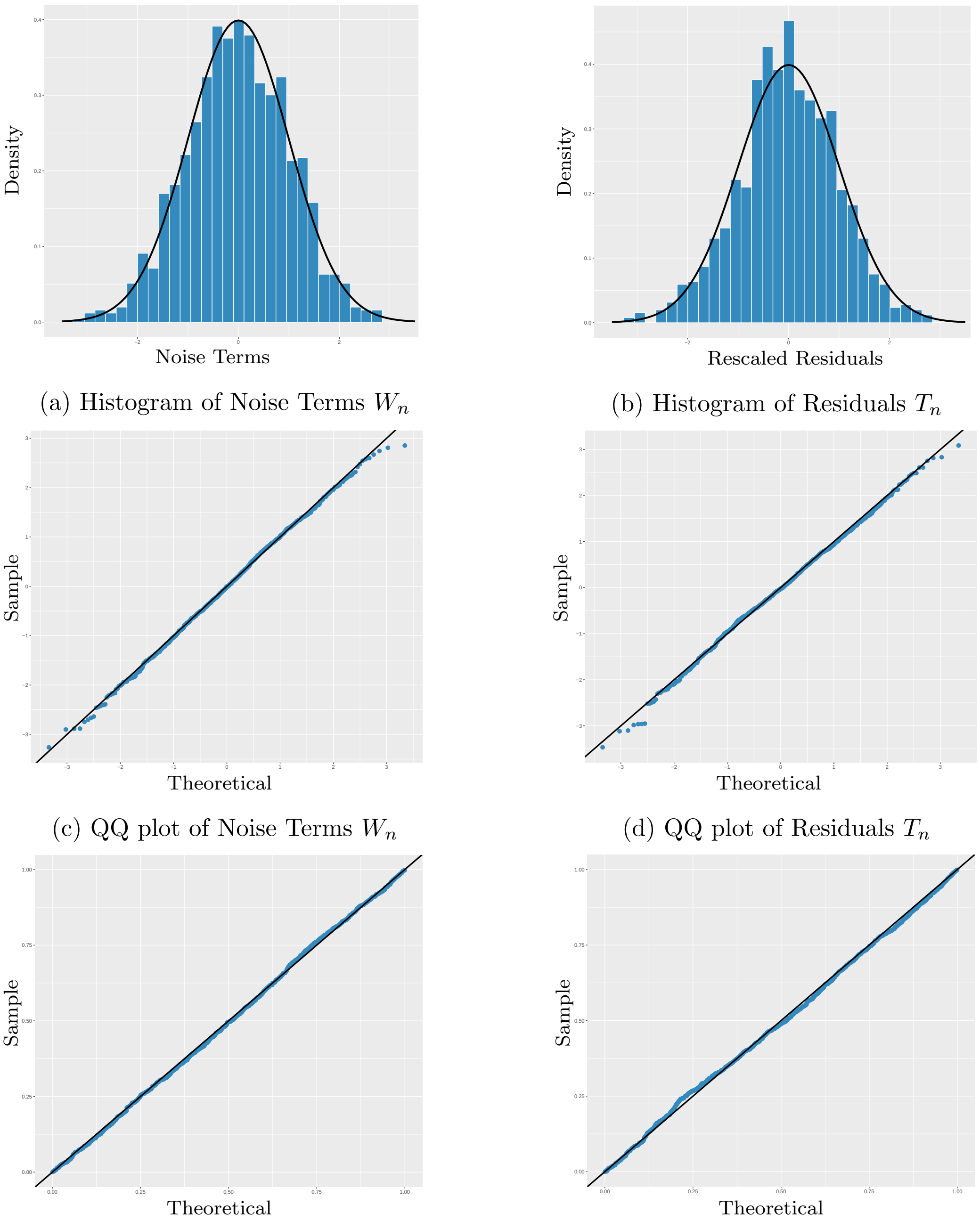
We used $r_0 = 6$ (length of first episode E_0) and $\beta = 1.3$ for lengths of other episodes $E_\ell \sim \beta^\ell$. For each $i \in [p]$ we do the following. Let $\theta_0 = (A_i^{(1)}, A_i^{(2)}, \dots, A_i^{(d)})^\mathsf{T} \in \mathbb{R}^{dp}$ encode the i^{th} rows of the matrices $A^{(\ell)}$ and compute the noise component of $\widehat{\theta}^{\mathsf{on}}$ as

$$W_n \equiv \frac{1}{\sqrt{n}} \sum_{\ell=0}^{K-1} M^{(\ell)} \Big(\sum_{t \in E_\ell} x_t \varepsilon_t \Big)\,, \tag{60}$$

the rescaled residual $T_n \in \mathbb{R}^{dp}$ with $T_{n,a} = \sqrt{\frac{n}{V_{n,a}}}(\widehat{\theta}^{\mathsf{on}}_a - \theta_{0,a})$, and $V_{n,a}$ given by Equation (23) and $\sigma = 1$. Left and right plots of Figure 9 denote the QQ-plot, PP-plot and histogram of noise terms W_n and rescaled residuals T_n of *all coordinates* (across all $i \in [p]$ and $a \in [dp]$) stacked together, respectively.

True and False Positive Rates. Consider the linear time-series model (5) with $A^{(i)}$ matrices having entries drawn independently from the distribution $b \cdot \text{Bern}(q) \cdot \text{Unif}(\{+1, -1\})$ and noise terms be gaussian with covariance matrix Σ_ζ. In this example, we evaluate the performance of our proposed online debiasing method for constructing confidence intervals and hypothesis testing as discussed in Section 5. We

[6] The link address is removed from the current blinded version of the manuscript



(a) Histogram of Noise Terms W_n

(b) Histogram of Residuals T_n

(c) QQ plot of Noise Terms W_n

(d) QQ plot of Residuals T_n

(e) PP plot of Noise Terms W_n

(f) PP plot of Residuals T_n

Figure 9: A simple example of an online debiased Var(3) process with dimension $p = 20$ and $T = 50$ sample data points. Plots 9a, 9c, 9e demonstrate respectively the histogram, QQ-plot, and PP plot of noise values of all $dp^2 = 1200$ entries of A_i matrices in linear time series model (5). Plots 9b, 9d, 9f are histogram, QQ-plot, and PP-plot of rescaled residuals of all coordinates as well. Alignment of data points in these plots with their corresponding standard normal $(0, 1)$ line corroborates our theoretical results on the asymptotic normal behavior of noise terms and rescaled residuals discussed in corollary 3.7 and proposition 3.8, respectively.

consider four metrics: True Positive Rate (TPR), False Positive Rate (FPR), Average length of confidence intervals (Avg CI length), and coverage rate of confidence intervals. Tables 3 and 4 summarize the results for various configurations of the Var(d) processes and significance level $\alpha = 0.05$. Table 3 corresponds to the cases where noise covariance has the structure $\Sigma_\zeta(i,j) = 0.1^{|i-j|}$ and Table 4 corresponds to the case of $\Sigma_\zeta(i,j) = 0.1^{\mathbb{I}(i \neq j)}$. The reported measures for each configuration (each row of the table) are average over 20 different realizations of the $\mathsf{VAR}(d)$ model.

Table 3: Evaluation of the online debiasing approach for statistical inference on the coefficients of a $\mathsf{VAR}(d)$ model under different configurations. Here the noise terms ζ_i are gaussian with covariance matrix $\Sigma_\zeta(i,j) = 0.1^{|i-j|}$. The results are reported in terms of four metrics: FPR (False Positive Rate), TPR (True Positive Rate), Coverage rate and Average length of confidence intervals (Avg CI length) at significance level $\alpha = 0.05$

	p	T	q	b	FPR	TPR	Avg CI length	Coverage rate
$d = 1$	40	30	0.01	2	0.0276	1	3.56	0.9725
	35	30	0.01	2	0.0354	0.9166	3.7090	0.9648
	60	55	0.01	0.9	0.0314	0.7058	2.5933	0.9686
$d = 2$	55	100	0.01	0.8	0.0424	0.8000	1.9822	0.9572
	40	75	0.01	0.9	0.0343	0.9166	2.5166	0.9656
	50	95	0.01	0.7	0.0368	0.6182	2.4694	0.963
$d = 3$	45	130	0.005	0.9	0.0370	0.6858	2.070	0.9632
	40	110	0.01	0.7	0.0374	0.6512	2.1481	0.9623
	50	145	0.005	0.85	0.0369	0.6327	2.2028	0.9631

Table 4: Evaluation of the online debiasing approach for statistical inference on the coefficients of a $\mathsf{VAR}(d)$ model under different configurations. Here the noise terms ζ_i are gaussian with covariance matrix $\Sigma_\zeta(i,j) = 0.1^{\mathbb{I}(i\neq j)}$. The results are reported in terms of four metrics: FPR (False Positive Rate), TPR (True Positive Rate), Coverage rate and Average length of confidence intervals (Avg CI length) at significance level $\alpha = 0.05$

	p	T	q	b	FPR	TPR	Avg CI length	Coverage rate
$d=1$	40	30	0.01	2	0.0402	1	3.5835	0.96
	40	35	0.02	1.2	0.0414	0.8125	2.6081	0.9575
	50	40	0.015	0.9	0.0365	0.7435	2.0404	0.9632
$d=2$	35	65	0.01	0.9	0.0420	0.8077	2.4386	0.9580
	45	85	0.01	0.9	0.0336	0.7298	2.5358	0.9655
	50	70	0.01	0.95	0.0220	0.8333	2.4504	0.9775
$d=3$	40	115	0.01	0.9	0.0395	0.7906	1.6978	0.9598
	45	130	0.005	0.95	0.0359	0.7714	2.1548	0.9641
	50	145	0.005	0.85	0.0371	0.5918	2.1303	0.9624

E.1 Real data experiments: a marketing application

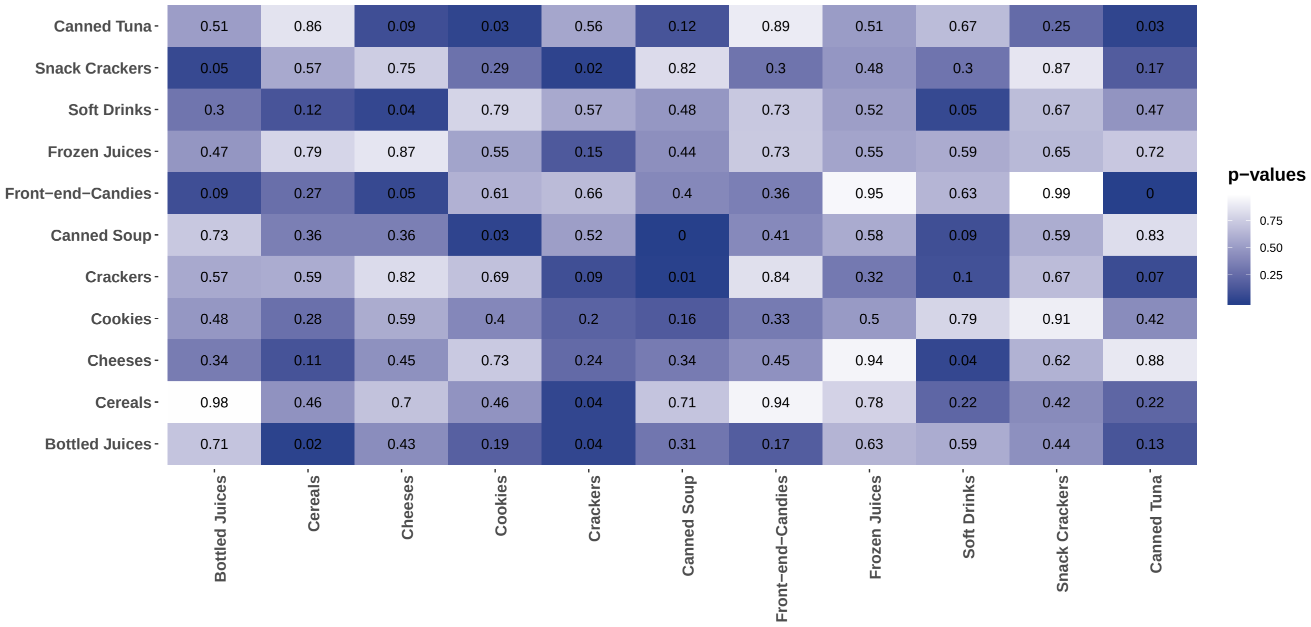
Retailers often offer sales of various categories of products and for an effective management of the business, they need to understand the cross-category effect of products on each other, e.g., how the price, promotion or sale of category A will effect the sales of category B after some time.

We used data of sales, prices and promotions of Chicago-area grocery store chain Dominick's that is publicly available at `https://research.chicagobooth.edu/kilts/marketing-databases/dominicks`. The same data set has been used in [12] where a sparse VAR model is fit to data and also in [39] where a VARX model is employed to estimate the demand effects (VARX models incorporate the effect of unmodeled exogenous variables (X) into the VAR). In this experiment, we use the proposed online debiasing approach to provide p-values for the category effects.

We consider 11 categories of products[7] over 71 weeks, so for each week t, we have information $x_t \in \mathbb{R}^{33}$ for sales, prices and promotions of the 11 categories. For thorough explanation on calculating sales, prices and promotions, we refer to [33] and [12]. We posit $\mathsf{VAR}(2)$ model as the generating process for covariates x_i and then apply our proposed online debiasing method to calculate two-sided p-values for the null hypothesis of form $H_0 : \theta_{0,a} = 0$ with $\theta_{0,a}$ an entry in the VAR model, as discussed earlier in Section 5 (See Eq. (42)). By running the Benjamini–Yekutieli procedure [4] (with log factor correction to account for dependence among p-values), we obtain the following statistically significant cross category associations at level 0.05: sales of canned tuna on sales of front-end-candies after one week with p-val= 5.8e-05, and price of crackers on sales of canned tuna after one week with p-val= 5.5e-04. In [12], sparse VAR models are used to construct networks of interlinked product categories, but they are not accompanied by statistical measures such as p-values. Our online debiasing method here provides p-values for individual possible cross-category associations.

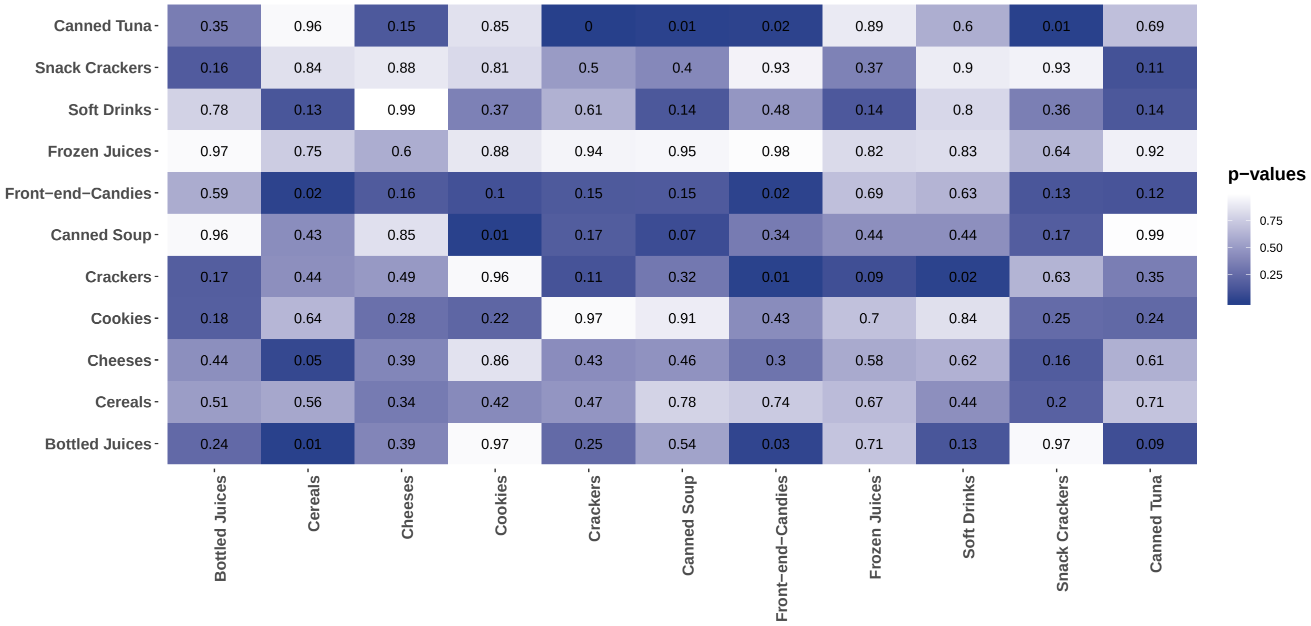
In the rest of this section we report the p-values obtained by the online debiasing for the cross-category effects. Figures 10, 11, 12 provide the p-values corresponding to the effect of price, sale, and promotions of different categories on the other categories, after one week ($d = 1$) and two weeks ($d = 2$). The darker cells indicate smaller p-values and hence higher statistical significance.

[7] Bottled Juices, Cereals, Cheeses, Cookies, Crackers, Canned Soup, Front-end-Candies, Frozen Juices, Soft Drinks, Snack Crackers and Canned Tuna



	Bottled Juices	Cereals	Cheeses	Cookies	Crackers	Canned Soup	Front−end−Candies	Frozen Juices	Soft Drinks	Snack Crackers	Canned Tuna
Canned Tuna	0.51	0.86	0.09	0.03	0.56	0.12	0.89	0.51	0.67	0.25	0.03
Snack Crackers	0.05	0.57	0.75	0.29	0.02	0.82	0.3	0.48	0.3	0.87	0.17
Soft Drinks	0.3	0.12	0.04	0.79	0.57	0.48	0.73	0.52	0.05	0.67	0.47
Frozen Juices	0.47	0.79	0.87	0.55	0.15	0.44	0.73	0.55	0.59	0.65	0.72
Front−end−Candies	0.09	0.27	0.05	0.61	0.66	0.4	0.36	0.95	0.63	0.99	0
Canned Soup	0.73	0.36	0.36	0.03	0.52	0	0.41	0.58	0.09	0.59	0.83
Crackers	0.57	0.59	0.82	0.69	0.09	0.01	0.84	0.32	0.1	0.67	0.07
Cookies	0.48	0.28	0.59	0.4	0.2	0.16	0.33	0.5	0.79	0.91	0.42
Cheeses	0.34	0.11	0.45	0.73	0.24	0.34	0.45	0.94	0.04	0.62	0.88
Cereals	0.98	0.46	0.7	0.46	0.04	0.71	0.94	0.78	0.22	0.42	0.22
Bottled Juices	0.71	0.02	0.43	0.19	0.04	0.31	0.17	0.63	0.59	0.44	0.13

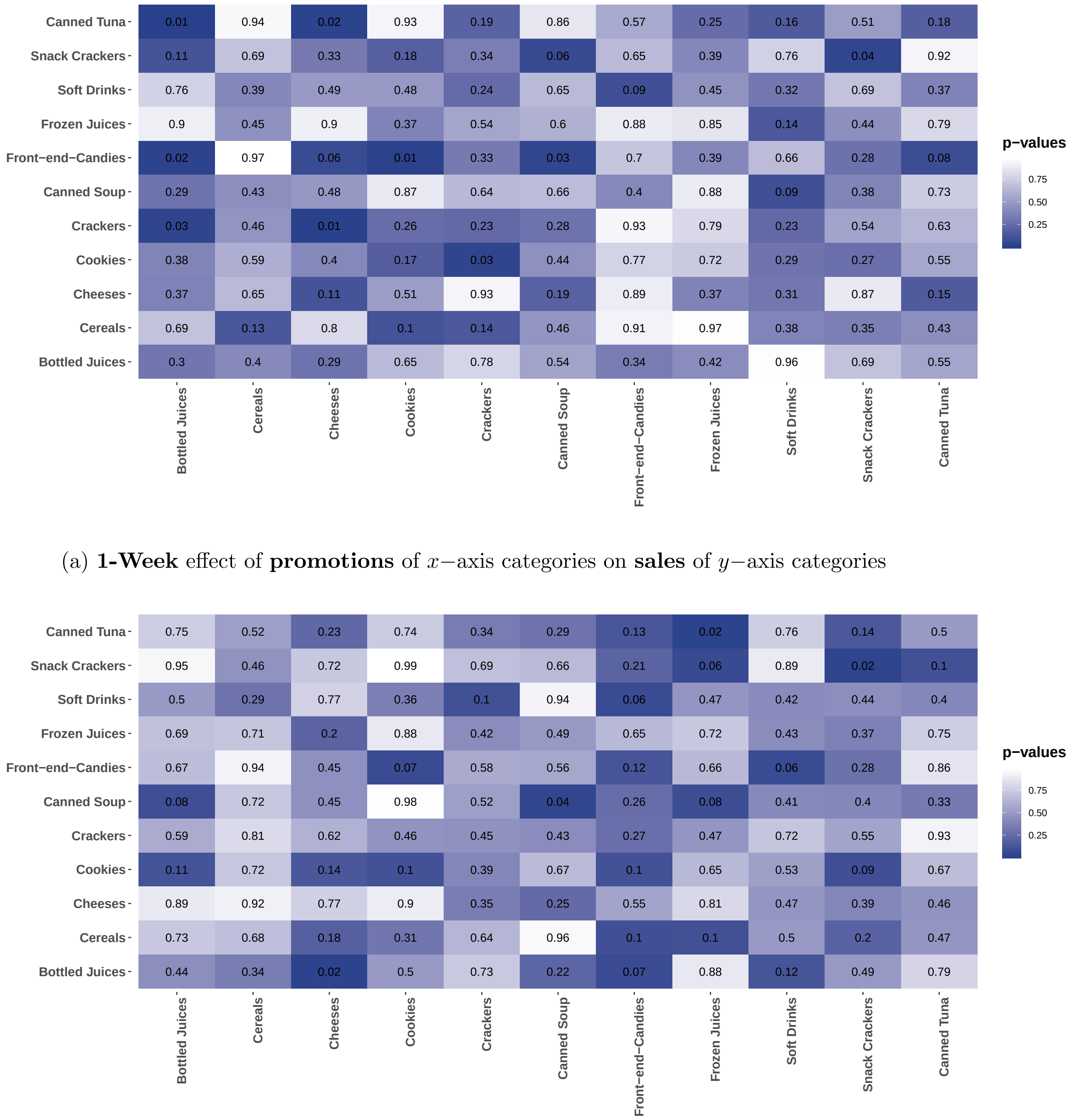
(a) **1-Week** effect of **sales** of $x-$axis categories on **sales** of $y-$axis categories



	Bottled Juices	Cereals	Cheeses	Cookies	Crackers	Canned Soup	Front−end−Candies	Frozen Juices	Soft Drinks	Snack Crackers	Canned Tuna
Canned Tuna	0.35	0.96	0.15	0.85	0	0.01	0.02	0.89	0.6	0.01	0.69
Snack Crackers	0.16	0.84	0.88	0.81	0.5	0.4	0.93	0.37	0.9	0.93	0.11
Soft Drinks	0.78	0.13	0.99	0.37	0.61	0.14	0.48	0.14	0.8	0.36	0.14
Frozen Juices	0.97	0.75	0.6	0.88	0.94	0.95	0.98	0.82	0.83	0.64	0.92
Front−end−Candies	0.59	0.02	0.16	0.1	0.15	0.15	0.02	0.69	0.63	0.13	0.12
Canned Soup	0.96	0.43	0.85	0.01	0.17	0.07	0.34	0.44	0.44	0.17	0.99
Crackers	0.17	0.44	0.49	0.96	0.11	0.32	0.01	0.09	0.02	0.63	0.35
Cookies	0.18	0.64	0.28	0.22	0.97	0.91	0.43	0.7	0.84	0.25	0.24
Cheeses	0.44	0.05	0.39	0.86	0.43	0.46	0.3	0.58	0.62	0.16	0.61
Cereals	0.51	0.56	0.34	0.42	0.47	0.78	0.74	0.67	0.44	0.2	0.71
Bottled Juices	0.24	0.01	0.39	0.97	0.25	0.54	0.03	0.71	0.13	0.97	0.09

(b) **1-Week** effect of **prices** of $x-$axis categories on **sales** of $y-$axis categories

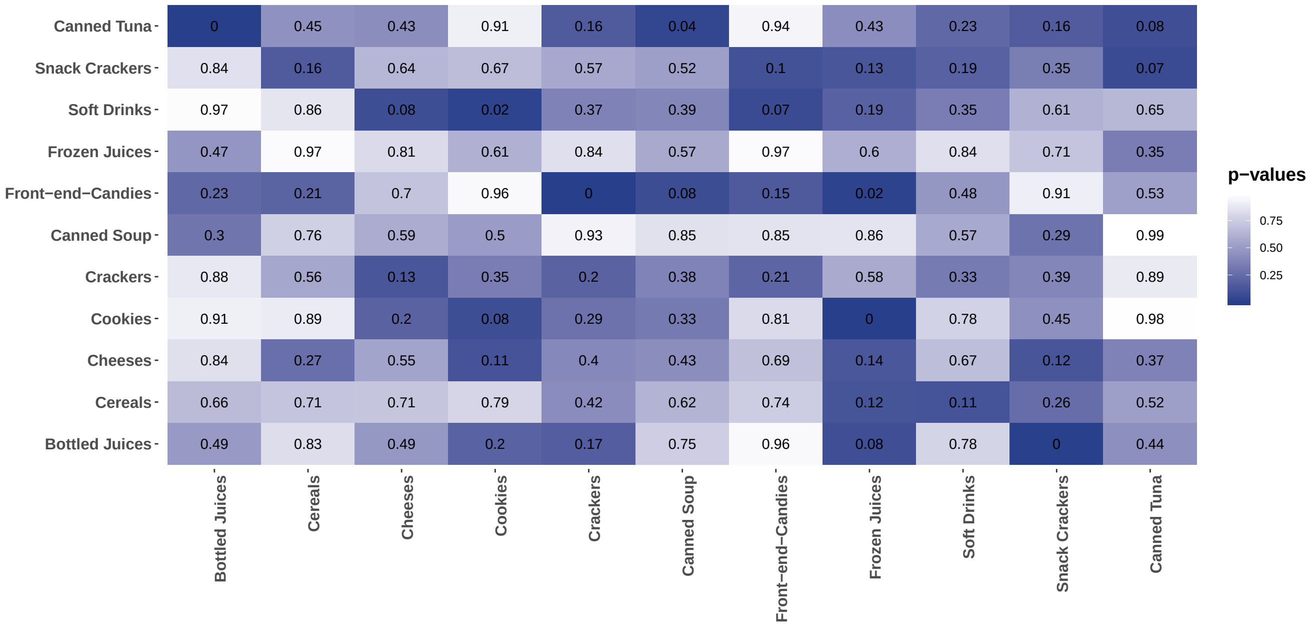
Figure 10: Figures 10a, and 10b respectively show the p-values for cross-category effects of sales and prices of $x-$axis categories on sales of $y-$axis categories after one week.



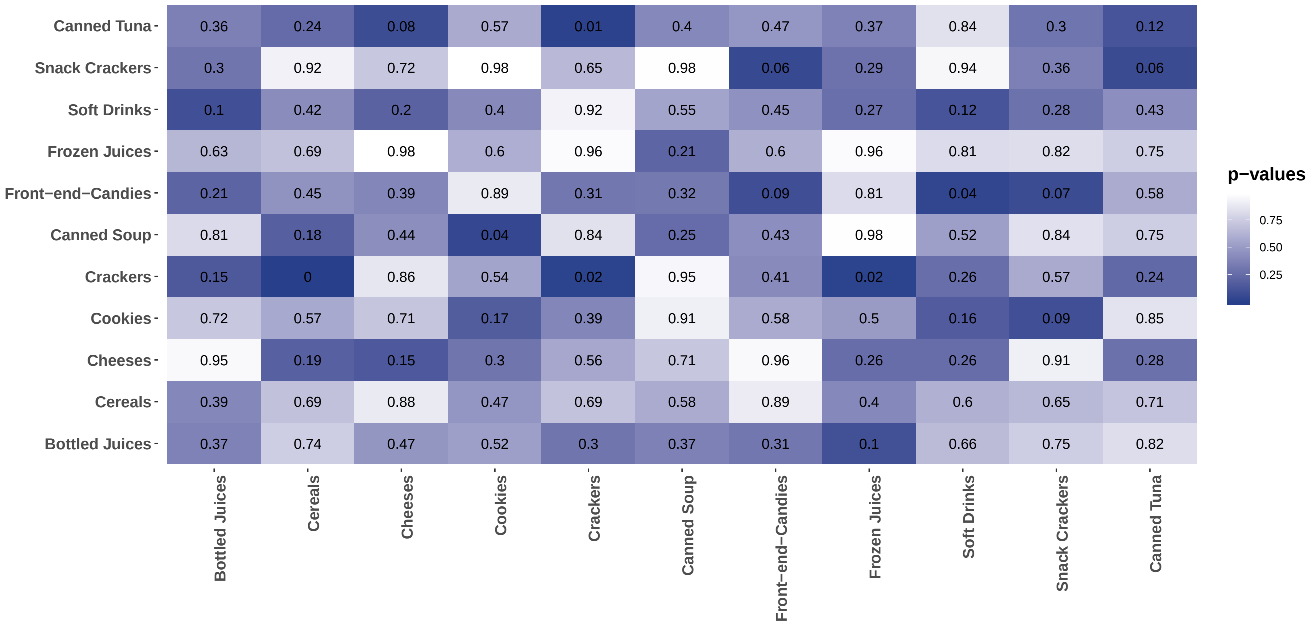
(a) **1-Week** effect of **promotions** of $x-$axis categories on **sales** of $y-$axis categories

(b) **2-Week** effect of **promotions** of $x-$axis categories on **sales** of $y-$axis categories

Figure 11: Figures 11a, and 11b show $p-$values for cross-category effects of promotions of $x-$axis categories on sales of $y-$axis categories, after one week and two weeks.



(a) **2-Week** effect of **sales** of $x-$axis categories on **sales** of $y-$axis categories



(b) **2-Week** effect of **prices** of $x-$axis categories on **sales** of $y-$axis categories

Figure 12: Figures 12a, and 12b respectively show p-values for cross-category effects of sales and prices of x-axis categories on sales of $y-$axis categories after two weeks.

F Proofs of Section 3

F.1 Technical preliminaries

Recall the definition of the regression design from Eqs.(11) in the time series case:

$$
\begin{aligned}
\theta_0 &= (A_i^{(1)}, A_i^{(2)}, \dots, A_i^{(d)})^\mathsf{T}, \\
X &= \begin{bmatrix} z_d^\mathsf{T} & z_{d-1}^\mathsf{T} & \dots & z_1^\mathsf{T} \\ z_{d+1}^\mathsf{T} & z_d^\mathsf{T} & \dots & z_2^\mathsf{T} \\ \vdots & \vdots & \ddots & \vdots \\ z_{T-1}^\mathsf{T} & z_{T-2}^\mathsf{T} & \dots & z_{T-d}^\mathsf{T} \end{bmatrix}, \\
y &= (z_{d+1,i}, z_{d+2,i}, \dots, z_{T,i}), \\
\varepsilon &= (\zeta_{d+1,i}, \zeta_{d+2,i}, \dots, \zeta_{T,i}).
\end{aligned}
$$

We first establish some preliminary results for stable time series. For the stationary process $x_t = (z_{t+d-1}^\mathsf{T}, \dots, z_t^\mathsf{T})^\mathsf{T}$ (rows of X), let $\Gamma_x(s) = \mathrm{Cov}(x_t, x_{t+s})$, for $t, s \in \mathbb{Z}$ and define the spectral density $f_x(r) \equiv 1/(2\pi) \sum_{\ell=-\infty}^{\infty} \Gamma_X(\ell) e^{-j\ell r}$, for $r \in [-\pi, \pi]$. The measure of stability of the process is defined as the maximum eigenvalue of the density

$$
M(f_x) \equiv \sup_{r \in [-\pi,\pi]} \sigma_{\max}(f_x(r)) \,. \tag{61}
$$

Likewise, the minimum eigenvalue of the spectrum is defined as $m(f_x) \equiv \inf\limits_{r \in [-\pi,\pi]} \sigma_{\min}(f_x(r))$, which captures the dependence among the covariates. (Note that for the case of i.i.d. samples, $M(f_x)$ and $m(f_x)$ reduce to the maximum and minimum eigenvalue of the population covariance.)

The p-dimensional $\mathsf{VAR}(d)$ model (5) can be represented as a dp-dimensional $\mathsf{VAR}(1)$ model. Recall our notation $x_t = (z_{t+d-1}^\mathsf{T}, \dots, z_t^\mathsf{T})^\mathsf{T}$ (rows of X in (11)). Then (5) can be written as

$$
x_t = \widetilde{A} x_{t-1} + \tilde{\zeta}_t \,, \tag{62}
$$

with

$$\widetilde{A} = \left(\begin{array}{cccc|c} A_1 & A_2 & \dots & A_{d-1} & A_d \\ \hline & & I_{(d-1)p} & & 0 \end{array}\right), \qquad \tilde{\zeta}_t = \begin{pmatrix} \zeta_{t+d-1} \\ 0 \end{pmatrix}. \tag{63}$$

The reverse characteristic polynomial for the $\mathsf{VAR}(1)$ model reads as $\tilde{\mathcal{A}} = I - \tilde{A}z$.

The following lemma controls $M(f_x), m(f_x)$ in terms of the spectral properties of the noise Σ_ζ and the characteristic polynomials $\mathcal{A}, \tilde{\mathcal{A}}$.

Lemma F.1 ([2])**.** *We have:*

$$\begin{aligned} \frac{1}{2\pi}\lambda_{\max}(\Sigma) \le M(f_x) &\le \frac{\lambda_{\max}(\Sigma_\zeta)}{\mu_{\min}(\tilde{\mathcal{A}})}, \\ \lambda_{\min}(\Sigma) &\ge \frac{\lambda_{\min}(\Sigma_\zeta)}{\mu_{\max}(\mathcal{A})}. \end{aligned} \tag{64}$$

We also use the following bound on $M(f_x)$ in terms of characteristic polynomial $\mathcal{A}$ of the time series z_t.

Lemma F.2. *The following holds:*

$$\frac{1}{2\pi}\lambda_{\max}(\Sigma) \le M(f_x) \le dM(f_z) \le \frac{d\lambda_{\max}(\Sigma_\zeta)}{\mu_{\min}(\mathcal{A})}.$$

Proof. Let $\Gamma_x(\ell) = \mathbb{E}[x_t x_{t+\ell}^{\mathsf{T}}]$ to refer the autocovariance of the dp-dimensional process x_t. Therefore $\Sigma = \Gamma_x(0)$. Likewise, the autocovariance $\Gamma_z(\ell)$ is defined for the p-dimensional process z_t. We represent $\Gamma_x(\ell)$ in terms of d^2 blocks, each of which is a $p \times p$ matrix. The block in position (r, s) is $\Gamma_z(\ell + r - s)$. Now, for a vector $v \in \mathbb{R}^{dp}$ with unit ℓ_2 norm, decompose it as d blocks of p dimensional vectors $v = (v_1^{\mathsf{T}}, v_2^{\mathsf{T}}, \dots, v_d^{\mathsf{T}})^{\mathsf{T}}$, by which we have

$$v^{\mathsf{T}}\Gamma_z(\ell)v = \sum_{1 \le r,s \le d} v_r^{\mathsf{T}}\Gamma_x(\ell + r - s)v_s\,. \tag{65}$$

Since the spectral density $f_z(\theta)$ is the Fourier transform of the autocorrelation function, we have by Equation (65),

$$\begin{aligned}
\langle v, f_z(\theta) v\rangle &= \frac{1}{2\pi}\sum_{\ell=-\infty}^{\infty} \langle v, \Gamma_z(\ell) e^{-j\ell\theta} v\rangle \\
&= \frac{1}{2\pi}\sum_{\ell=-\infty}^{\infty} \sum_{1\le r,s\le d} \langle v_r, \Gamma_z(\ell + r - s) e^{-j\ell\theta} v_s\rangle \\
&= \sum_{1\le r,s\le d} \langle v_r, \Big(\frac{1}{2\pi}\sum_{\ell=-\infty}^{\infty} \Gamma_x(\ell + r - s) e^{-j(\ell+r-s)\theta}\Big) v_s e^{j(r-s)\theta}\rangle \\
&= \sum_{1\le r,s\le d} \langle v_r, f_x(\theta) e^{j(r-s)\theta} v_s\rangle \\
&= V(\theta)^* f_x(\theta) V(\theta),
\end{aligned}$$

with $V(\theta) = \sum_{r=1}^{d} e^{-jr\theta} v_r$. Now, we have:

$$\|V(\theta)\|_2 \le \sum_{r=1}^{d} \|v_r\|_2 \le \Big(d \sum_{r=1}^{d} \|v_r\|_2^2\Big)^{1/2} \le \sqrt{d}.$$

Combining this with the Rayleigh quotient calculation above, yields $M(f_x) \le d M(f_z)$. Now, by using [2, Equation (4.1)] for the process z_t, with reverse characteristic polynomial $\mathcal{A}$, we obtain

$$\lambda_{\max}(\Sigma) \le 2\pi M(f_x) \le 2\pi d M(f_z) \le \frac{d\lambda_{\max}(\Sigma_\zeta)}{\mu_{\min}(\mathcal{A})}\,. \tag{66}$$

□

The following proposition is a straightforward consequence of the spectral bounds above and [2, Proposition 2.4].

Proposition F.3. *There exists a constant $c > 0$, such that for any vectors $u, v \in \mathbb{R}^{dp}$ with $\|u\| \le 1$, $\|v\| \le 1$, and any $\eta \ge 0$,*

$$\mathbb{P}\left(|u^{\mathsf{T}}(\widehat{\Sigma}^{(\ell)} - \Sigma)v| > \frac{d\lambda_{\max}(\Sigma_\zeta)}{\mu_{\min}(\mathcal{A})}\eta\right) \le 6\exp\left(-cn_\ell \min\{\eta^2, \eta\}\right)\,. \tag{67}$$

F.2 Remarks on proof of Theorem 3.2

The key part of establishing Theorem 3.2 is to establish an appropriate 'restricted eigenvalue' condition as follows:

Proposition F.4. *Let $\{z_1, \ldots, z_T\}$ be generated according to the (stable)* $\mathsf{VAR}(d)$ *process* (5) *and let $n = T - d$. Then there exist constants $c \in (0,1)$ and $C > 1$ such that for all $n \geq C\omega^2 \log(dp)$, with probability at least $1 - \exp(-cn/\omega^2)$, satisfies*

$$\langle v, (X^{\mathsf{T}} X/n)v \rangle \geq \alpha \|v\|^2 - \alpha\tau \|v\|_1^2.$$

Here, α, ω and τ are given by:

$$\begin{aligned} \omega &= \frac{d\lambda_{\max}(\Sigma_\zeta)\mu_{\max}(\mathcal{A})}{\lambda_{\min}(\Sigma_\zeta)\mu_{\min}(\mathcal{A})}\,, \\ \alpha &= \frac{\lambda_{\min}(\Sigma_\zeta)}{2\mu_{\max}(\mathcal{A})}\,, \\ \tau &= \omega^2 \sqrt{\frac{\log(dp)}{n}}\,. \end{aligned} \tag{68}$$

Given Proposition F.4, the estimation result of Theorem 3.2 is standard (see [6]). Proposition F.4 can be proved analogous to [2, Proposition 4.2], with the following considerations and minor modifications:

1. [2] writes the $\mathsf{VAR}(d)$ model as a $\mathsf{VAR}(1)$ model and then vectorize the obtained equation to get a linear regression form (cf. Section 4.1 of [2]). This way, they prove $I \otimes (X^{\mathsf{T}} X/n)$ satisfies a restricted eigenvalue property. Towards this, the first step in their proof is to show that $X^{\mathsf{T}} X/n$ satisfies a restricted eigenvalue property, i.e. Proposition F.4.

2. [2, Proposition 4.2] assumes $n \geq Ck \max\{\omega^2, 1\} \log(dp)$, with $k = \sum_{\ell=1}^{d} \|\text{vec}(A^{(\ell)})\|_0$, the total number of nonzero entries of matrices A_ℓ and then it is later used to get $\tau \leq 1/(Ck)$. However, as the restricted eigenvalue condition is independent of the sparsity of matrices $A^{(\ell)}$, we can use their result with $k = 1$.

3. The proof involves upper bounding $M(f_x)$, for which we use Lemma F.2 in lieu of Lemma F.1.

F.3 Proof of Lemma 3.3

The idea is to use Proposition F.3 along with the union bound. Fix $i, j \in [dp]$ and let $u = \frac{\Omega e_i}{\|\Omega e_i\|}$ and $v = e_j$. Then:

$$\begin{aligned}
|(\Omega\widehat{\Sigma}^{(\ell)} - I)_{ij}| &= |\langle \Omega e_i, (\widehat{\Sigma}^{(\ell)} - \Sigma)e_j\rangle| \\
&= \|\Omega e_i\| |\langle u, (\widehat{\Sigma}^{(\ell)} - \Sigma)v\rangle| \\
&\le \lambda_{\max}(\Omega) |\langle u, (\widehat{\Sigma}^{(\ell)} - \Sigma)v\rangle| \\
&\le \frac{\mu_{\max}(\mathcal{A})}{\lambda_{\min}(\Sigma_\zeta)} |\langle u, (\widehat{\Sigma}^{(\ell)} - \Sigma)v\rangle|,
\end{aligned}$$

where the last line uses Lemma F.1 to bound $\lambda_{\min}(\Sigma)$ from below. Combining this with Proposition F.3, for $\eta \le 1$:

$$\begin{aligned}
\mathbb{P}\Big\{|(\Omega\widehat{\Sigma}^{(\ell)} - I)_{ij}| \ge d\lambda_{\max}(\Sigma_\zeta)\eta/\mu_{\min}(\mathcal{A})\Big\} &\le \mathbb{P}\Big\{|\langle u, (\widehat{\Sigma}^{(\ell)} - \Sigma)v\rangle| \ge \omega\eta\Big\} \\
&\le 6\exp(-cn_\ell\eta^2).
\end{aligned}$$

Setting $\eta = C\sqrt{\log(dp)/n_\ell}$ for a large enough constant C, the probability bound above is smaller than $(dp)^{-8}$. With a union bound over $i, j \in [dp]$:

$$\begin{aligned}
\mathbb{P}\Big\{\|\Omega\widehat{\Sigma}^{(\ell)} - I\|_\infty \ge C\omega\sqrt{\frac{\log(dp)}{n_\ell}}\Big\} &\le (dp)^2 \sup_{i,j} \mathbb{P}\Big\{|(\Omega\widehat{\Sigma}^{(\ell)} - I)_{ij}| \ge C\omega\sqrt{\frac{\log(dp)}{n_\ell}}\Big\} \\
&\le (dp)^{-6}.
\end{aligned}$$

This completes the proof.

F.4 Proof of Theorem 3.4

Starting from the decomposition (20), we have

$$\sqrt{n}(\widehat{\theta}^{\mathsf{on}} - \theta_0) = \Delta_n + W_n\,,$$

with $\Delta_n = B_n(\widehat{\theta}^{\mathsf{L}} - \theta_0)$. As explained below (20), W_n is a martingale with respect to filtration $\mathcal{F}_j = \{\varepsilon_1, \dots, \varepsilon_j\}$, $j \in \mathbb{N}$ and hence $\mathbb{E}(W_n) = 0$.

We also note that $\|\Delta_n\|_\infty \le \|B_n\|_\infty \|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1$. Our next lemma bounds $\|B_n\|_\infty$.

Lemma F.5. *Suppose that the decorrelating matrices $M^{(\ell)}$ are computed according to Eq.*(14)*, with $\mu_\ell = C\omega\sqrt{(\log(dp)/n_\ell}$ and $L \ge \|\Omega\|_1$. Let ω and γ be:*

$$\omega = \frac{d\mu_{\max}(\mathcal{A})\lambda_{\max}(\Sigma_\zeta)}{\mu_{\min}(\mathcal{A})\lambda_{\min}(\Sigma_\zeta)},$$
$$\gamma = \frac{d\lambda_{\max}(\Sigma_\zeta)}{\mu_{\min}(\mathcal{A})}.$$

Then, for B_n given by (18)*, the following bound holds with probability at least $1-(dp)^{-8}$:*

$$\|B_n\|_\infty \le \frac{r_0}{\sqrt{n}} + \sum_{\ell=1}^{K-1}\left[\frac{r_\ell\mu_\ell}{\sqrt{n}} + CL\gamma\sqrt{\frac{\log(dp)}{n}}\Big(\frac{r_\ell}{\sqrt{n_\ell}} + \sqrt{r_\ell}\Big)\right], \tag{69}$$
$$\le \frac{r_0}{\sqrt{n}} + C(\omega + L\gamma)\sqrt{\frac{\log(dp)}{n}}\sum_{\ell=1}^{K-1}\Big(\frac{r_\ell}{\sqrt{n_\ell}} + \sqrt{r_\ell}\Big). \tag{70}$$

The bound (70) holds for general batch sizes $r_0, \dots, r_{K-1}$. A natural approach to choose the values r_ℓ is by minimizing this upper bound. However, this is not a convex function in terms of r_ℓ. Focusing just on the last term in the bound, we have $\sum_{\ell=1}^{K-1}\sqrt{r_\ell} \ge (\sum_{\ell=1}^{K-1} r_\ell)^{1/2} = \sqrt{n}$. Therefore, the provided bound on $\|B_n\|_\infty$ is at least of order $\sqrt{\log(dp)}$. We next propose a choice of batch sizes r_ℓ for which the bound (70) achieves this order. Let $r_0 = \sqrt{n}$, $r_\ell = \beta^\ell$ for some $\beta > 1$ and $\ell = 1, \dots, K-2$. Finally we choose r_{K-1} so that the total lengths of batches add up to n (that is $r_0 + r_1 + \ldots + r_{K-1} = n$). Following this choice, bound (70) simplifies to:

$$\|B_n\|_\infty \le C_\beta(\omega + \gamma L)\sqrt{\log(dp)}\,, \tag{71}$$

for some constant $C_\beta > 0$ that depends on the constant β.

Next by combining Theorem 3.2 and Lemma F.5 we obtain that, with probability at least $1 - 2(dp)^{-6}$

$$\|\Delta_n\|_\infty \le C_\beta(\omega + L\gamma)\sqrt{\log(dp)} \cdot \Big(\frac{s_0\lambda_n}{\alpha}\Big)$$
$$\le C_\beta\frac{\lambda_0(\omega + L\gamma)}{\alpha}\frac{s_0\log(dp)}{\sqrt{n}}. \tag{72}$$

This implies the claim by selecting a β bounded away from 1, say $\beta = 1.3$.

It remains to prove the claim on the bias $\mathbb{E}\{\widehat{\theta}^{\mathsf{on}} - \theta_0\}$. For this, define G to be the event where Δ_n satisfies the upper bound in Eq.(72). Therefore:

$$\begin{aligned}
\|\mathbb{E}\{\widehat{\theta}^{\mathsf{on}} - \theta_0\}\|_\infty &= \frac{\|\mathbb{E}\{\Delta_n\}\|_\infty}{\sqrt{n}} \\
&\le \frac{\|\mathbb{E}\{\Delta_n \mathbb{I}(G)\}\|_\infty}{\sqrt{n}} + \mathbb{E}\{\|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1 \mathbb{I}(G^c)\}.
\end{aligned}$$

For the first term we use the bound Eq.(72). For the second, we use Lemma I.7:

$$\|\mathbb{E}\{\widehat{\theta}^{\mathsf{on}} - \theta_0\}\|_\infty \le \frac{C\lambda_0(\omega + L\gamma)}{\alpha}\frac{s_0 \log p}{n} + \frac{\mathbb{E}\{\|\varepsilon\|^2 \mathbb{I}(G^c)\}}{n\lambda_n} + 2\|\theta_0\|_1 \mathbb{P}(G^c).$$

It suffices, therefore, to show that the final two terms are at most $C\|\theta_0\|_1/(dp)^6$. By Holder inequality and $\mathbb{P}(G^c) \le 2(dp)^{-6}$:

$$\begin{aligned}
\frac{\mathbb{E}\{\|\varepsilon\|^2 \mathbb{I}(G^c)\}}{n\lambda_n} + 2\|\theta_0\|_1 \mathbb{P}(G^c) &\le \frac{\mathbb{E}\{\|\varepsilon\|^4\}^{1/2} \mathbb{P}(G^c)^{1/2}}{n\lambda_n} + 2\|\theta_0\|_1 \mathbb{P}(G^c) \\
&\le C\frac{\lambda_{\max}(\Sigma_\zeta)^2}{(dp)^3 \lambda_0 \sqrt{n \log(dp)}} + C\frac{\|\theta_0\|_1}{(dp)^6}.
\end{aligned}$$

In the high-dimensional regime, the first term is negligible in comparison to $s_0 \log(dp)/n$, which yields, after adjusting C appropriately:

$$\|\mathbb{E}\{\widehat{\theta}^{\mathsf{on}} - \theta_0\}\|_\infty \le \frac{C_1\lambda_0(\omega + L\gamma)}{\alpha}\frac{s_0 \log p}{n} + C_2\frac{\|\theta_0\|_1}{(dp)^6},$$

as required.

It remains to prove Lemma F.5:

Proof of Lemma F.5. For each episode ℓ, let

$$R^{(\ell)} := \frac{1}{r_\ell} \sum_{t \in E_\ell} x_t x_t^{\mathsf{T}}$$

be the sample covariance in episode ℓ. Fix $a \in [dp]$ and define $B_{n,a} \equiv \sqrt{n} e_a - \frac{1}{\sqrt{n}} \sum_{\ell=1}^{K-1} r_\ell R^{(\ell)} m_a^\ell$. We then have

$$B_{n,a} = \sqrt{n} e_a - \frac{1}{\sqrt{n}} \sum_{\ell=1}^{K-1} r_\ell R^{(\ell)} m_a^\ell = \frac{r_0}{\sqrt{n}} e_a + \sum_{\ell=1}^{K-1} \frac{r_\ell}{\sqrt{n}} \Big(e_a - R^{(\ell)} m_a^\ell \Big), \tag{73}$$

where we used that $\sum_{\ell=0}^{K-1} r_\ell = n$. By triangle inequality, followed by Holder inequality:

$$\begin{aligned}
\|B_{n,a}\|_\infty &\le \frac{r_0}{\sqrt{n}} + \frac{1}{\sqrt{n}} \sum_{\ell=1}^{K-1} r_\ell \|e_a - R^{(\ell)} m_a^\ell\|_\infty \\
&\le \frac{r_0}{\sqrt{n}} + \sum_{\ell=1}^{K-1} \frac{r_\ell}{\sqrt{n}} \big(\|e_a - \widehat{\Sigma}^{(\ell)} m_a^\ell\|_\infty + \|(\widehat{\Sigma}^{(\ell)} - \Sigma) m_a^\ell\|_\infty + \|(\Sigma - R^{(\ell)}) m_a^\ell\|_\infty \big) \\
&\le \frac{r_0}{\sqrt{n}} + \sum_{\ell=1}^{K-1} \frac{r_\ell}{\sqrt{n}} \big(\|e_a - \widehat{\Sigma}^{(\ell)} m_a^\ell\|_\infty + \|\widehat{\Sigma}^{(\ell)} - \Sigma\|_\infty \|m_a^\ell\|_1 + \|\Sigma - R^{(\ell)}\|_\infty \|m_a^\ell\|_1 \big)
\end{aligned}$$

We now bound each of the three terms appearing in the sum above:

1. By the construction of decorrelating vectors m_a^ℓ as in optimization (14), we have

$$\|\widehat{\Sigma}^{(\ell)} m_a^\ell - e_a\|_\infty \le \mu_\ell \,, \quad \ell = 0, \dots, K-1 \,. \tag{74}$$

2. Also by construction, $\|m_a^\ell\|_1 \le L$. From an argument similar to that of Lemma 3.3, $\|\widehat{\Sigma}^{(\ell)} - \Sigma\|_\infty \le C\gamma\sqrt{\log(dp)/n_\ell}$ with probability at least $1 - K(dp)^{-9}$, where $\gamma = d\lambda_{\max}(\Sigma_\zeta)/\mu_{\min}(\mathcal{A})$. Therefore, with the same probability, the second term is at most $CL\gamma\sqrt{\log(dp)/n_\ell}$.

3. Again, by construction $\|m_a^\ell\|_1 \le L$. Similar to Lemma 3.3, $\|R^{(\ell)} - \Sigma\|_\infty$ is at most $C\gamma\sqrt{\log(dp)/r_\ell}$ with probability at least $1 - K(dp)^{-9}$.

Combining these and the fact that we set $\mu_\ell = C\omega\sqrt{\log(dp)/n}$ we have that, with probability at least $1 - 2K(dp)^{-9}$,

$$\begin{aligned}
\|B_{n,a}\|_\infty &\le \frac{r_0}{\sqrt{n}} + \frac{C}{\sqrt{n}} \sum_{\ell=1}^{K-1} r_\ell \left(\omega\sqrt{\frac{\log(dp)}{n_\ell}} + L\gamma\sqrt{\frac{\log(dp)}{n_\ell}} + L\gamma\sqrt{\frac{\log(dp)}{r_\ell}} \right) \\
&\le \frac{r_0}{\sqrt{n}} + C(\omega + L\gamma)\sqrt{\frac{\log(dp)}{n}} \sum_{\ell=1}^{K-1} \Big(\frac{r_\ell}{\sqrt{n_\ell}} + \sqrt{r_\ell} \Big).
\end{aligned}$$

This bound holds uniformly over $a \in [dp]$, and since $\|B_n\|_\infty = \sup_a \|B_{n,a}\|_\infty$, the same bound holds for $\|B_n\|_\infty$. This completes the proof. □

F.5 Proof of Lemma 3.6

We start by proving Claim (23). Let $m_a = \Omega e_a$ be the first column of the inverse (stationary) covariance. Using the fact that $\mathbb{E}\{x_t x_t^{\mathsf{T}}\} = \Sigma$ we have $\langle m_a, \mathbb{E}\{x_t x_t^{\mathsf{T}}\} m_a\rangle = \Omega_{a,a}$, which is to be the dominant term in the conditional variance $V_{n,a}$. Recall the shorthand $\sigma^2 \equiv \Sigma_{\zeta_{i,i}}$, with $i \in [p]$ the fixed coordinate in (11). Therefore, we decompose the difference as follows:

$$
\begin{aligned}
V_{n,a} - \sigma^2 \Omega_{a,a} &= \frac{\sigma^2}{n} \sum_{\ell=1}^{K-1} \sum_{t \in E_\ell} \Big[\langle m_a^\ell, x_t \rangle^2 - \Omega_{a,a} \Big] - \frac{r_0 \sigma^2}{n} \Omega_{a,a} \\
&= \frac{\sigma^2}{n} \sum_{\ell=1}^{K-1} \sum_{t \in E_\ell} \Big[\langle m_a^\ell, x_t \rangle^2 - \langle m_a, \mathbb{E}\{x_t x_t^{\mathsf{T}}\} m_a \rangle \Big] - \frac{r_0 \sigma^2}{n} \Omega_{a,a} \\
&= \frac{\sigma^2}{n} \sum_{\ell=1}^{K-1} \sum_{t \in E_\ell} [\langle m_a^\ell, x_t \rangle^2 - \langle m_a, x_t \rangle^2] \\
&\quad + \frac{1}{n} \sum_{t=0}^{n-1} \langle m_a, (x_t x_t^{\mathsf{T}} - \mathbb{E}\{x_t x_t^{\mathsf{T}}\}) m_a \rangle - \frac{r_0 \sigma^2}{n} \Omega_{a,a} \,. \qquad (75)
\end{aligned}
$$

We treat each of these three terms separately. Write

$$
\begin{aligned}
\bigg| \frac{1}{n} \sum_{\ell=1}^{K-1} \sum_{t \in E_\ell} [\langle m_a^\ell, x_t \rangle^2 - \langle m_a, x_t \rangle^2] \bigg| &= \frac{1}{n} \bigg| \sum_{\ell=1}^{K-1} \sum_{t \in E_\ell} [\langle m_a^\ell - m_a, x_t \rangle \langle m_a^\ell + m_a, x_t \rangle] \bigg| \\
&\le \frac{1}{n} \bigg\| \sum_{\ell=1}^{K-1} \sum_{t \in E_\ell} \langle m_a^\ell - m_a, x_t \rangle x_t \bigg\|_\infty \| m_a^\ell + m_a \|_1 \\
&\le \frac{2L}{n} \bigg\| \sum_{\ell=1}^{K-1} \sum_{t \in E_\ell} \langle m_a^\ell - m_a, x_t \rangle x_t \bigg\|_\infty \,. \qquad (76)
\end{aligned}
$$

To bound the last quantity, note that

$$\begin{aligned}
\frac{1}{n}\left\|\sum_{\ell=1}^{K-1}\sum_{t\in E_\ell}\langle m_a^\ell - m_a, x_t\rangle x_t\right\|_\infty &\le \left\|e_a - \frac{1}{n}\sum_{\ell=1}^{K-1}\sum_{t\in E_\ell}\langle m_a^\ell, x_t\rangle x_t\right\|_\infty \\
&\quad + \left\|e_a - \frac{1}{n}\sum_{\ell=1}^{K-1}\sum_{t\in E_\ell}\langle m_a, x_t\rangle x_t\right\|_\infty \\
&= \left\|e_a - \frac{1}{n}\sum_{\ell=1}^{K-1} r_\ell R^{(\ell)} m_a^\ell\right\|_\infty + \left\|e_a - \widehat{\Sigma}^{(K)} m_a\right\|_\infty \\
&= \frac{1}{\sqrt{n}}\|B_{n,a}\|_\infty + \left\|e_a - \widehat{\Sigma}^{(K)} m_a\right\|_\infty \\
&\le CL\gamma\sqrt{\frac{\log(dp)}{n}} + C\omega\sqrt{\frac{\log(dp)}{n}} \le C(L\gamma + \omega)\sqrt{\frac{\log(dp)}{n}}\,,
\end{aligned} \tag{77}$$

for some constant C. The last inequality follows from the positive events of Lemma F.5 and Lemma 3.3. Combining Equations (76) and (77), we obtain

$$\left|\frac{1}{n}\sum_{\ell=1}^{K-1}\sum_{t\in E_\ell}[\langle m_a^\ell, x_t\rangle^2 - \langle m_a, x_t\rangle^2]\right| \le CL(\omega + L\gamma)\sqrt{\frac{\log(dp)}{n}}. \tag{78}$$

For the second term in (75), we can use Proposition F.3 with $v = u = m_a/\|m_a\|, \eta = C\sqrt{\log(dp)/n}$ to obtain

$$\begin{align*}
\left|\frac{1}{n}\sum_{t=0}^{n-1}\langle m_a, (x_t x_t^\mathsf{T} - \mathbb{E}\{x_t x_t^\mathsf{T}\}) m_a\rangle\right| &= \left|\langle m_a, (\widehat{\Sigma}^{(K-1)} - \Sigma) m_a\rangle\right| \\
&\le \frac{Cd\lambda_{\max}(\Sigma_\zeta)}{\mu_{\min}(\mathcal{A})}\|m_a\|^2\sqrt{\frac{\log(dp)}{n}} \\
&\le \frac{Cd\lambda_{\max}(\Sigma_\zeta)}{\mu_{\min}(\mathcal{A})\lambda_{\min}(\Sigma)^2}\sqrt{\frac{\log(dp)}{n}} \tag{79} \\
&\le \frac{C\omega}{\alpha}\sqrt{\frac{\log(dp)}{n}}\,, \tag{80}
\end{align*}$$

where we used that $\|m_a\| = \|\Omega e_a\| \le \lambda_{\max}(\Omega) = \lambda_{\min}(\Sigma)^{-1} \le 1/\alpha$. For the third term, we have $r_0 = \sqrt{n}$. Also, $\Omega_{a,a} \le \lambda_{\max}(\Omega) \le 1/\alpha$. Therefore, this term is $O(1/\alpha\sqrt{n})$. Combining this bound with (78) and (80) in Equation (75) we get the Claim (23).

We next prove Claim (24). Note that $|\varepsilon_t| = |\zeta_{t+d,i}|$ is bounded with $\sigma\sqrt{2\log(n)}$, with high probability for $t \in [n]$, by tail bound for Gaussian variables. In addition,

$\max_\ell|\langle m_a^\ell, x_t\rangle| \le \|m_a^\ell\|_1\|x_t\|_\infty \le L\|x_t\|_\infty \le L\|X\|_\infty$. Note that variance of each entry $x_{t,i}$ is bounded by $\Sigma_{ii} \le \lambda_{\max}(\Sigma)$. Hence, by tail bound for Gaussian variables and union bounding we have

$$\mathbb{P}\left(\|X\|_\infty < \sqrt{2\lambda_{\max}(\Sigma)\log(dpn)}\right) \ge 1-(pdn)^{-2}\,, \tag{81}$$

Putting these bounds together we get

$$\begin{aligned}
&\max\Big\{\frac{1}{\sqrt{n}}|\langle m_a^\ell, x_t\rangle\varepsilon_t| :\, \ell\in[K-2],\, t\in[n]\Big\}\\
&\le \frac{1}{\sqrt{n}}L\sqrt{2\lambda_{\max}(\Sigma)\log(dpn)}\sigma\sqrt{2\log(n)}\\
&\le 2L\sigma\sqrt{\lambda_{\max}(\Sigma)}\,\frac{\log(dpn)}{\sqrt{n}}\\
&\le 2L_0\sigma\|\Omega\|_1\left(\frac{2\pi d\lambda_{\max}(\Sigma_\zeta)}{\mu_{\min}(\mathcal{A})}\right)^{1/2}\frac{\log(dpn)}{\sqrt{n}} = o(1)\,,
\end{aligned}$$

where in the last inequality we used Lemma F.2 to upper bound $\lambda_{\max}(\Sigma_\zeta)$. The conclusion that the final expression is $o(1)$ follows from Assumption 3.5.

F.6 Proof of Proposition 3.8

We prove that for all $x\in\mathbb{R}$,

$$\lim_{n\to\infty}\sup_{\|\theta_0\|_0\le s_0}\mathbb{P}\Big\{\frac{\sqrt{n}(\widehat{\theta}_a^{\mathsf{on}}-\theta_{0,a})}{\sqrt{V_{n,a}}}\le x\Big\}\le\Phi(x)\,. \tag{82}$$

We can obtain a matching lower bound by a similar argument which implies the result.

Invoking the decomposition (21) we have

$$\frac{\sqrt{n}(\widehat{\theta}_a^{\mathsf{on}}-\theta_{0,a})}{\sqrt{V_{n,a}}} = \frac{W_n}{\sqrt{V_{n,a}}}+\frac{\Delta_n}{\sqrt{V_{n,a}}}\,.$$

By Corollary 3.7, we have that $\widetilde{W}_n \equiv W_n/\sqrt{V_{n,a}} \to \mathsf{N}(0,1)$ in distribution. Fix an arbitrary $\varepsilon>0$ and write

$$\begin{aligned}
\mathbb{P}\Big\{\frac{\sqrt{n}(\widehat{\theta}_a^{\mathsf{on}}-\theta_{0,a})}{\sqrt{V_{n,a}}}\le x\Big\} &= \mathbb{P}\Big\{\widetilde{W}_n+\frac{\Delta_n}{\sqrt{V_{n,a}}}\le x\Big\}\\
&\le \mathbb{P}\{\widetilde{W}_n\le x+\varepsilon\}+\mathbb{P}\Big\{\frac{|\Delta_a|}{\sqrt{V_{n,a}}}\ge\varepsilon\Big\}
\end{aligned}$$

By taking the limit and using Equation (21), we get

$$\lim_{n\to\infty}\sup_{\|\theta_0\|_0\le s_0}\mathbb{P}\Big\{\frac{\sqrt{n}(\widehat{\theta}^{\sf on}_a-\theta_{0,a})}{\sqrt{V_{n,a}}}\le x\Big\}\le\Phi(x+\varepsilon)+\lim_{n\to\infty}\sup_{\|\theta_0\|_0\le s_0}\mathbb{P}\Big\{\frac{|\Delta_a|}{\sqrt{V_{n,a}}}\ge\varepsilon\Big\} \tag{83}$$

We show that the limit on the right hand side vanishes for any $\varepsilon > 0$. By virtue of Lemma 3.6 (Equation (23)), we have

$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}\Big\{\frac{|\Delta_a|}{\sqrt{V_{n,a}}}\ge\varepsilon\Big\}&\le\lim_{n\to\infty}\mathbb{P}\Big\{\frac{|\Delta_a|}{\sigma\sqrt{\Omega_{a,a}}}\ge\varepsilon\Big\}\\&\le\lim_{n\to\infty}\mathbb{P}\Big\{|\Delta_a|\ge\varepsilon\sigma\sqrt{\Omega_{a,a}}\Big\}\\&\le\lim_{n\to\infty}(dp)^{-4}=0\,.\end{aligned} \tag{84}$$

Here, in the last inequality we used that $s_0(L\gamma+\omega)=o(\sqrt{n}/\log(dp))$ and therefore, for large enough n, $\varepsilon\sigma\sqrt{\Omega_{a,a}}$ exceeds the bound (22) of Theorem 3.4.

Using (84) in bound (83) and then taking the limit $\varepsilon\to 0$, we obtain (82).

G Proofs of Section D

G.1 Proof of Lemma D.1

Rewrite the optimization problem (14) as follows:

$$\begin{aligned}\text{minimize}\quad & m^{\sf T}\widehat{\Sigma}^{(\ell)}m\\ \text{subject to}\quad & \langle z,\widehat{\Sigma}^{(\ell)}m-e_a\rangle\le\mu_\ell,\quad\|m\|_1\le L,\quad\|z\|_1=1\,,\end{aligned} \tag{85}$$

The Lagrangian is given by

$$\mathcal{L}(m,z,\lambda)=m^{\sf T}\widehat{\Sigma}^{(\ell)}m+\lambda(\langle z,\widehat{\Sigma}^{(\ell)}m-e_a\rangle-\mu_\ell),\quad\|z\|_1=1,\quad\|m\|_1\le L\,, \tag{86}$$

If $\lambda\le 2L$, minimizing Lagrangian over m is equivalent to $\frac{\partial\mathcal{L}}{\partial m}=0$ and we get $m_*=-\lambda z_*/2$. The dual problem is then given by

$$\begin{aligned}\text{maximize}\quad & -\frac{\lambda^2}{4}z^{\sf T}\widehat{\Sigma}^{(\ell)}z-\lambda\langle z,e_a\rangle-\lambda\mu_\ell\\ \text{subject to}\quad & \frac{\lambda}{2}\le L,\quad\|z\|_1=1\,,\end{aligned} \tag{87}$$

As $\|z\|_1 = 1$, by introducing $\beta = -\frac{\lambda}{2} z$, we get $\|\beta\|_1 = \frac{\lambda}{2}$. Rewrite the dual optimization problem in terms of β to get

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\beta^{\mathsf{T}}\widehat{\Sigma}^{(\ell)}\beta - \langle \beta, e_a\rangle + \mu_\ell \|\beta\|_1 \\ \text{subject to} \quad & \|\beta\|_1 \le L\,, \end{aligned} \tag{88}$$

Given β_* as the minimizer of the above optimization problem, from the relation of β and z we realize that $m_* = \beta_*$.

Also note that since optimization (88) is the dual to problem (85), we have that if (85) is feasible then the problem (88) is bounded.

G.2 Proof of Lemma A.1

By virtue of Proposition F.4, the sample covariance $\widehat{\Sigma}$ satisfies RE condition, $\widehat{\Sigma} \sim \text{RE}(\alpha, \tau)$, where

$$\alpha = \frac{\lambda_{\min}(\Sigma_\zeta)}{2\mu_{\max}(\mathcal{A})}\,, \qquad \tau = C\omega^2\sqrt{\frac{\log(dp)}{n}}\,, \tag{89}$$

and by the sample size condition we have $s_\Omega < 1/32\tau$.

Hereafter, we use the shorthand $m_a^* = \Omega e_a$ and let $\mathcal{L}(m)$ be the objective function in the optimization (50). By optimality of m_a, we have $\mathcal{L}(m_a^*) \le \mathcal{L}(m_a)$. Defining the error vector $\nu \equiv m_a - m_a^*$ and after some simple algebraic calculation we obtain the equivalent inequality

$$\frac{1}{2}\nu^{\mathsf{T}}\widehat{\Sigma}\nu \le \langle \nu, e_a - \widehat{\Sigma} m_a^*\rangle + \mu_n(\|m_a^*\|_1 - \|m_a^* + \nu\|_1)\,. \tag{90}$$

In the following we first upper bound the right hand side. By Lemma 3.3 (for $\ell = K$ and $n_K = n$), we have that with high probability

$$\langle \nu, e_a - \widehat{\Sigma} m_a^*\rangle \le \|\nu\|_1 a\sqrt{\frac{\log(dp)}{n}} = (\|\nu_S\|_1 + \|\nu_{S^c}\|_1)\frac{\mu_n}{2}\,,$$

where $S = \text{supp}(\Omega e_a)$ and hence $|S| \le s_\Omega$. On the other hand,

$$\|m_a + \nu\|_1 - \|m_a^*\|_1 \ge (\|m_{a,S}^*\|_1 - \|\nu_S\|_1) + \|\nu_{S^c}\|_1 - \|m_a^*\|_1 = \|\nu_{S^c}\|_1 - \|\nu_S\|_1\,.$$

Combining these pieces we get that the right-hand side of (90) is upper bounded by

$$(\|\nu_S\|_1 + \|\nu_{S^c}\|_1)\frac{\mu_n}{2} + \mu_n\left(\|\nu_S\|_1 - \|\nu_{S^c}\|_1\right) = \frac{3}{2}\mu_n\|\nu_S\|_1 - \frac{1}{2}\mu_n\|\nu_{S^c}\|_1\,, \tag{91}$$

Given that $\widehat{\Sigma} \succeq 0$, the left hand side of (90) is non-negative, which implies that $\|\nu_{S^c}\|_1 \le 3\|\nu_S\|_1$ and hence

$$\|\nu\|_1 \le 4\|\nu_S\|_1 \le 4\sqrt{s_\Omega}\|\nu_S\|_2 \le 4\sqrt{s_\Omega}\|\nu\|_2\,. \tag{92}$$

Next by using the restricted eigenvalue condition for $\widehat{\Sigma}$ we write

$$\nu^{\mathsf{T}}\widehat{\Sigma}\nu \ge \alpha\|\nu\|_2^2 - \alpha\tau\|\nu\|_1^2 \ge \alpha(1-16s_\Omega\tau)\|\nu\|_2^2 \ge \frac{\alpha}{2}\|\nu\|_2^2\,, \tag{93}$$

where we used $\tau \le 1/(32s_\Omega)$ in the final step.

Putting (90), (91) and (93) together, we obtain

$$\frac{\alpha}{4}\|\nu\|_2^2 \le \frac{3}{2}\mu_n\|\nu_S\|_1 \le 6\sqrt{s_\Omega}\mu_n\|\nu\|_2\,.$$

Simplifying the bound and using equation 92, we get

$$\begin{aligned}\|\nu\|_2 &\le \frac{24}{\alpha}\sqrt{s_\Omega}\mu_n\,,\\ \|\nu\|_1 &\le \frac{96}{\alpha}s_\Omega\mu_n\,,\end{aligned}$$

which completes the proof.

G.3 Proof of Theorem A.2

Continuing from the decomposition (48) we have

$$\sqrt{n}(\widehat{\theta}^{\mathsf{off}} - \theta_0) = \Delta_1 + \Delta_2 + Z\,, \tag{94}$$

with $Z = \Omega X^{\mathsf{T}}\varepsilon/\sqrt{n}$. By using Lemma 3.3 (for $\ell = K$) and recalling the choice of $\mu = \tau\sqrt{\log(dp)/n}$ we have that the following optimization is feasible, with high probability:

$$\begin{aligned}&\text{minimize} \quad m^{\mathsf{T}}\widehat{\Sigma}m\\ &\text{subject to} \quad \|\widehat{\Sigma}m - e_a\|_\infty \le \mu\,.\end{aligned}$$

Therefore, optimization (50) (which is shown to be its dual in Lemma (D.1)) has bounded solution. Hence, its solution should satisfy the KKT condition which reads as

$$\widehat{\Sigma} m_a - e_a + \mu \text{sign}(m_a) = 0\,, \tag{95}$$

which implies $\|\widehat{\Sigma} m_a - e_a\|_\infty \le \mu$. Invoking the estimation error bound of Lasso for time series (Proposition 3.2), we bound Δ_1 as

$$\|\Delta_1\|_\infty \le C\sqrt{n}\mu s_0 \sqrt{\frac{\log p}{n}} = O_P\Big(s_0 \frac{\log(dp)}{\sqrt{n}}\Big)\,. \tag{96}$$

We next bound the bias term Δ_2. By virtue of [2, Proposition 3.2] we have the deviation bound $\|X^\mathsf{T}\varepsilon\|_\infty/\sqrt{n} = O_P(\sqrt{\log(dp)})$, which in combination with Lemma A.1 gives us the following bound

$$\|\Delta_2\|_\infty \le \left(\max_{i\in[dp]} \|(M-\Omega)e_i)\|\right)\left(\frac{1}{\sqrt{n}}\|X^\mathsf{T}\varepsilon\|_\infty\right) = O_P\Big(s_\Omega \frac{\log(dp)}{\sqrt{n}}\Big)\,. \tag{97}$$

Therefore, letting $\Delta = \Delta_1 + \Delta_2$, we have $\|\Delta\|_\infty = o_P(1)$, by recalling our assumption $s_0 = o(\sqrt{n}/\log(dp))$ and $s_\Omega = o(\sqrt{n}/\log(dp))$.

Our next lemma is analogous to Lemma 3.6 for the covariance of the noise component in the offline debiased estimator, and its proof is deferred to Section G.1.

Lemma G.1. *Assume that $s_\Omega = o(\sqrt{n}/\log(dp))$ and $\Lambda_{\min}(\Sigma_\epsilon)/\mu_{\max}(\mathcal{A}) > c_{\min} > 0$ for some constant $c_{min} > 0$. For $\mu = \tau\sqrt{\log(dp)/n}$ and the decorrelating vectors m_i constructed by* (50)*, the following holds. For any fixed sequence of integers $a(n) \in [dp]$, we have*

$$m_a^\mathsf{T} \widehat{\Sigma} m_a = \Omega_{a,a} + o_P(1/\sqrt{\log(dp)})\,. \tag{98}$$

We are now ready to prove the theorem statement. We show that

$$\lim_{n\to\infty} \sup_{\|\theta_0\|_0 \le s_0} \mathbb{P}\left\{\frac{\sqrt{n}(\widehat{\theta}^{\mathsf{off}}_a - \theta_{0,a})}{\sqrt{V_{n,a}}} \le u\right\} \le \Phi(u)\,. \tag{99}$$

A similar lower bound can be proved analogously. By the decomposition (94) we have

$$\frac{\sqrt{n}(\widehat{\theta}^{\mathsf{off}}_a - \theta_{0,a})}{\sqrt{V_{n,a}}} = \frac{\Delta_a}{\sqrt{V_{n,a}}} + \frac{Z_a}{\sqrt{V_{n,a}}}\,.$$

Define

$$\widetilde{Z}_a \equiv \frac{Z_a}{\sigma\sqrt{\Omega_{a,a}}} = \frac{1}{\sigma\sqrt{n\Omega_{a,a}}}(\Omega X^{\mathsf{T}}\varepsilon)_a = \frac{1}{\sigma\sqrt{n\Omega_{a,a}}}\sum_{i=1}^{n} e_a^{\mathsf{T}}\Omega x_i \varepsilon_i \,.$$

Since ε_i is independent of x_i, the summand $\sum_{i=1}^{n} e_a^{\mathsf{T}}\Omega x_i\varepsilon_i$ is a martingale. Furthermore, $\mathbb{E}[(e_a^{\mathsf{T}}\Omega x_i\varepsilon_i)^2] = \sigma^2\Omega_{a,a}$. Hence, by a martingale central limit theorem [14, Corollary 3.2], we have that $\widetilde{Z}_a \to \mathsf{N}(0,1)$ in distribution. In other words,

$$\lim_{n\to\infty} \mathbb{P}\{\widetilde{Z}_a u\} = \Phi(u)\,. \tag{100}$$

Next, fix $\delta \in (0,1)$ and write

$$\begin{aligned}
\mathbb{P}\left\{\frac{\sqrt{n}(\widehat{\theta}_a^{\mathsf{off}} - \theta_{0,a})}{\sqrt{V_{n,a}}} \le u\right\} &= \mathbb{P}\left\{\frac{\sqrt{\Omega_{a,a}}}{\sqrt{V_{n,a}}}\widetilde{Z}_a + \frac{\Delta_a}{\sqrt{V_{n,a}}} \le u\right\} \\
&\le \mathbb{P}\left\{\frac{\sqrt{\Omega_{a,a}}}{\sqrt{V_{n,a}}}\widetilde{Z}_a \le u+\delta\right\} + \mathbb{P}\left\{\frac{\Delta_a}{\sqrt{V_{n,a}}} \ge \delta\right\} \\
&\le \mathbb{P}\left\{\widetilde{Z}_a \le u + 2\delta + \delta|u|\right\} + \mathbb{P}\left\{\Big|\frac{\sqrt{\Omega_{a,a}}}{\sqrt{V_{n,a}}} - 1\Big| \ge \delta\right\} \\
&\quad + \mathbb{P}\left\{\frac{\Delta_a}{\sqrt{V_{n,a}}} \ge \delta\right\}\,.
\end{aligned}$$

Now by taking the limit of both sides and using (100) and Lemma G.1, we obtain

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{\|\theta_0\|_0 \le s_0} \mathbb{P}\left\{\frac{\sqrt{n}(\widehat{\theta}_a^{\mathsf{off}} - \theta_{0,a})}{\sqrt{V_{n,a}}} \le u\right\} \le \\
&\qquad \Phi(u + 2\delta + \delta|u|) + \limsup_{n\to\infty} \sup_{\|\theta_0\|_0 \le s_0} \mathbb{P}\left\{\frac{\Delta_a}{\sqrt{V_{n,a}}} \ge \delta\right\}\,.
\end{aligned} \tag{101}$$

Since $\delta \in (0,1)$ was chosen arbitrarily, it suffices to show that the limit on the right hand side vanishes. To do that, we use Lemma G.1 again to write

$$\begin{aligned}
\lim_{n\to\infty} \sup_{\|\theta_0\|_0 \le s_0} \mathbb{P}\Big\{\frac{|\Delta_a|}{\sqrt{V_{n,a}}} \ge \delta\Big\} &\le \lim_{n\to\infty} \sup_{\|\theta_0\|_0 \le s_0} \mathbb{P}\Big\{\frac{|\Delta_a|}{\sigma\sqrt{(\Omega_{a,a}}} \ge \delta\Big\} \\
&\le \lim_{n\to\infty} \sup_{\|\theta_0\|_0 \le s_0} \mathbb{P}\Big\{|\Delta_a| \ge \delta\sigma\sqrt{\Omega_{a,a}}\Big\} = 0\,,
\end{aligned}$$

where the last step follows since we showed $\|\Delta\|_\infty = o_P(1)$. The proof is complete.

G.3.1 Proof of Lemma G.1

By invoking bound (64) on minimum eigenvalue of the population covariance, we have

$$\lambda_{\min}(\Sigma) \ge \frac{\lambda_{\min}(\Sigma_\zeta)}{\mu_{\max}(\mathcal{A})}, \tag{102}$$

bounded away from 0 by our assumption. Therefore, $\lambda_{\max}(\Omega) = \lambda_{\min}(\Sigma)^{-1}$ is bounded away from ∞. Since $\Omega \succcurlyeq 0$, we have $|\Omega_{a,b}| \le \sqrt{\Omega_{a,a}\Omega_{b,b}}$ for any two indices $a, b \in [dp]$. Hence, $|\Omega|_\infty \le 1/\lambda_{\min}(\Sigma)$. This implies that $\|\Omega e_a\|_1 \le s_\Omega/\lambda_{\min}(\Sigma)$. Using this observation along with the bound established in Lemma A.1, we obtain

$$\|m_a\|_1 \le \|\Omega e_a\| + \|m_a - \Omega e_a\|_1 \le \frac{s_\Omega}{\lambda_{\min}(\Sigma)} + \frac{192\tau}{\alpha} s_\Omega \sqrt{\frac{\log(dp)}{n}} = O(s_\Omega)\,. \tag{103}$$

We also have

$$\|m_a - \Omega e_a\|_\infty \le \|m_a - \Omega e_a\|_1 = O\Big(s_\Omega \sqrt{\frac{\log(dp)}{n}}\Big)\,. \tag{104}$$

In addition, by the KKT condition (95) we have

$$\|\widehat{\Sigma} m_a - e_a\|_\infty \le \mu\,. \tag{105}$$

Combining bounds (103), (104) and (105), we have

$$\begin{aligned} |m_a^\mathsf{T} \widehat{\Sigma} m_a - \Omega_{a,a}| &\le |(m_a^\mathsf{T} \widehat{\Sigma} - e_a^\mathsf{T}) m_a| + |e_a^\mathsf{T} m_a - \Omega_{a,a}| \\ &\le \|m_a^\mathsf{T} \widehat{\Sigma} - e_a^\mathsf{T}\|_\infty \|m_a\|_1 + \|m_a - \Omega e_a\|_\infty \\ &= O\Big(s_\Omega \sqrt{\frac{\log(dp)}{n}}\Big) = o(1/\sqrt{\log(dp)})\,, \end{aligned}$$

which completes the proof.

H Proofs of Section 4

H.1 Consistency results for LASSO under adaptively collected samples

Theorem 4.2 shows that, under an appropriate compatibility condition, the LASSO estimate admits ℓ_1 error at a rate of $s_0\sqrt{\log p/n}$. Importantly, despite the adaptivity introduced by the sampling of data, the error of LASSO estimate has the same asymptotic rate as expected without adaptivity. With slightly stronger restricted-eigenvalue conditions on the covariances $\mathbb{E}\{xx^{\mathsf{T}}\}$ and $\mathbb{E}\{xx^{\mathsf{T}}|\langle x, \widehat{\theta}^1\rangle \geq \varsigma\}$, it is also possible to extend Theorem 4.2 to show ℓ_2 error of order $s_0 \log p/n$, analogous to the non-adaptive setting. However, since the ℓ_2 error rate will not be used for our analysis of online debiasing, we do not pursue this direction here.

H.1.1 Proof of Theorem 4.2

The important technical step is to prove that, under the conditions specified in Theorem 4.2, the *sample covariance* $\widehat{\Sigma} = (1/n)\sum_i x_i x_i^{\mathsf{T}}$ is $(\phi_0/4, \mathrm{supp}(\theta_0))$ compatible.

Proposition H.1. *With probability exceeding* $1 - p^{-4}$ *the sample covariance* $\widehat{\Sigma}$ *is* $(\phi_0/4, \mathrm{supp}(\theta_0))$ *compatible when* $n_1 \vee n_2 \geq C(\kappa^4/\phi_0^2)s_0^2 \log p$*, for an absolute constant* $C > 0$.

Let $\widehat{\Sigma}^{(1)}$ and $\widehat{\Sigma}^{(2)}$ denote the sample covariances of each batch, i.e. $\widehat{\Sigma}^{(1)} = (1/n_1)\sum_{i\leq n_1} x_i x_i^{\mathsf{T}}$ and similarly $\widehat{\Sigma}^{(2)} = (1/n_2)\sum_{i>n_1} x_i x_i^{\mathsf{T}}$. We also let $\Sigma^{(2)}$ be the conditional covariance $\Sigma^{(2)} = \Sigma^{(2)}(\widehat{\theta}^1) = \mathbb{E}\{xx^{\mathsf{T}}|\langle x, \widehat{\theta}^1\rangle \geq \varsigma\}$. We first prove that at least one of the sample covariances $\widehat{\Sigma}^{(1)}$ and $\widehat{\Sigma}^{(2)}$ closely approximate their population counterparts, and that this implies they are $(\phi_0/2, \mathrm{supp}(\theta_0))$-compatible.

Lemma H.2. *With probability at least* $1 - p^{-4}$

$$\|\widehat{\Sigma}^{(1)} - \Sigma\|_\infty \wedge \|\widehat{\Sigma}^{(2)} - \Sigma^{(2)}\|_\infty \leq 12\kappa^2\sqrt{\frac{\log p}{n}},$$

Proof. Since $n = n_1 + n_2 \leq 2\max(n_1, n_2)$, at least one of n_1 and n_2 exceeds $n/2$. We assume that $n_2 \geq n/2$, and prove that $\|\widehat{\Sigma}^{(2)} - \Sigma^{(2)}\|_\infty$ satisfies the bound in the claim. The case $n_1 \geq n/2$ is similar. Since we are proving the case $n_2 \geq n/2$, for notational convenience, we assume probabilities and expectations in the rest of the proof are conditional on the first batch $(y_1, x_1), \dots (y_{n_1}, x_{n_1})$, and omit this in the notation.

For a fixed pair $(a, b) \in [p] \times [p]$:

$$\widehat{\Sigma}^{(2)}_{a,b} - \Sigma^{(2)}_{a,b} = \frac{1}{n_2}\sum_{i>n_1} x_{i,a}x_{i,b} - \mathbb{E}\{x_{i,a}x_{i,b}\}$$

Using Lemma I.4 we have that $\|x_{i,a}x_{i,b}\|_{\psi_1} \le 2\|x_i\|_{\psi_2}^2 \le 2\kappa^2$ almost surely. Then using the tail inequality Lemma I.5 we have for any $\varepsilon \le 2e\kappa^2$

$$\mathbb{P}\Big\{|\widehat{\Sigma}^{(2)}_{a,b} - \Sigma^{(2)}_{a,b}| \ge \varepsilon\Big\} \le 2\exp\Big\{-\frac{n_2\varepsilon^2}{6e\kappa^4}\Big\}$$

With $\varepsilon = \varepsilon(p, n_2, \kappa) = 12\kappa^2\sqrt{\log p/n_2} \le 20\kappa^2\sqrt{\log p/n}$ we have that $\mathbb{P}\{|\widehat{\Sigma}^{(2)}_{a,b} - \Sigma^{(2)}_{a,b}| \ge \varepsilon(p, n_2, \kappa)\} \le p^{-8}$, whence the claim follows by union bound over pairs (a, b). □

Lemma H.3 ([6, Corollary 6.8]). *Suppose that Σ is (ϕ_0, S)-compatible. Then any matrix Σ' such that $\|\Sigma' - \Sigma\|_\infty \le \phi_0/(32|S|)$ is $(\phi_0/2, S)$-compatible.*

We can now prove Proposition H.1.

Proof of Proposition H.1. Combining Lemmas H.2 and H.3 yields that, with probability $1 - p^{-4}$, at least one of $\widehat{\Sigma}^{(1)}$ and $\widehat{\Sigma}^{(2)}$ are $(\phi_0/2, \mathrm{supp}(\theta_0))$-compatible provided

$$12\kappa^2\sqrt{\frac{\log p}{n}} \le \frac{\phi_0}{32 s_0},$$

$$\text{which is implied by } n \ge \Big(\frac{400\kappa^2 s_0}{\phi_0}\sqrt{\log p}\Big)^2.$$

Since $\widehat{\Sigma} = (n_1/n)\widehat{\Sigma}^{(1)} + (n_2/n)\widehat{\Sigma}^{(2)}$ and at least one of n_1/n and n_2/n exceed 1/2, this implies that $\widehat{\Sigma}$ is $(\phi_0/4, \mathrm{supp}(\theta_0))$-compatible with probability exceeding $1 - p^{-4}$. □

The following lemma shows that $X^\mathsf{T}\varepsilon$ is small entrywise.

Lemma H.4. *For any $\lambda_n \ge 40\kappa\sigma\sqrt{(\log p)/n}$, with probability at least $1-p^{-4}$, $\|X^\mathsf{T}\varepsilon\|_\infty \le n\lambda_n/2$.*

Proof. The a^{th} coordinate of the vector $X^\mathsf{T}\varepsilon$ is $\sum_i x_{ia}\varepsilon_i$. As the rows of X are uniformly κ-subgaussian and $\|\varepsilon_i\|_{\psi_2} = \sigma$, Lemma I.4 implies that the sequence $(x_{ia}\varepsilon_i)_{1\le i\le n}$ is uniformly $2\kappa\sigma$-subexponential. Applying the Bernstein-type martingale tail bound Lemma I.6, for $\varepsilon \le 12e\kappa\sigma$:

$$\mathbb{P}\Big\{\Big|\sum_i x_{ia}\varepsilon_i\Big| \ge \varepsilon n\Big\} \le 2\exp\Big\{-\frac{n\varepsilon^2}{24e\kappa^2\sigma^2}\Big\}$$

Set $\varepsilon = \varepsilon(p,n,\kappa,\sigma) = 20\kappa\sigma\sqrt{(\log p)/n}$, the exponent on the right hand side above is at least $5\log p$, which implies after union bound over a that

$$\begin{aligned}
\mathbb{P}\{\|X^{\mathsf{T}}\varepsilon\|_\infty \ge \varepsilon n\} &= \mathbb{P}\Big\{\max_a \Big|\sum_i x_{ia}\varepsilon_i\Big| \ge \varepsilon n\Big\} \\
&\le \sum_a \mathbb{P}\Big\{\Big|\sum_i x_{ia}\varepsilon_i\Big| \ge \varepsilon n\Big\} \\
&\le 2p^{-6}.
\end{aligned}$$

This implies the claim for p large enough. □

The rest of the proof is standard, cf. [15] and is given below for the reader's convenience.

Proof of Theorem 4.2. Throughout we condition on the intersection of good events in Proposition H.1 and Lemma H.4, which happens with probability at least $1-2p^{-4}$. On this good event, the sample covariance $\widehat{\Sigma}$ is $(\phi_0/4, \mathrm{supp}(\theta_0))$-compatible and $\|X^{\mathsf{T}}\varepsilon\|_\infty \le 20\kappa\sigma\sqrt{n\log p} \le n\lambda_n/2$.

By optimality of $\widehat{\theta}^{\mathsf{L}}$:

$$\frac{1}{2}\|y - X\widehat{\theta}^{\mathsf{L}}\|^2 + \lambda_n\|\widehat{\theta}^{\mathsf{L}}\|_1 \le \frac{1}{2}\|y - X\theta_0\|^2 + \lambda_n\|\theta_0\|_1.$$

Using $y = X\theta_0 + \varepsilon$, the shorthand $\nu = \widehat{\theta}^{\mathsf{L}} - \theta_0$ and expanding the squares leads to

$$\begin{aligned}
\frac{1}{2}\langle \nu, \widehat{\Sigma}\nu\rangle &\le \frac{1}{n}\langle X^{\mathsf{T}}\varepsilon, \nu\rangle + \lambda_n(\|\theta_0\|_1 - \|\widehat{\theta}^{\mathsf{L}}\|_1) \\
&\le \frac{1}{n}\|\nu\|_1\|X^{\mathsf{T}}\varepsilon\|_\infty + \lambda_n(\|\theta_0\|_1 - \|\widehat{\theta}^{\mathsf{L}}\|_1) \\
&\le \lambda_n\Big\{\frac{1}{2}\|\nu\|_1 + \|\theta_0\|_1 - \|\widehat{\theta}^{\mathsf{L}}\|_1\Big\}. \qquad (106)
\end{aligned}$$

First we show that the error vector ν satisfies $\|\nu_{S_0^c}\|_1 \le 3\|\nu_{S_0}\|_1$, where $S_0 \equiv \mathrm{supp}(\theta_0)$. Note that $\|\widehat{\theta}^{\mathsf{L}}\|_1 = \|\theta_0 + \nu\|_1 = \|\theta_0 + \nu_{S_0}\|_1 + \|\nu_{S_0^c}\|_1$. By triangle inequality, therefore:

$$\begin{aligned}
\|\theta_0\|_1 - \|\widehat{\theta}^{\mathsf{L}}\|_1 &= \|\theta_0\|_1 - \|\theta_0 + \nu_{S_0}\|_1 - \|\nu_{S_0^c}\|_1 \\
&\le \|\nu_{S_0}\|_1 - \|\nu_{S_0^c}\|_1.
\end{aligned}$$

Combining this with the basic lasso inequality Eq.(106) we obtain

$$\begin{aligned}\frac{1}{2}\langle \nu, \widehat{\Sigma}\nu\rangle &\le \lambda_n\Big\{\frac{1}{2}\|\nu\|_1 + \|\nu_{S_0}\|_1 - \|\nu_{S_0^c}\|_1\Big\}\\ &= \frac{\lambda_n}{2}\Big\{3\|\nu_{S_0}\|_1 - \|\nu_{S_0^c}\|.\Big\}\end{aligned}$$

As $\widehat{\Sigma}$ is positive-semidefinite, the LHS above is non-negative, which implies $\|\nu_{S_0^c}\|_1 \le 3\|\nu_{S_0}\|_1$. Now, we can use the fact that $\widehat{\Sigma}$ is $(\phi_0/4, S_0)$-compatible to lower bound the LHS by $\|\nu\|_1^2\phi_0/2s_0$. This leads to

$$\frac{\phi_0\|\nu\|_1^2}{2s_0} \le \frac{3\lambda_n\|\nu_{S_0}\|_1}{2} \le \frac{3\lambda_n\|\nu\|_1}{2}.$$

Simplifying this results in $\|\nu\|_1 = \|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1 \le 3s_0\lambda_n/\phi_0$ as required.

□

H.2 Bias control: Proof of Theorem 4.7

Recall the decomposition (30) from which we obtain:

$$\begin{aligned}\Delta_n &= B_n(\widehat{\theta}^{\mathsf{L}} - \theta_0),\\ B_n &= \sqrt{n}\Big(I_p - \frac{n_1}{n}M^{(1)}\widehat{\Sigma}^{(1)} - \frac{n_2}{n}M^{(2)}\widehat{\Sigma}^{(2)}\Big),\\ W_n &= \frac{1}{\sqrt{n}}\sum_{i\le n_1} M^{(1)}x_i\varepsilon_i + \frac{1}{\sqrt{n}}\sum_{n_1 < i \le n} M^{(2)}x_i\varepsilon_i.\end{aligned}$$

By construction $M^{(1)}$ is a function of X_1 and hence is independent of $\varepsilon_1, \dots, \varepsilon_{n_1}$. In addition, $M^{(2)}$ is independent of $\varepsilon_{n_1+1}, \dots, \varepsilon_n$. Therefore $\mathbb{E}\{W_n\} = 0$ as required. The key is to show the bound on $\|\Delta_n\|_\infty$. We start by using Hölder inequality

$$\|\Delta_n\|_\infty \le \|B_n\|_\infty\|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1.$$

Since the ℓ_1 error of $\widehat{\theta}^{\mathsf{L}}$ is bounded in Theorem 4.2, we need only to show the bound on B_n. For this, we use triangle inequality and that $M^{(1)}$ and $M^{(2)}$ are feasible for the online debiasing program:

$$\begin{aligned}\|B_n\|_\infty &= \sqrt{n}\Big\|\frac{n_1}{n}(I_p - M^{(1)}\widehat{\Sigma}^{(1)}) + \frac{n_2}{n}(I_p - M^{(2)}\widehat{\Sigma}^{(2)})\Big\|_\infty\\ &\le \sqrt{n}\Big(\frac{n_1}{n}\|I_p - M^{(1)}\widehat{\Sigma}^{(1)}\|_\infty + \frac{n_2}{n}\|I_p - M^{(2)}\widehat{\Sigma}(2)\|_\infty\Big)\\ &\le \sqrt{n}\Big(\frac{n_1\mu_1}{n} + \frac{n_2\mu_2}{n}\Big).\end{aligned}$$

The following lemma shows that, with high probability, we can take μ_1, μ_2 so that the resulting bound on B_n is of order $\sqrt{\log p}$.

Lemma H.5. *Denote by* $\Omega = (\mathbb{E}\{xx^{\mathsf{T}}\})^{-1}$ *and* $\Omega^{(2)}(\widehat{\theta}) = (\mathbb{E}\{xx^{\mathsf{T}}|\langle x, \widehat{\theta}\rangle \geq \varsigma\})^{-1}$ *be the population precision matrices for the first and second batches. Suppose that* $n_1 \wedge n_2 \geq 2\Lambda_0/\kappa^2 \log p$*. Then, with probability at least* $1 - p^{-4}$

$$\|I_p - \Omega\widehat{\Sigma}^{(1)}\|_\infty \leq 15\kappa{\Lambda_0}^{-1/2}\sqrt{\frac{\log p}{n_1}},$$
$$\|I_p - \Omega^{(2)}\widehat{\Sigma}^{(2)}\|_\infty \leq 15\kappa{\Lambda_0}^{-1/2}\sqrt{\frac{\log p}{n_2}}.$$

In particular, with the same probability, the online debiasing program (28) *is feasible with* $\mu_\ell = 15\kappa^2{\Lambda_0}^{-1}\sqrt{(\log p)/n_\ell} < 1/2$*.*

It follows from the lemma, Theorem 4.2 and the previous display that, with probability at least $1 - 2p^{-3}$

$$\begin{aligned}
\|\Delta_n\|_\infty &\leq \|B_n\|_\infty \|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1 \\
&\leq 15\kappa\Lambda_0^{-1/2}\sqrt{n}\Big(\frac{n_1}{n}\sqrt{\frac{\log p}{n_1}} + \frac{n_2}{n}\sqrt{\frac{\log p}{n_2}}\Big) \cdot 120\kappa\sigma\phi_0^{-1}s_0\sqrt{\frac{\log p}{n}}, \\
&\leq 2000\frac{\kappa^2\sigma}{\sqrt{\Lambda_0}\phi_0}\frac{s_0\log p}{n}(\sqrt{n_1} + \sqrt{n_2}) \\
&\leq 4000\frac{\kappa^2\sigma}{\sqrt{\Lambda_0}\phi_0}\frac{s_0\log p}{\sqrt{n}}. \qquad (107)
\end{aligned}$$

This implies the first claim that, with probability rapidly converging to one, $\Delta_n/\sqrt{n}$ is of order $s_0 \log p/n$.

We should also expect $\|\mathbb{E}\{\widehat{\theta}^{\mathsf{on}} - \theta_0\}\|_\infty$ to be of the same order. To prove this, however, we need some control (if only rough) on $\widehat{\theta}^{\mathsf{on}}$ in the exceptional case when the LASSO error is large or the online debiasing program is infeasible. Let G_1 denote the good event of Lemma H.4 and G_2 denote the good event of Theorem 4.2 as below:

$$G_1 = \Bigg\{ \text{ For } \ell = 1, 2 : \|I_p - \Omega^{(\ell)}\widehat{\Sigma}^{(\ell)}\|_\infty \leq 15\kappa{\Lambda_0}^{-1/2}\sqrt{\frac{\log p}{n_\ell}}\Bigg\},$$
$$G_2 = \Bigg\{\|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1 \leq \frac{3s_0\lambda_n}{\phi_0} = \frac{120\kappa\sigma}{\phi_0}s_0\sqrt{\frac{\log p}{n}}.\Bigg\}.$$

On the intersection $G = G_1 \cap G_2$, Δ_n satisfies the bound (107). For the complement: we will use the following rough bound on the LASSO error:

Now, since W_n is unbiased:

$$\begin{aligned}\|\mathbb{E}\{\widehat{\theta}^{\mathsf{on}} - \theta_0\}\|_\infty &= \Big\|\frac{\mathbb{E}\{\Delta_n\}}{\sqrt{n}}\Big\|_\infty \\ &= \Big\|\frac{\mathbb{E}\{\Delta_n \mathbb{I}(G)\}}{\sqrt{n}}\Big\|_\infty + \Big\|\frac{\mathbb{E}\{\Delta_n \mathbb{I}(G^c)\}}{\sqrt{n}}\Big\|_\infty \\ &\leq 4000 \frac{\kappa^2\sigma}{\sqrt{\Lambda_0}\phi_0}\frac{s_0\log p}{n} + \mathbb{E}\{\|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1 \mathbb{I}(G^c)\}.\end{aligned}$$

For the second term, we can use Lemma I.7, Cauchy Schwarz and that $\mathbb{P}\{G^c\} \leq 4p^{-3}$ to obtain:

$$\begin{aligned}\mathbb{E}\{\|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1 \mathbb{I}(G^c)\} &\leq \mathbb{E}\Big\{\frac{\|\varepsilon\|^2\mathbb{I}(G^c)}{2n\lambda_n} + 2\|\theta_0\|_1\mathbb{I}(G^c)\Big\} \\ &\leq \frac{\mathbb{E}\{\|\varepsilon\|^4\}^{1/2}\mathbb{P}(G^c)^{1/2}}{2n\lambda_n} + 2\|\theta_0\|_1\mathbb{P}\{G^c\} \\ &\leq \frac{\sqrt{3}\sigma^2}{\sqrt{n}p^{1.5}\lambda_n} + 8\|\theta_0\|_1 p^{-3} \leq 10c\frac{s_0\log p}{n},\end{aligned}$$

for n, p large enough . This implies the claim on the bias.

It remains only to prove the intermediate Lemma H.5.

Proof of Lemma H.5. We prove the claim for the second batch, and in the rest of the proof, we assume that all probabilities and expectations are conditional on the first batch (in particular, the intermediate estimate $\widehat{\theta}^1$). The (a, b) entry of $I_p - \Omega^{(2)}\widehat{\Sigma}^{(2)}$ reads

$$\begin{aligned}(I_p - \Omega^{(2)}\widehat{\Sigma}^{(2)})_{a,b} &= \mathbb{I}(a = b) - \langle\Omega^{(2)}e_a, \widehat{\Sigma}^{(2)}e_b\rangle \\ &= \frac{1}{n_2}\sum_{i>n_1}\mathbb{I}(a = b) - \langle e_a, \Omega^{(2)}x_i\rangle x_{ib}.\end{aligned}$$

Now, $\mathbb{E}\{\langle e_a, \Omega^{(2)}x_i\rangle x_{i,b}\rangle\} = \mathbb{I}(a = b)$ and $\langle e_a, \Omega^{(2)}x_i\rangle$ is $(\|\Omega^{(2)}\|_2\kappa)$-subgaussian. Since $\Sigma^{(2)} \succcurlyeq \Lambda_0 I_p$, we have that $\|\Omega^{(2)}\|_2 \leq {\Lambda_0}^{-1}$. This observation, coupled with Lemma I.4, yields $\langle e_a, \Omega^{(2)}x_i\rangle x_{i,b}$ is $2\kappa^2/\Lambda_0$-subexponential. Then we may apply Lemma I.5 for $\varepsilon \leq 12\kappa^2/\Lambda_0$ as below:

$$\mathbb{P}\{(I_p - \Omega^{(2)}\widehat{\Sigma}^{(2)})_{a,b} \geq \varepsilon\} \leq \exp\Big(-\frac{n_2\varepsilon^2}{36\kappa^2{\Lambda_0}^{-1}}\Big).$$

Keeping $\varepsilon = \varepsilon(p, n_2, \kappa, \Lambda_0) = 15\kappa{\Lambda_0}^{-1/2}\sqrt{(\log p)/n_2}$ we obtain:

$$\mathbb{P}\Big\{(I_p - \Omega^{(2)}\widehat{\Sigma}^{(2)})_{a,b} \geq 15\kappa{\Lambda_0}^{-1/2}\sqrt{\frac{\log p}{n_2}}\Big\} \leq p^{-6}.$$

Union bounding over the pairs (a, b) yields the claim. The requirement $n_2 \geq 2(\Lambda_0/\kappa^2)\log p$ ensures that the choice ε above satisfies $\varepsilon \leq 12\kappa^2/\Lambda_0$.

□

H.3 Central limit asymptotics: proofs of Proposition 4.9 and Theorem 4.10

Our approach is to apply a martingale central limit theorem to show that $W_{n,a}$ is approximately normal. An important first step is to show that the conditional covariance $V_{n,a}$ is stable, or approximately constant. Recall that $V_{n,a}$ is defined as

$$V_{n,a} = \sigma^2\Big(\frac{n_1}{n}\langle m_a^{(1)}, \widehat{\Sigma}^{(1)}m_a^{(1)}\rangle + \frac{n_2}{n}\langle m_a^{(2)}, \widehat{\Sigma}^{(2)}m_a^{(2)}\rangle\Big).$$

We define its deterministic equivalent as follows. Consider the function $f : \mathbb{S}^n \to \mathbb{R}$ by:

$$f(\Sigma) = \{\min\,\langle m, \Sigma m\rangle : \|\Sigma m - e_a\|_\infty \leq \mu\,,\ \ \|m\|_1 \leq L\}.$$

We begin with two lemmas about the stability of the optimization program used to obtain the online debiasing matrices.

Lemma H.6. *On its domain (and uniformly in μ, e_a), f is L^2-Lipschitz with respect to the $\|\cdot\|_\infty$ norm.*

Proof. For two matrices Σ, Σ' in the domain, let m, m' be the respective optimizers (which exist by compactness of the set $\{m : \|\Sigma m - v\|_\infty \leq \mu, \|m\|_1 \leq L\}$. We prove that $|f(\Sigma) - f(\Sigma')| \leq L^2\|\Sigma - \Sigma'\|_\infty$.

$$\begin{aligned} f(\Sigma) - f(\Sigma') &= \langle \Sigma, mm^{\mathsf{T}}\rangle - \langle \Sigma', m'(m')^{\mathsf{T}}\rangle \\ &\leq \langle \Sigma, m'(m')^{\mathsf{T}}\rangle - \langle \Sigma', m'(m')^{\mathsf{T}}\rangle \\ &= \langle (\Sigma - \Sigma')m', m'\rangle \\ &\leq \|(\Sigma - \Sigma')m'\|_\infty \|m'\|_1 \\ &\leq \|\Sigma - \Sigma'\|_\infty \|m'\|_1^2 \leq L^2\|\Sigma - \Sigma'\|_\infty. \end{aligned}$$

Here the first inequality follows from optimality of m and the last two inequalities are Hölder inequality. The reverse inequality $f(\Sigma) - f(\Sigma') \geq -L^2\|\Sigma - \Sigma'\|_\infty$ is proved in the same way. □

Lemma H.7. *We have the following lower bound on the optimization value reached to compute $f(\Sigma)$:*

$$\frac{(1-\mu)^2}{\lambda_{\max}(\Sigma)} \leq f(\Sigma) \leq \frac{1}{\lambda_{\min}(\Sigma)}.$$

Proof. We first prove the lower bound for $f(\Sigma)$. Suppose m is an optimizer for the program. Then

$$\|\Sigma m\|_2 \geq \|\Sigma m\|_\infty \geq \|e_a\|_\infty - \mu = 1 - \mu.$$

On the other hand, the value is given by

$$\langle m, \Sigma m\rangle = \langle \Sigma m, \Sigma^{-1}(\Sigma m)\rangle \geq \lambda_{\min}(\Sigma^{-1})\|\Sigma m\|_2^2 = \|\Sigma m\|_2^2\, \lambda_{\max}(\Sigma)^{-1}.$$

Combining these gives the lower bound.

For the upper bound, it suffices to consider any feasible point; we choose $m = \Sigma^{-1}e_a$, which is feasible since $\|\Sigma^{-1}\|_1 \leq L$. The value is then $\langle e_a, \Sigma^{-1}e_a\rangle \leq \lambda_{\max}(\Sigma^{-1})$ which gives the upper bound. □

Lemma H.8. *(Stability of $W_{n,a}$) Define $\Sigma^{(2)}(\theta) = \mathbb{E}\{xx^{\mathsf{T}}|\langle x_1, \theta\rangle \geq \varsigma\}$. Then, under Assumptions 4.5 and 4.8*

$$\lim_{n\to\infty}\Big|V_{n,a} - \sigma^2\Big(\frac{n_1 f(\Sigma)}{n} + \frac{n_2 f(\Sigma^2(\theta_0))}{n}\Big)\Big| = 0, \qquad \textit{in probability.}$$

Proof. Using Lemma H.6:

$$\begin{aligned}
&\Big|V_{n,a} - \sigma^2\Big(\frac{n_1}{n}f(\Sigma) + \frac{n_2}{n}f(\Sigma(\theta_0)\Big)\Big| \\
&= \frac{\sigma^2 n_1}{n}(f(\widehat{\Sigma}^{(1)}) - f(\Sigma)) + \frac{\sigma^2 n_2}{n}(f(\widehat{\Sigma}^{(2)} - f(\Sigma(\theta_0)))) \\
&\leq L^2\frac{\sigma^2 n_1}{n}\|\Sigma - \widehat{\Sigma}^{(1)}\|_\infty + L^2\frac{\sigma^2 n_2}{n}\|\Sigma^{(2)}(\theta_0) - \widehat{\Sigma}^{(2)}\|_\infty \\
&\leq L^2\frac{\sigma^2 n_1}{n}\|\Sigma - \widehat{\Sigma}^{(1)}\|_\infty + L^2\frac{\sigma^2 n_2}{n}\big(\|\Sigma^{(2)}(\theta_0) - \Sigma^{(2)}(\widehat{\theta}^1)\|_\infty + \|\Sigma^{(2)}(\widehat{\theta}^1) - \widehat{\Sigma}^{(2)}\|_\infty\big) \\
&\leq \sigma^2 L^2\|\Sigma - \widehat{\Sigma}^{(1)}\|_\infty + \sigma^2 L^2\big(K\|\widehat{\theta}^1 - \theta_0\|_1 + \|\Sigma^{(2)}(\widehat{\theta}^1) - \widehat{\Sigma}^{(2)}\|_\infty\big)\,.
\end{aligned}$$

Using Lemma H.2 the first and third term vanish in probability. It is straightforward to apply Theorem 4.2 to the intermediate estimate $\widehat{\theta}^1$; indeed Assumption 4.8 guarantees that $n_1 \geq cn$ for a universal c. Therefore the intermediate estimate has an error $\|\widehat{\theta}^1 - \theta_0\|_1$ of order $\kappa\sigma\phi_0^{-1}\sqrt{(s_0^2 \log p)/n}$ with probability converging to one. In particular, the second term is, with probability converging to one, of order $KL^2\sigma^3\kappa\phi_0^{-1}\sqrt{s_0^2(\log p)/n} = o(1)$ by Assumption 4.8. □

Lemma H.9. *Under Assumptions 4.5 and 4.8, with probability at least* $1 - p^{-2}$

$$\max_i |\langle m_a, x_i\rangle| \leq 10L\kappa\sqrt{\log p},$$

In particular $\lim_{n\to\infty} \max_i |\langle m_a, x_i\rangle| = 0$ *in probability.*

Proof. By Hölder inequality, $\max_i\langle|\langle m_a, x_i\rangle| \leq \max_i \|m_a\|_1\|x_i\|_\infty \leq L\max_i \|x_i\|_\infty$. Therefore, it suffices to prove that, with the required probability $\max_{i,a}|x_{i,a}| \leq 10\kappa\sqrt{\log p}$. Let $u = 10\kappa\sqrt{\log p}$. Since x_i are uniformly κ-subgaussian, we obtain for $q > 0$:

$$\begin{aligned}\mathbb{P}\{|x_{i,a}| \geq u\} &\leq u^{-q}\mathbb{E}\{|x_{i,a}|^q\} \leq (\sqrt{q}\kappa/u)^q \\ &= \exp\Big(-\frac{q}{2}\log\frac{u^2}{\kappa^2 q}\Big) \leq \exp\Big(-\frac{u^2}{2\kappa^2}\Big) \leq p^{-5},\end{aligned}$$

where the last line follows by choosing $q = u^2/e\kappa^2$. By union bound over $i \in [n], a \in [p]$, we obtain:

$$\mathbb{P}\{\max_{i,a}|x_{i,a}| \geq u\} \leq \sum_{i,a}\mathbb{P}\{|x_{i,a}| \geq u\} \leq p^{-3},$$

which implies the claim (note that $p \geq n$ as we are focusing on the high-dimensional regime). □

With these in hand we can prove Proposition 4.9 and Theorem 4.10.

Proof of Proposition 4.9. Consider the minimal filtration $\mathfrak{F}_i$ so that

1. For $i < n_1$, $y_1, \dots, y_i$, $x_1, \dots x_{n_1}$ and $\varepsilon_1, \dots, \varepsilon_i$ are measurable with respect to $\mathfrak{F}_i$.

2. For $i \geq n_1$ $y_1, \dots, y_i$, $x_1, \dots, x_n$ and $\varepsilon_1, \dots \varepsilon_i$ are measurable with respect to $\mathfrak{F}_i$.

The martingale W_n (and therefore, its $a^{\rm th}$ coordinate $W_{n,a}$) is adapted to the filtration $\mathfrak{F}_i$. We can now apply the martingale central limit theorem [14, Corollary 3.1] to $W_{n,a}$ to obtain the result. From Lemmas H.7 and H.8 we know that $V_{n,a}$ is bounded away from 0, asymptotically. The stability and conditional Lindeberg conditions of [14, Corollary 3.1] are verified by Lemmas H.8 and H.9. □

Proof of Theorem 4.10. This is a straightforward corollary of the bias bound of 4.7 and Proposition 4.9. We will show that:

$$\lim_{n\to\infty} \mathbb{P}\Big\{\sqrt{\frac{n}{V_{n,a}}}(\widehat{\theta}^{\sf on}_a - \theta_{0,a}) \le x\Big\} \le \Phi(x).$$

The reverse inequality follows using the same argument.

Fix a $\delta > 0$. We decompose the difference above as:

$$\sqrt{\frac{n}{V_{n,a}}}(\widehat{\theta}^{\sf on}_a - \theta_{0,a}) = \frac{W_{n,a}}{\sqrt{V_{n,a}}} + \frac{\Delta_{n,a}}{\sqrt{V_{n,a}}}\,.$$

Therefore,

$$\mathbb{P}\Big\{\sqrt{\frac{n}{V_{n,a}}}(\widehat{\theta}^{\sf on}_a - \theta_{0,a}) \le x\Big\} \le \mathbb{P}\Big\{\frac{W_{n,a}}{\sqrt{V_{n,a}}} \le x + \delta\Big\} + \mathbb{P}\{|\Delta_{n,a}| \ge \sqrt{V_{n,a}}\delta\}.$$

By Proposition 4.9 the first term converges to $\Phi(x+\delta)$. To see that the second term vanishes, observe first that Lemma H.7 and Lemma H.8, imply that $V_{n,a}$ is bounded away from 0 in probability. Using this:

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}\{|\Delta_{n,a}| \ge \sqrt{V_{n,a}}\delta\} &\le \lim_{n\to\infty} \mathbb{P}\{\|\Delta_n\|_\infty \ge \sqrt{V_{n,a}}\delta\} \\ &\le \lim_{n\to\infty} \mathbb{P}\Big\{\|\Delta_n\|_\infty \ge 4000\frac{\kappa^2\sigma}{\sqrt{\Lambda_0}\phi_0}\frac{s_0\log p}{\sqrt{n}}\Big\} = 0\end{aligned}$$

by applying Theorem 4.7 and that for n large enough, $\sqrt{V_{n,a}}\delta$ exceeds the bound on $\|\Delta_n\|_\infty$ used. Since δ is arbitrary, the claim follows. □

H.4 Proofs for Gaussian designs

In this Section we prove that Gaussian designs of Example 4.6 satisfy the requirements of Theorem 4.2 and Theorem 4.7.

The following distributional identity will be important.

Lemma H.10. *Consider the parametrization* $\varsigma = \bar{\varsigma}\langle\widehat{\theta}, \Sigma\widehat{\theta}\rangle^{1/2}$*. Then*

$$x|_{\langle x,\widehat{\theta}\rangle\geq\varsigma} \stackrel{\mathrm{d}}{=} \frac{\Sigma\widehat{\theta}}{\langle\widehat{\theta}, \Sigma\widehat{\theta}\rangle^{1/2}}\xi_1 + \Big(\Sigma - \frac{\Sigma\widehat{\theta}\widehat{\theta}^{\mathsf{T}}\Sigma}{\langle\widehat{\theta}, \Sigma\widehat{\theta}\rangle}\Big)^{1/2}\xi_2,$$

where ξ_1, ξ_2 *are independent,* $\xi_2 \sim \mathsf{N}(0, I_p)$ *and* ξ_1 *has the density:*

$$\frac{\mathrm{d}\mathbb{P}_{\xi_1}}{\mathrm{d}u}(u) = \frac{1}{\sqrt{2\pi}\Phi(-\bar{\varsigma})}\exp(-u^2/2)\mathbb{I}(u \geq \bar{\varsigma}).$$

Proof. This follows from the distribution of $x|\langle x, \widehat{\theta}\rangle$ being $\mathsf{N}(\mu', \Sigma')$ with

$$\mu' = \frac{\Sigma\widehat{\theta}}{\langle\widehat{\theta}, \Sigma\widehat{\theta}\rangle}\langle x, \widehat{\theta}\rangle, \quad \Sigma' = \Sigma - \frac{\Sigma\widehat{\theta}\widehat{\theta}^{\mathsf{T}}\Sigma}{\langle\widehat{\theta}, \Sigma\widehat{\theta}\rangle}.$$

□

The following lemma shows that they satisfy compatibility.

Lemma H.11. *Let* $\mathbb{P}_x = \mathsf{N}(0, \Sigma)$ *for a positive definite covariance* Σ*. Then, for any vector* $\widehat{\theta}$ *and subset* $S \subseteq [p]$*, the second moments* $\mathbb{E}\{xx^{\mathsf{T}}\}$ *and* $\mathbb{E}\{xx^{\mathsf{T}}|\langle x, \widehat{\theta}\rangle \geq \varsigma\}$ *are* (ϕ_0, S)*-compatible with* $\phi_0 = \lambda_{\min}(\Sigma)/16$*.*

Proof. Fix an $S \subseteq [p]$. We prove that $\Sigma = \mathbb{E}\{x_1x_1^{\mathsf{T}}\}$ is (ϕ_0, S)-compatible with $\phi_0 = \lambda_{\min}(\Sigma)/16$. Note that, for any v satisfying $\|v_{S^c}\|_1 \leq 3\|v_S\|$, its ℓ_1 norm satisfies $\|v\|_1 \leq 4\|v_S\|_1$. Further $\Sigma \succcurlyeq \lambda_{\min}(\Sigma)I_p$ implies:

$$\frac{|S|\langle v, \Sigma v\rangle}{\|v\|_1^2} \geq \lambda_{\min}(\Sigma)\frac{|S|\|v\|^2}{\|v\|_1^2} \geq \lambda_{\min}(\Sigma)\frac{|S|\|v_S\|^2}{16\|v_S\|_1^2} \geq \frac{\lambda_{\min}(\Sigma)}{16}.$$

For $\mathbb{E}\{xx^{\mathsf{T}}|\langle x, \widehat{\theta}\rangle \geq \varsigma\}$, we use Lemma H.10 to obtain

$$\mathbb{E}\{xx^{\mathsf{T}}|\langle x, \widehat{\theta}\rangle \geq \varsigma\} = \Sigma + (\mathbb{E}\{\xi_1^2\} - 1)\frac{\Sigma\widehat{\theta}\widehat{\theta}^{\mathsf{T}}\Sigma}{\langle\widehat{\theta}, \Sigma\widehat{\theta}\rangle},$$

where ξ_1 is as in Lemma H.10. Since $\mathbb{E}\{\xi_1^2\} = 1 + \bar{\varsigma}\varphi(\bar{\varsigma})/\Phi(-\bar{\varsigma}) \geq 1 + \bar{\varsigma}^2$ whenever $\bar{\varsigma} \geq 0$:

$$\mathbb{E}\{xx^{\mathsf{T}}|\langle x, \widehat{\theta}\rangle \geq \varsigma\} \geq \Sigma + \bar{\varsigma}^2\frac{\Sigma\widehat{\theta}\widehat{\theta}^{\mathsf{T}}\Sigma}{\langle\widehat{\theta}, \Sigma\widehat{\theta}\rangle} \succcurlyeq \lambda_{\min}(\Sigma)I_p\,.$$

The rest of the proof is as for Σ.

□

Lemma H.12. *Let $\mathbb{P}_x = \mathsf{N}(0, \Sigma)$ for a positive definite covariance Σ. Then, for any vector $\widehat{\theta}$ and subset $S \subseteq [p]$, the random vectors x and $x|_{\langle x, \widehat{\theta}\rangle \geq \varsigma}$ are κ-subgaussian with $\kappa = 3\lambda_{\max}(\Sigma)^{1/2}(\bar{\varsigma} \vee \bar{\varsigma}^{-1})$, where $\bar{\varsigma} = \varsigma / \langle \widehat{\theta}, \Sigma\widehat{\theta}\rangle^{1/2}$.*

Proof. By definition, $\langle x, v\rangle \sim \mathsf{N}(0, v^{\mathsf{T}}\Sigma v)$ is $\sqrt{v^{\mathsf{T}}\Sigma v}$-subGaussian. Optimizing over all unit vectors v, x is $\lambda_{\max}^{1/2}(\Sigma)$-subgaussian.

For $x|_{\langle x, \widehat{\theta}\rangle \geq \varsigma}$, we use the decomposition of Lemma H.10:

$$x|_{\langle x, \widehat{\theta}\rangle \geq \varsigma} \overset{d}{=} \frac{\Sigma\widehat{\theta}}{\langle \widehat{\theta}, \Sigma\widehat{\theta}\rangle^{1/2}}\xi_1 + \Big(\Sigma - \frac{\Sigma\widehat{\theta}\widehat{\theta}^{\mathsf{T}}\Sigma}{\langle \widehat{\theta}, \Sigma\widehat{\theta}\rangle}\Big)^{1/2}\xi_2.$$

Clearly, ξ_2 is 1-subgaussian, which means the second term is $\lambda_{\max}^{1/2}(\Sigma)$-subgaussian. For the first term, we claim that ξ_1 is 1-subgaussian and therefore the first term is $\lambda_{\max}^{1/2}(\Sigma)$-subgaussian. To show this, we start with the moment generating function of ξ_1. Recall that $\bar{\varsigma} = \varsigma / \langle \widehat{\theta}, \Sigma\widehat{\theta}\rangle^{1/2}$:

$$\mathbb{E}\{e^{\lambda\xi_1}\} = \int_{\bar{\varsigma}}^{\infty} e^{\lambda u} e^{-u^2/2} \frac{\mathrm{d}u}{\sqrt{2\pi}\Phi(-\bar{\varsigma})} = e^{\lambda^2/2}\frac{\Phi(\lambda - \bar{\varsigma})}{\Phi(-\bar{\varsigma})}.$$

Here φ and Φ are the density and c.d.f. of the standard normal distribution. It follows that:

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\log\mathbb{E}\{e^{\lambda\xi_1}\} &= \frac{1}{2} + \frac{(\lambda - \bar{\varsigma})\varphi(\lambda - \bar{\varsigma})}{\Phi(\lambda - \bar{\varsigma})} - \frac{\varphi(\lambda - \bar{\varsigma})^2}{\Phi(\lambda - \bar{\varsigma})^2} \\
&\leq \frac{1}{2} + \sup_{\lambda \geq \bar{\varsigma}} \frac{(\lambda - \bar{\varsigma})\varphi(\lambda - \bar{\varsigma})}{\Phi(\lambda - \bar{\varsigma})} \\
&\leq \frac{1}{2} + \sup_{\lambda \geq 0} \frac{\lambda\varphi(\lambda)}{\Phi(\lambda)} < 1\,.
\end{aligned}$$

Now, consider the centered version $\xi_1' = \xi_1 - \mathbb{E}\{\xi_1\}$. The above bound also holds for $\mathrm{d}^2/\mathrm{d}\lambda^2(\log\mathbb{E}\{e^{\lambda\xi_1'}\})$. Therefore, by integration, $\mathrm{d}\log\mathbb{E}\{e^{\lambda\xi_1'}\}/\mathrm{d}\lambda \leq \lambda + C$, for some constant C independent of λ. Now

$$\frac{\mathrm{d}\log\mathbb{E}\{e^{\lambda\xi_1'}\}}{\mathrm{d}\lambda}\Big|_{\lambda=0} = \mathbb{E}\{\xi_1'\} = 0.$$

Therefore, we can take the constant C to be 0. Repeating this integration argument, we obtain $\log\mathbb{E}\{e^{\lambda\xi_1'}\} \leq \lambda^2/2$, which implies that $\xi_1' = \xi_1 - \mathbb{E}\{\xi_1\}$ is 1-subgaussian.

It follows, by triangle inequality, that ξ_1 is $(1+\mathbb{E}\{\xi_1\})$-subgaussian. It only remains to bound $\mathbb{E}\{\xi_1\}$ as below:

$$\mathbb{E}\{\xi_1\} = \frac{\varphi(\bar{\varsigma})}{\Phi(-\bar{\varsigma})} \leq \frac{1+\bar{\varsigma}^2}{\bar{\varsigma}} \leq 2(\bar{\varsigma} \vee \bar{\varsigma}^{-1}).$$

Therefore, the subgaussian constant of $x|_{\langle x,\widehat{\theta}\rangle \geq \varsigma}$ is at most $\lambda_{\max}(\Sigma)^{1/2}(2\bar{\varsigma} \vee \bar{\varsigma}^{-1} + 1) \leq 3\lambda_{\max}(\Sigma)^{1/2}(\bar{\varsigma} \vee \bar{\varsigma}^{-1})$.

□

For Example 4.6, it remains only to show the constraint on the approximate sparsity of the inverse covariance. We show this in the following

Lemma H.13. *Let $\mathbb{P}_x = \mathsf{N}(0,\Sigma)$ and $\widehat{\theta}$ be any vector such that $\|\widehat{\theta}\|_1\|\widehat{\theta}\|_\infty \leq L\lambda_{\min}(\Sigma)\|\widehat{\theta}\|^2/2$ and $\|\Sigma^{-1}\|_1 \leq L/2$. Then, with $\Omega = \mathbb{E}\{xx^{\mathsf{T}}\}^{-1}$ and $\Omega^{(2)}(\widehat{\theta}) = \mathbb{E}\{xx^{\mathsf{T}}|\langle x,\widehat{\theta}\rangle \geq \varsigma\}^{-1}$:*

$$\|\Omega\|_1 \vee \|\Omega^{(2)}\|_1 \leq L.$$

Proof. By assumption $\|\Omega\|_1 \leq L/2$, so we only require to prove the claim for $\Omega^{(2)} = \mathbb{E}\{xx^{\mathsf{T}}|\langle x,\widehat{\theta}\rangle \geq \varsigma\}^{-1}$. Using Lemma H.10, we can compute the precision matrix:

$$\begin{aligned}\Omega^{(2)} &= \mathbb{E}\{xx^{\mathsf{T}}|\langle x,\widehat{\theta}\rangle \geq \varsigma\}^{-1} \\ &= \Big(\Sigma + (\mathbb{E}\{\xi_1^2\}-1)\frac{\Sigma\widehat{\theta}\widehat{\theta}^{\mathsf{T}}\Sigma}{\langle\widehat{\theta},\Sigma\widehat{\theta}\rangle}\Big)^{-1} \\ &= \Omega + (\mathbb{E}\{\xi_1^2\}^{-1}-1)\frac{\widehat{\theta}\widehat{\theta}^{\mathsf{T}}}{\langle\widehat{\theta},\Sigma\widehat{\theta}\rangle}\,,\end{aligned}$$

where the last step follows by an application of Sherman–Morrison formula. Since $\mathbb{E}\{\xi_1^2\} = 1 + \bar{\varsigma}\varphi(\bar{\varsigma})/\Phi(-\bar{\varsigma})$, where $\bar{\varsigma} = \varsigma/\langle\widehat{\theta},\Sigma\widehat{\theta}\rangle^{1/2}$ this yields:

$$\Omega^{(2)} = \Omega - \frac{\bar{\varsigma}\varphi(\bar{\varsigma})}{\Phi(-\bar{\varsigma})+\bar{\varsigma}\varphi(\bar{\varsigma})}\frac{\widehat{\theta}\widehat{\theta}^{\mathsf{T}}}{\langle\widehat{\theta},\Sigma\widehat{\theta}\rangle}.$$

By triangle inequality, for any $\bar{\varsigma} \geq 0$:

$$\begin{aligned}\|\Omega^{(2)}\|_1 &\leq \|\Omega\|_1 + \frac{\|\widehat{\theta}\widehat{\theta}^{\mathsf{T}}\|_1}{\langle\widehat{\theta},\Sigma\widehat{\theta}\rangle} \\ &\leq \frac{L}{2} + \frac{\|\widehat{\theta}\|_1\|\widehat{\theta}\|_\infty}{\lambda_{\min}(\Sigma)\|\widehat{\theta}\|^2} \leq L.\end{aligned}$$

□

Next we show that the conditional covariance of x is appropriately Lipschitz.

Lemma H.14. *Suppose* $\varsigma = \bar{\varsigma}\langle\theta, \Sigma\theta\rangle^{1/2}$ *for a constant* $\bar{\varsigma} \geq 0$. *Then The conditional covariance function* $\Sigma^{(2)}(\theta) = \mathbb{E}\{xx^{\mathsf{T}} | \langle x, \theta\rangle \geq \varsigma\}$ *satisfies:*

$$\|\Sigma^{(2)}(\theta') - \Sigma^{(2)}(\theta)\|_\infty \leq K\|\theta' - \theta\|,$$

where $K = \sqrt{8}(1 + \bar{\varsigma}^2)\lambda_{\max}(\Sigma)^{3/2}/\lambda_{\min}(\Sigma)^{1/2}$.

Proof. Using Lemma H.10,

$$\Sigma^{(2)}(\theta) = \Sigma + (\mathbb{E}\{\xi_1^2\} - 1)\frac{\Sigma\theta\theta^{\mathsf{T}}\Sigma}{\langle\theta, \Sigma\theta\rangle}.$$

Let $v = \Sigma^{1/2}\theta/\|\Sigma^{1/2}\theta\|$ and $v' = \Sigma^{1/2}\theta'/\|\Sigma^{1/2}\theta'\|$. With this,

$$\begin{aligned}
\|\Sigma^{(2)}(\theta') - \Sigma^{(2)}(\theta)\|_\infty &= (\mathbb{E}\{\xi_1^2\} - 1)\|\Sigma^{1/2}(vv^{\mathsf{T}} - v'v'^{\mathsf{T}})\Sigma^{1/2}\|_\infty \\
&\leq (\mathbb{E}\{\xi_1^2\} - 1)\,\lambda_{\max}(\Sigma)\|vv^{\mathsf{T}} - v'v'^{\mathsf{T}}\|_2 \\
&\leq (\mathbb{E}\{\xi_1^2\} - 1)\lambda_{\max}(\Sigma)\|vv^{\mathsf{T}} - v'v'^{\mathsf{T}}\|_F \\
&\overset{(a)}{\leq} \sqrt{2}(\mathbb{E}\{\xi_1^2\} - 1)\lambda_{\max}(\Sigma)\|v - v'\| \\
&\overset{(b)}{\leq} \frac{\sqrt{8}\lambda_{\max}(\Sigma)^{3/2}}{\lambda_{\min}(\Sigma)^{1/2}}(\mathbb{E}\{\xi_1^2\} - 1)\|\theta - \theta'\| \\
&\overset{(c)}{\leq} \frac{\sqrt{8}\lambda_{\max}(\Sigma)^{3/2}}{\lambda_{\min}(\Sigma)^{1/2}}(\bar{\varsigma}^2 + 1)\|\theta - \theta'\|\,.
\end{aligned}$$

Here, (a) follows by noting that for two unit vectors v, v', we have

$$\|vv^{\mathsf{T}} - v'v'^{\mathsf{T}}\|_F^2 = 2 - 2(v^{\mathsf{T}}v')^2 = 2(1 - v^{\mathsf{T}}v')(1 + v^{\mathsf{T}}v') \leq 2\|v - v'\|^2\,.$$

Also, (b) holds using the following chain of triangle inequalities

$$\begin{aligned}
\|v - v'\| &= \Big\|\frac{\Sigma^{1/2}\theta}{\|\Sigma^{1/2}\theta\|} - \frac{\Sigma^{1/2}\theta'}{\|\Sigma^{1/2}\theta'\|}\Big\| \\
&\leq \frac{\|\Sigma^{1/2}(\theta - \theta')\|}{\|\Sigma^{1/2}\theta\|} + \|\Sigma^{1/2}\theta'\|\Big|\frac{1}{\|\Sigma^{1/2}\theta\|} - \frac{1}{\|\Sigma^{1/2}\theta'\|}\Big| \\
&\leq 2\frac{\|\Sigma^{1/2}(\theta - \theta')\|}{\|\Sigma^{1/2}\theta\|} \leq 2\sqrt{\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}}\,\|\theta - \theta'\|
\end{aligned}$$

Finally (c) holds since

$$\mathbb{E}\{\xi_1^1\} - 1 = \bar{\varsigma}\varphi(\bar{\varsigma})/\Phi(-\bar{\varsigma}) \leq \bar{\varsigma}^2 + 1\,,$$

using standard tail bound $\varphi(\bar{\varsigma})\frac{\bar{\varsigma}}{\bar{\varsigma}^2+1} \leq \Phi(-\bar{\varsigma})$. □

I Technical preliminaries

Definition I.1. *(Subgaussian norm) The subgaussian norm of a random variable X, denoted by $\|X\|_{\psi_2}$, is defined as*

$$\|X\|_{\psi_2} \equiv \sup_{q\geq 1} q^{-1/2}\mathbb{E}\{|X|^q\}^{1/q}.$$

For a random vector X the subgaussian norm is defined as

$$\|X\|_{\psi_2} \equiv \sup_{\|v\|=1} \|\langle X, v\rangle\|_{\psi_2}.$$

Definition I.2. *(Subexponential norm) The subexponential norm of a random variable X is defined as*

$$\|X\|_{\psi_1} \equiv \sup_{q\geq 1} q^{-1}\mathbb{E}\{|X|^q\}^{1/q}.$$

For a random vector X the subexponential norm is defined by

$$\|X\|_{\psi_1} \equiv \sup_{\|v\|=1} \|\langle X, v\rangle\|_{\psi_1}.$$

Definition I.3. *(Uniformly subgaussian/subexponential sequences) We say a sequence of random variables $\{X_i\}_{i\geq 1}$ adapted to a filtration $\{\mathcal{F}_i\}_{i\geq 0}$ is* uniformly K-subgaussian *if, almost surely:*

$$\sup_{i\geq 1}\sup_{q\geq 1} q^{-1/2}\mathbb{E}\{|X_i|^q|\mathcal{F}_{i-1}\}^{1/q} \leq K.$$

A sequence of random vectors $\{X_i\}_{i\geq 1}$ is uniformly K-subgaussian if, almost surely,

$$\sup_{i\geq 1}\sup_{\|v\|=1}\sup_{q\geq 1} \mathbb{E}\{|\langle X_i, v\rangle|^q|\mathcal{F}_{i-1}\}^{1/q} \leq K.$$

Subexponential sequences are defined analogously, replacing the factor $q^{-1/2}$ with q^{-1} above.

Lemma I.4. *For a pair of random variables X, Y, $\|XY\|_{\psi_1} \leq 2\|X\|_{\psi_2}\|Y\|_{\psi_2}$.*

Proof. By Cauchy Schwarz:

$$\begin{aligned}
\|XY\|_{\psi_1} &= \sup_{q\ge 1} q^{-1}\mathbb{E}\{|XY|^q\}^{1/q} \\
&\le \sup_{q\ge 1} q^{-1}\mathbb{E}\{|X|^{2q}\}^{1/2q}\mathbb{E}\{|Y|^{2q}\}^{1/2q} \\
&\le 2\big(\sup_{q\ge 2}(2q)^{-1/2}\mathbb{E}\{|X|^{2q}\}^{1/2q}\big)\cdot\big(\sup_{q\ge 2}(2q)^{-1/2}\mathbb{E}\{|Y|^{2q}\}^{1/2q}\big) \\
&\le 2\|X\|_{\psi_2}\|Y\|_{\psi_2}.
\end{aligned}$$

□

The following lemma from [37] is a Bernstein-type tail inequality for sub-exponential random variables.

Lemma I.5 ([37, Proposition 5.16])**.** *Let $X_1, X_2, \dots, X_n$ be a sequence of independent random variables with* $\max_i\|X_i\|_{\psi_1} \le K$*. Then for any $\varepsilon \ge 0$:*

$$\mathbb{P}\Big\{\Big|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}\{X_i\}\Big| \ge \varepsilon\Big\} \le 2\exp\Big\{-\frac{n\varepsilon}{6eK}\min\Big(\frac{\varepsilon}{eK},1\Big)\Big\} \tag{108}$$

We also use a martingale generalization of [37, Proposition 5.16], whose proof is we omit.

Lemma I.6. *Suppose $(\mathcal{F}_i)_{i\ge 0}$ is a filtration, $X_1, X_2, \dots, X_n$ is a uniformly K-subexponential sequence of random variables adapted to $(\mathcal{F}_i)_{i\ge 0}$ such that almost surely $\mathbb{E}\{X_i|\mathcal{F}_{i-1}\} = 0$. Then for any $\varepsilon \ge 0$:*

$$\mathbb{P}\Big\{\Big|\frac{1}{n}\sum_{i=1}^n X_i\Big| \ge \varepsilon\Big\} \le 2\exp\Big\{-\frac{n\varepsilon}{6eK}\min\Big(\frac{\varepsilon}{eK},1\Big)\Big\} \tag{109}$$

The following is a rough bound on the LASSO error.

Lemma I.7 (Rough bound on LASSO error)**.** *For LASSO estimate $\widehat{\theta}^{\mathsf{L}}$ with regularization λ_n the following bound holds:*

$$\|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1 \le \frac{\|\varepsilon\|^2}{2n\lambda_n} + 2\|\theta_0\|_1 .$$

Proof of Lemma I.7. We first bound the size of $\widehat{\theta}^{\mathsf{L}}$. By optimality of $\widehat{\theta}^{\mathsf{L}}$:

$$\begin{aligned}\lambda_n\|\widehat{\theta}^{\mathsf{L}}\|_1 &\le \frac{1}{2n}\|\varepsilon\|_2^2 + \lambda_n\|\theta_0\|_1 - \frac{1}{2n}\|y - X\widehat{\theta}^{\mathsf{L}}\|_2^2\\ &\le \frac{1}{2n}\|\varepsilon\|_2^2 + \lambda_n\|\theta_0\|_1.\end{aligned}$$

We now use triangle inequality and the bound above to get the claim:

$$\begin{aligned}\|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1 &\le \|\widehat{\theta}^{\mathsf{L}}\|_1 + \|\theta_0\|_1\\ &\le \frac{1}{2n\lambda_n}\|\varepsilon\|^2 + 2\|\theta_0\|_1\,.\end{aligned}$$

□

References

[1] S. Basu, S. Das, G. Michailidis, and A. K. Purnanandam. A system-wide approach to measure connectivity in the financial sector. *Available at SSRN 2816137*, 2017.

[2] S. Basu and G. Michailidis. Regularized estimation in sparse high-dimensional time series models. *Annals of Statistics*, 43(4):1535–1567, 2015.

[3] A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[4] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[5] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[6] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[7] T. T. Cai and Z. Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Annals of statistics*, 45(2):615–646, 2017.

[8] Y. Deshpande, L. Mackey, V. Syrgkanis, and M. Taddy. Accurate inference for adaptive linear models. In *International Conference on Machine Learning*, pages 1202–1211, 2018.

[9] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279. ACM, 2008.

[10] A. Dvoretzky. Asymptotic normality for sums of dependent random variables. In *Proc. 6th Berkeley Symp. Math. Statist. Probab*, volume 2, pages 513–535, 1972.

[11] A. Fujita, J. R. Sato, H. M. Garay-Malpartida, R. Yamaguchi, S. Miyano, M. C. Sogayar, and C. E. Ferreira. Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC systems biology*, 1(1):39, 2007.

[12] S. Gelper, I. Wilms, and C. Croux. Identifying demand effects in a large network of product categories. *Journal of Retailing*, 92(1):25–39, 2016.

[13] D. GUIDANCE. Adaptive designs for clinical trials of drugs and biologics. *Center for Biologics Evaluation and Research (CBER)*, 2018.

[14] P. Hall and C. C. Heyde. *Martingale limit theory and its application.* Academic press, 2014.

[15] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations.* Chapman and Hall/CRC, 2015.

[16] D. Holtz-Eakin, W. Newey, and H. S. Rosen. Estimating vector autoregressions with panel data. *Econometrica: Journal of the Econometric Society*, pages 1371–1395, 1988.

[17] A. Javanmard. *Inference and estimation in high-dimensional data analysis.* PhD thesis, PhD Thesis, Stanford University, 2014.

[18] A. Javanmard and H. Javadi. False discovery rate control via debiased lasso. *Electronic Journal of Statistics*, 13(1):1212–1253, 2019.

[19] A. Javanmard and J. D. Lee. A flexible framework for hypothesis testing in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):685–718, 2020.

[20] A. Javanmard and A. Montanari. Nearly optimal sample size in hypothesis testing for high-dimensional regression. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1427–1434. IEEE, 2013.

[21] A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[22] A. Javanmard and A. Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554, 2014.

[23] A. Javanmard and A. Montanari. Debiasing the lasso: Optimal sample size for gaussian designs. *Annals of Statistics*, 46(6A):2593–2622, 2018.

[24] E. S. Kim, R. S. Herbst, I. I. Wistuba, J. J. Lee, G. R. Blumenschein, A. Tsao, D. J. Stewart, M. E. Hicks, J. Erasmus, S. Gupta, et al. The battle trial: personalizing therapy for lung cancer. *Cancer discovery*, 1(1):44–53, 2011.

[25] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[26] T. L. Lai and C. Z. Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Annals of Statistics*, pages 154–166, 1982.

[27] A. Montanari. Graphical models concepts in compressed sensing. *Compressed Sensing: Theory and Applications*, page 394, 2012.

[28] X. Nie, T. Xiaoying, J. Taylor, and J. Zou. Why adaptively collected data have negative bias and how to correct for it. 2017.

[29] Y. Ning and H. Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195, 2017.

[30] A. K. Seth, A. B. Barrett, and L. Barnett. Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, 35(8):3293–3297, 2015.

[31] J. Shin, A. Ramdas, and A. Rinaldo. On the bias, risk and consistency of sample means in multi-armed bandits. *arXiv preprint arXiv:1902.00746*, 2019.

[32] R. H. Shumway and D. S. Stoffer. *Time series analysis and its applications: with R examples*. Springer Science & Business Media, 2006.

[33] S. Srinivasan, K. Pauwels, D. M. Hanssens, and M. G. Dekimpe. Do promotions benefit manufacturers, retailers, or both? *Management Science*, 50(5):617–629, 2004.

[34] J. H. Stock and M. W. Watson. Vector autoregressions. *Journal of Economic perspectives*, 15(4):101–115, 2001.

[35] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

[36] S. Van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202, 2014.

[37] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge University Press, 2012.

[38] S. Villar, J. Bowden, and J. Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

[39] I. Wilms, S. Basu, J. Bien, and D. S. Matteson. Interpretable vector autoregressions with exogenous time series. *arXiv preprint arXiv:1711.03623*, 2017.

[40] M. Xu, T. Qin, and T.-Y. Liu. Estimation bias in multi-armed bandit algorithms for search advertising. In *Advances in Neural Information Processing Systems*, pages 2400–2408, 2013.

[41] C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[42] L. Zheng, G. Raskutti, et al. Testing for high-dimensional network parameters in autoregressive models. *Electronic Journal of Statistics*, 13(2):4977–5043, 2019.

[43] X. Zhou, S. Liu, E. S. Kim, R. S. Herbst, and J. J. Lee. Bayesian adaptive design for targeted therapy development in lung cancer—a step toward personalized medicine. *Clinical Trials*, 5(3):181–193, 2008.