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

Yash Deshpande* Adel Javanmard† Mohammad Mehrabi‡§

May 7, 2020

Abstract

Adaptive collection of data is commonplace in applications throughout science and engineering. From the point of view of statistical inference however, adaptive data collection induces memory and correlation in the samples, and poses significant challenge.

We consider the high-dimensional linear regression, where the samples are collected adaptively, and the sample size n can be smaller than p, the number of covariates. In this setting, there are two distinct sources of bias: the first due to regularization imposed for consistent estimation, e.g. using the LASSO, and the second due to adaptivity in collecting the samples. We propose *'online debiasing'*, a general procedure for estimators such as the LASSO, which addresses both sources of bias. In two concrete contexts (i) time series analysis and (ii) batched data collection, we demonstrate that online debiasing optimally debiases the LASSO estimate when the underlying parameter θ_0 has sparsity of order $o(\sqrt{n}/\log p)$. In this regime, the debiased estimator can be used to compute p-values and confidence intervals of optimal size.

1 Introduction

Modern data collection, experimentation and modeling are often adaptive in nature. For example, clinical trials are run in phases, wherein the data from a previous phase inform and influence the design of future phases. In commercial recommendation engines, algorithms collect data by eliciting feedback from their users; data which is ultimately used to improve the algorithms underlying the recommendations and so influence the future data. In such applications, adaptive data collection is often carried out for objectives correlated to, but distinct from statistical inference. In clinical trials, an ethical experimenter might prefer to assign more patients a treatment that they might benefit from, instead of the control treatment. In e-commerce, recommendation engines aim to minimize the revenue loss. In other applications, collecting data is potentially costly, and practitioners may choose to collect samples that are a priori deemed most informative. Since such objectives are intimately related to statistical estimation, it is not surprising that adaptively collected data can

*Institute for Data, Systems and Society, Massachusetts Institute of Technology, Email: yash@mit.edu

†Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Email: ajavanma@usc.edu

‡Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Email: mehrabim@usc.edu

§The names of the authors are in alphabetical order.

be used to derive statistically consistent estimates, often using standard estimators. The question of statistical inference however, is more subtle: on the one hand, consistent estimation indicates that the collected samples are informative enough. On the other hand, adaptive collection induces endogenous correlation in the samples, resulting in bias in the estimates. In this paper, we address the following natural question raised by this dichotomy:

Can adaptively collected data be used for ex post statistical inference?

We will focus on the linear model, where the samples $(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)$ satisfy:

$$y_i = \langle x_i, \theta_0 \rangle + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2). \tag{1}$$

Here $\theta_0 \in \mathbb{R}^p$ is an unknown parameter vector relating the covariates x_i to the response y_i, and the noise ε_i are i.i.d. $\mathsf{N}(0, \sigma^2)$ random variables. In vector form, we write Eq.(1) as

$$y = X\theta_0 + \varepsilon, \tag{2}$$

where $y = (y_1, y_2, \dots, y_n)$, $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$ and the design matrix $X \in \mathbb{R}^{n \times p}$ has rows $x_1^{\mathsf{T}}, \dots, x_n^{\mathsf{T}}$. When the samples are adaptively collected, the data point (y_i, x_i) is obtained *after viewing the previous data points* $(y_1, x_1), \dots, (y_{i-1}, x_{i-1})$[1].

In the 'sample-rich' regime when $p < n$, the standard approach would be to compute the least squares estimate $\widehat{\theta}^{\mathsf{LS}} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}y$, and assess the uncertainty in $\widehat{\theta}^{\mathsf{LS}}$ using a central limit approximation $(X^{\mathsf{T}}X)^{1/2}(\widehat{\theta}^{\mathsf{LS}} - \theta_0) \approx \mathsf{N}(0, \mathrm{I}_p)$ [LW82]. However, while the estimator $\widehat{\theta}^{\mathsf{LS}}$ is consistent under fairly weak conditions, adaptive data collection complicates the task of characterizing its distribution. One hint for this is the observation that, in stark contrast with the non-adaptive setting, $\widehat{\theta}^{\mathsf{LS}} = \theta_0 + (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\varepsilon$ is in general *a biased estimate* of θ_0. Adaptive data collection creates correlation between the responses y_i (therefore ε_i) and covariate vectors $x_{i+1}, x_{i+2}, \dots, x_n$ observed in the future. In the context of multi-armed bandits, where the estimator $\widehat{\theta}^{\mathsf{LS}}$ for model (1) reduces to sample averages, [XQL13, VBW15] observed such bias empirically, and [NXTZ17, SRR19] characterized and developed upper bounds on the bias. While bias is an important problem, estimates may also show higher-order distributional defects that complicate inferential tasks.

This phenomenon is exacerbated in the high-dimensional or 'feature-rich' regime when $p > n$. Here the design matrix X becomes rank-deficient, and consistent parameter estimation requires (i) additional structural assumptions on θ_0 and (ii) regularized estimators beyond $\widehat{\theta}^{\mathsf{LS}}$, such as the LASSO [Tib96]. Such estimators are non-linear, non-explicit and, consequently it is difficult to characterize their distribution even with strong random design assumptions [BM12, JM14b]. In analogy to the low-dimensional regime, it is relatively easier to develop consistency guarantees for estimation using the LASSO when $p > n$. Given the sample $(y_1, x_1), \dots (y_n, x_n)$ one can compute the LASSO estimate $\widehat{\theta}^{\mathsf{L}} = \widehat{\theta}^{\mathsf{L}}(y, X; \lambda_n)$

$$\widehat{\theta}^{\mathsf{L}} = \arg\min_{\theta} \Big\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1 \Big\}, \tag{3}$$

If θ_0 is sparse with at most $s_0 \ll p$ non-zero entries and the design X satisfies some technical conditions, the LASSO estimate, for an appropriate choice of λ_n has estimation error $\|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_2^2$ of

[1] Formally, we assume a filtration $(\mathfrak{F}_i)_{i \le n}$ to which the sequence $(y_i, x_i)_{i \le n}$ is adapted, and with respect to which the sequence $(x_i)_{i \le n}$ is predictable

order $\sigma^2 s_0(\log p)/n$, with high probability [BM15, BB15]. In particular the estimate is consistent provided the sparsity satisfies $s_0 = o(n/\log p)$. This estimator is biased though because of two distinct reasons. The first is the regularization imposed in Eq.(3), which disposes $\widehat{\theta}^{\mathsf{L}}$ to have small ℓ_1 norm. The second is the correlation induced between X and ε due to adaptive data collection. To address the first source, [ZZ11, JM14a, VdGBR+14] proposed a so-called "*debiased estimate*" of the form

$$\widehat{\theta}^{\mathsf{off}} = \widehat{\theta}^{\mathsf{L}} + \frac{1}{n} M X^{\mathsf{T}}(y - X\widehat{\theta}^{\mathsf{L}}), \tag{4}$$

where M is chosen as an 'approximate inverse' of the sample covariance $\widehat{\Sigma} = X^{\mathsf{T}}X/n$. The intuition for this idea is the following decomposition that follows directly from Eqs.(1), (4):[2]

$$\widehat{\theta}^{\mathsf{off}} - \theta_0 = (I_p - M\widehat{\Sigma})(\widehat{\theta}^{\mathsf{L}} - \theta_0) + \frac{1}{n} M X^{\mathsf{T}}\varepsilon. \tag{5}$$

When the data collection is non-adaptive, X and ε are independent and therefore, conditional on the design X, $MX^{\mathsf{T}}\varepsilon/n$ is distributed as $\mathsf{N}(0, \sigma^2 Q/n)$ where $Q = M\widehat{\Sigma}M^{\mathsf{T}}$. Further, the bias in $\widehat{\theta}^{\mathsf{off}}$ is isolated to the first term, which intuitively should be of smaller order than the second term, provided both $\widehat{\theta}^{\mathsf{L}} - \theta_0$ and $M\widehat{\Sigma} - \mathrm{I}_p$ are small in an appropriate sense. This intuition suggests that, if the second term dominates the first term in $\widehat{\theta}^{\mathsf{off}}$, we can produce confidence intervals for θ_0 in the usual fashion using the debiased estimate $\widehat{\theta}^{\mathsf{off}}$ [JM14a, JM14b, VdGBR+14]. For instance, with $Q = M\widehat{\Sigma}M^{\mathsf{T}}$, the interval $\big[\widehat{\theta}^{\mathsf{off}}_1 - 1.96\sigma\sqrt{Q_{11}/n}, \widehat{\theta}^{\mathsf{off}}_1 + 1.96\sigma\sqrt{Q_{11}/n}\big]$ forms a standard 95% confidence interval for the parameter $\theta_{0,1}$. In the so-called 'random design' setting –when the rows of X are drawn i.i.d. from a broad class of distributions– this approach to inference via the debiased estimate $\widehat{\theta}^{\mathsf{off}}$ enjoys several optimality guarantees: the resulting confidence intervals have minimax optimal size [Jav14, JM14a, CG17], and are semi-parametrically efficient [VdGBR+14].

This line of argument breaks down when the samples are adaptively collected, as the debiased estimate $\widehat{\theta}^{\mathsf{off}}$ *still suffers the second source of bias.* Indeed, this is exactly analogous to $\widehat{\theta}^{\mathsf{LS}}$ in low dimensions. Since M, X and the noise ε are correlated, we can no longer assert that the term $MX^{\mathsf{T}}\varepsilon/n$ is unbiased. Indeed, characterizing its distribution can be quite difficult, given the intricate correlation between M, X and ε induced by the data collecting policy and the procedure for choosing M. We illustrate the failure of offline debiasing in two scenarios of interest in this paper: (*i*) batched data collection and (*ii*) autoregressive time series.

1.1 Why offline debiasing fails?

Batched data collection

Consider a stylized model of adaptive data collection wherein the experimenter (or analyst) collects data in two phases or batches. In the first phase, the experimenter collects an initial set of samples $(y_1, x_1), \dots, (y_{n_1}, x_{n_1})$ of size $n_1 < n$ where the responses follow Eq.(1) and the covariates are i.i.d. from a distribution $\mathbb{P}_x$. Following this, she computes an intermediate estimate $\widehat{\theta}^1$ of θ_0 and then collects additional samples $(y_{n_1+1}, x_{n_1+1}), \dots, (y_n, x_n)$ of size $n_2 = n - n_1$, where the covariates x_i are drawn independently from the law of x_1, conditional on the event $\{\langle x_1, \widehat{\theta}^1\rangle \geq \varsigma\}$, where ς is a

[2]The notation $\widehat{\theta}^{\mathsf{off}}$ stands for "offline" debiasing. We use this notation/terminology to highlight its main difference from the "online" debiasing that will be introduced later in this paper.

threshold, that may be data-dependent. This is a typical scenario where the response y_i represents an instantaneous reward that the experimenter wishes to maximize, as in multi-armed bandits [LR85, BCB+12]. For instance, clinical trials may be designed to be response-adaptive and allocate patients to treatments that they are likely to benefit from based on prior data [ZLK+08, KHW+11]. The multi-armed bandit problem is a standard formalization of this trade-off, and a variety of bandit algorithms are designed to operate in distinct phases of 'explore–then exploit'[RT10, DM12, BB15, PRC+16]. The model we describe above is a close approximation of data collected from one arm in a run of such an algorithm. With the full samples $(y_1, x_1), \dots, (y_n, x_n)$ at hand, the experimenter would like to perform inference on a fixed coordinate $\theta_{0,a}$ of the underlying parameter.

As a numerical example, we consider $\theta_0 \in \{0,1\}^{600}$ with exactly $s_0 = 10$ non-zero entries. We obtain the first batch $(y_1, x_1), \dots, (y_{500}, x_{500})$ of observations with $y_i = \langle x_i, \theta_0 \rangle + \varepsilon_i$, $x_i \overset{\text{iid}}{\sim} \mathsf{N}(0, \Sigma)$ and $\varepsilon_i \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$ where we use the covariance Σ as below:

$$\Sigma_{a,b} = \begin{cases} 1 & \text{if } a = b, \\ 0.1 & \text{if } |a-b| = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Based on this data, we construct an intermediate estimator $\widehat{\theta}^1$ on $(y^{(1)}, X_1)$ using two different strategies: (i) debiased LASSO and (ii) ridge regression with cross-validation. With this estimate we now sample new covariates $x_{501}, \dots, x_{1000}$ independently from the law of $x|_{\langle x, \widehat{\theta}^1 \rangle \geq \langle \widehat{\theta}^1, \Sigma \widehat{\theta}^1 \rangle^{1/2}}$ and the corresponding outcomes $y_{501}, \dots, y_{1000}$ are generated according to Eq.(1). Unconditionally, $\langle x, \widehat{\theta}^1 \rangle \sim \mathsf{N}(0, \langle \widehat{\theta}^1, \Sigma \widehat{\theta}^1 \rangle)$, so this choice of threshold corresponds to sampling covariates that correlate with $\widehat{\theta}^1$ at least one standard deviation higher than expected unconditionally. This procedure yields two batches of data, each of $n_1 = n_2 = 500$ data points, combining to a set of 1000 samples.

From the full dataset $(y_1, x_1), \dots, (y_{1000}, x_{1000})$ we compute the LASSO estimate $\widehat{\theta}^{\mathsf{L}} = \widehat{\theta}^{\mathsf{L}}(y, X; \lambda)$ with $\lambda = 2.5 \lambda_{\max}(\Sigma) \sqrt{(\log p)/n}$. Offline debiasing yields the following prescription to debias $\widehat{\theta}^{\mathsf{L}}$:

$$\widehat{\theta}^{\mathsf{off}} = \widehat{\theta}^{\mathsf{L}} + \frac{1}{n} \Omega(\widehat{\theta}^1) X^{\mathsf{T}} (y - X \widehat{\theta}^{\mathsf{L}}),$$

where $\Omega(\widehat{\theta})$ is the population precision matrix:

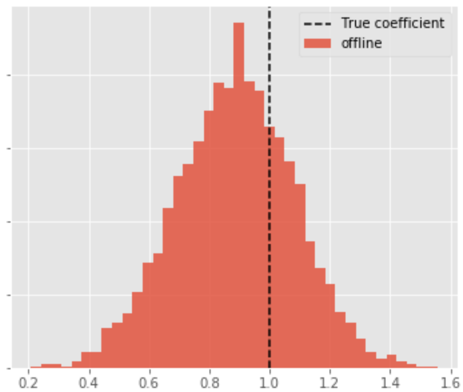
$$\Omega(\widehat{\theta}^1)^{-1} = \frac{1}{2} \mathbb{E}\{x x^{\mathsf{T}}\} + \frac{1}{2} \mathbb{E}\left\{ x x^{\mathsf{T}} \middle| \langle x, \widehat{\theta}^1 \rangle \geq \|\Sigma^{1/2} \widehat{\theta}^1\| \right\}.$$

We generate the dataset for 100 Monte Carlo iterations and compute the offline debiased estimate $\widehat{\theta}^{\mathsf{off}}$ for each iteration. Figure 1 shows the histogram of the entries $\widehat{\theta}^{\mathsf{off}}$ on the support of θ_0 for the two choices of $\widehat{\theta}^1$. As we see $\widehat{\theta}^{\mathsf{off}}$ still has considerable bias, due to adaptivity in the data collection.

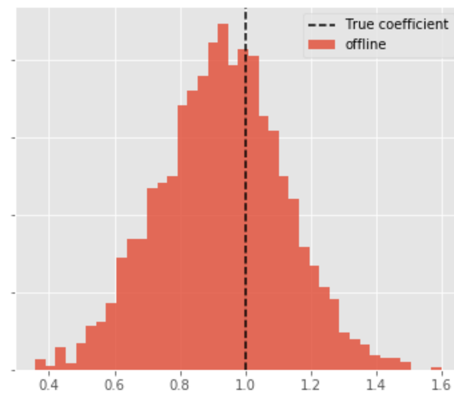
Autoregressive time series

A vector autoregressive (VAR) time series model posits that data points z_t evolve according to the dynamics:

$$z_t = \sum_{\ell=1}^{d} A^{(\ell)} z_{t-\ell} + \zeta_t \tag{6}$$



(a) with $\widehat{\theta}^1$ the debiased LASSO on first batch



(b) with $\widehat{\theta}^1$ the ridge estimate on first batch

Figure 1: Histograms of the offline debiased estimate $\widehat{\theta}^{\sf off}$ restricted to the support of θ_0. The dashed line indicates the true coefficient size. Recall that the second batch is chosen based on an intermediate estimator $\widehat{\theta}^1$ computed on the first batch. (Left) $\widehat{\theta}^1$ is debiased LASSO on the first batch, (Right) $\widehat{\theta}^1$ is ridge estimate on the first batch. As we observe the offline debiasing (even with access to the precision matrix Ω of the random designs) has a significant bias and dose not admit a Gaussian distribution.

where $A^{(\ell)} \in \mathbb{R}^{p\times p}$ are time invariant coefficients and ζ_t is the noise term satisfying $\mathbb{E}(\zeta_t) = 0$ (zero-mean), $\mathbb{E}(\zeta_t\zeta_t^{\sf T}) = \Sigma_\zeta$ (stationary covariance), and $\mathbb{E}(\zeta_t\zeta_{t-k}^{\sf T}) = 0$ for $k > 0$ (no serial correlation). Given the data $z_1, \dots, z_T$, the task of interest is to perform statistical inference on the model parameters, i.e., coefficient matrices $A^{(1)}, \dots, A^{(d)}$. Clearly, the samples z_t are 'adaptively collected', in the sense that there is serial correlation in the samples. Indeed, the data point z_t depends on the previous data points $z_{t-1}, z_{t-2}, \dots, z_1$.

As in the batched data example, we will carry out a simple illustration. We generate data from a $\mathsf{VAR}(d)$ model with $p = 15$, $d = 5$, $T = 60$, and diagonal $A^{(\ell)}$ matrices with value $b = 0.15$ on their diagonals. We also generate $\zeta_t \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \Sigma_\zeta)$. Note that this is a high-dimensional setting as the number of parameters dp^2 exceeds the sample size $(T-d)p$. We keep the covariance of the noise terms ζ_t as below:

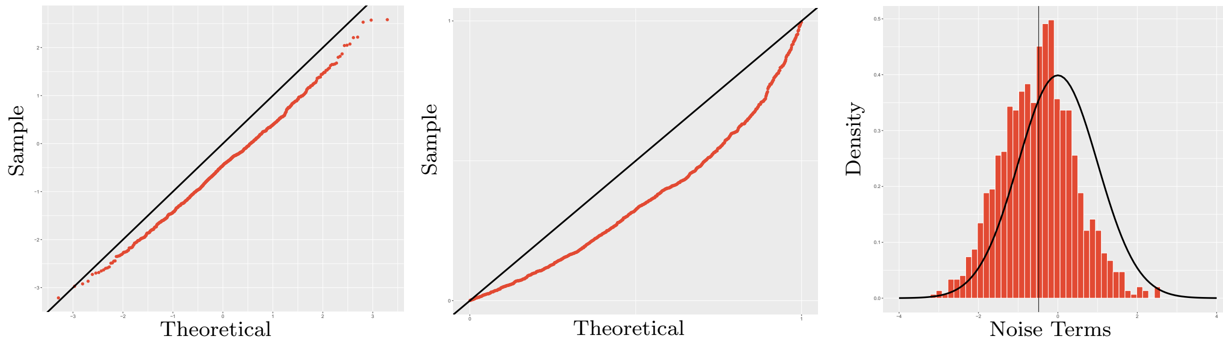
$$\Sigma_{\zeta,ij} = 0.5^{\mathbb{I}(i\neq j)}$$

To estimate the parameters, we define the covariate vectors $x_t = (z_{t+d-1}^{\sf T}, \dots, z_t^{\sf T})^{\sf T} \in \mathbb{R}^{dp}$, obtained by concatenating d consecutive data points and $\varepsilon = (\zeta_{d+1,i}, \zeta_{d+2,i} \dots, \zeta_{T,i})$. We focus on the noise component of the offline debiased estimate, i.e.,

$$W^{\rm off} = \frac{1}{\sqrt{n}} M \sum_{t=1}^{n} x_t \varepsilon_t \,, \tag{7}$$

with M denoting the decorrelating matrix in the debiased estimate as per (4).

In Figure 2, we show the QQ-plot, PP-plot and histogram of $W_1^{\rm off}$ (corresponding to the entry $(1,1)$ of matrix A_1) for 1000 different realizations of the noise ζ_t. As we observe, even the noise



(a) (b) (c)

Figure 2: Empirical behavior of noise term associated with the offline debiased estimate of a fixed coordinate of Gaussian $\mathsf{VAR}(d)$ model. In this example, $d = 5, p = 15, T = 60, \rho = 0.5$, $\Sigma_\zeta(i,j) = \rho^{|i-j|}$, and $A^{(\ell)}$ matrices are diagonal with value $b = 0.15$ on their diagonals. Plots 2a, 2b, and 2c show the QQ plot, PP plot, and the histogram of the offline debiased noise terms (red) over 1000 independent experiments, respectively and black curve/lines denote the ideal standard normal distribution. As we observe, even the noise component of the offline debiased estimator deviates from the standard normal distribution; This implies the failure of offline debiasing method for statistical inference purposes when the samples are correlated. The vertical black line in (c) indicates the mean of the noise component of the offline debiased estimator.

component $W^{\rm off}$ is biased because the offline construction of M depends on all features x_t and hence endogenous noise ζ_t. Recall that for the setting with an i.i.d sample, the noise component is zero mean gaussian for any finite sample size n. This further highlights the challenge of high-dimensional statistical inference with adaptively collected samples and demonstrate why the classical debiasing approach will not work in this case.

2 Online debiasing

We propose *online debiased* estimator $\widehat{\theta}^{\sf on} = \widehat{\theta}^{\sf on}(y, X; (M_i)_{i\le n}, \lambda)$ that takes the form

$$\widehat{\theta}^{\sf on} \equiv \widehat{\theta}^{\sf L} + \frac{1}{n}\sum_{i=1}^{n} M_i x_i (y_i - x_i^{\sf T}\widehat{\theta}^{\sf L}). \tag{8}$$

The term 'online' comes from the first crucial constraint of *predictability* imposed on the sequence $(M_i)_{i\le n}$.

Definition 2.1 (Predictability)**.** *Without loss of generality, there exists a filtration $(\mathfrak{F}_i)_{i\ge 0}$ so that, for $i = 1, 2, \dots, n$, (i) ε_i are adapted to $\mathfrak{F}_i$ and ε_i is independent of $\mathfrak{F}_j$ for $j < i$. We assume that the sequences $(x_i)_{i\ge1}$ and $(M_i)_{i\ge1}$ are* predictable *with respect to $\mathfrak{F}_i$, i.e. for each i, x_i and M_i are measurable with respect to $\mathfrak{F}_{i-1}$.*

With predictability, the data points (y_i, x_i) are adapted to the filtration $(\mathfrak{F}_i)_{i\le n}$ and, moreover, the covariates x_i are predictable with respect to $\mathfrak{F}_i$. Intuitively, the σ-algebra $\mathfrak{F}_i$ contains all information in the data, as well as potential external randomness, that is used to query the new data covariate x_{i+1}. Predictability ensures that only this information may be used to construct the matrix M_{i+1}. Analogous to Eq.(5) we can decompose $\widehat{\theta}^{\mathsf{on}}$ into two components:

$$\widehat{\theta}^{\mathsf{on}} = \theta_0 + \frac{1}{\sqrt{n}}\big(B_n(\widehat{\theta}^{\mathsf{L}} - \theta_0) + W_n\big) \tag{9}$$

$$\text{where } B_n \equiv \sqrt{n}\Big(I_p - \frac{1}{n}\sum_i M_i x_i x_i^{\mathsf{T}}\Big),$$

$$\text{and } W_n \equiv \frac{1}{\sqrt{n}}\sum_i M_i x_i \varepsilon_i.$$

Predictability of $(M_i)_{i\le n}$ ensures that W_n is unbiased and the bias in $\widehat{\theta}^{\mathsf{on}}$ is contained entirely in the first term $B_n(\widehat{\theta}^{\mathsf{L}} - \theta_0)$. Suppose that, analogous to offline debiasing, we prove that the bias term $B_n(\widehat{\theta}^{\mathsf{L}} - \theta_0)$ is of smaller order than the variance term W_n. We are then left with the problem of characterizing the asymptotic distribution of the sequence W_n. As the sequence $\sqrt{n}W_n = \sum_i M_i x_i \varepsilon_i$ is *a martingale* with respect to the filtration $\mathfrak{F}_i$, one might expect that W_n is asymptotically Gaussian. The following 'stability' property, identified first by Lai and Wei [LW82] in this context, is crucial to ensure that this intuition is correct.

Definition 2.2 (Stability)**.** *Consider a square integrable triangular martingale array $\{Z_{i,n}\}_{i\le n, n\ge 1}$ adapted to a filtration $\mathfrak{F}_i$ and its quadratic variation $V_n = \sum_{i\le n}\mathbb{E}\{(Z_{i,n} - Z_{i-1,n})^2|\mathfrak{F}_{i-1}\}$. Note that V_n is non-negative random variable, measurable with respect to $\mathfrak{F}_{n-1}$. We say that the martingale array $\{Z_{i,n}\}_{i\ge 1}$ is stable if there exists a constant $v_\infty > 0$ where $\lim_{n\to\infty} V_n = v_\infty$ in probability.*

An important contribution of our paper is to develop online debiasing estimators $\widehat{\theta}^{\mathsf{on}}$ whose underlying martingales are stable. The specifics of construction of predictable sequence $(M_i)_{i\le n}$ and deriving the distributional characterization of the debiased estimator $\widehat{\theta}^{\mathsf{on}}$ depend on the context of the problem at hand. In this paper, we instantiate this idea in two concrete contexts: (i) time series analysis (Section 3) and (ii) batched data collection (Section 4). For both of these settings,

1. We first establish estimation results for the LASSO estimate, showing that even with adaptive data collection, the LASSO estimate enjoys good estimation error (Theorems 3.2 and 4.1). These results draw significantly on prior work in high-dimensional estimation [BM15, BVDG11].

2. Next, we propose constructions for the online debiasing sequence $(M_i)_{i\le n}$, using an optimization program that trades off variance with bias, *while ensuring stability*. This optimization program is a novel modification of the approximate inverse construction in [JM14a]. The important change is the inclusion of an ℓ_1 constraint in the program, which ensures stability of the underlying martingales, and allows the use of a martingale CLT theorem to characterize the distribution of the online debiased estimator.

3. We establish a distributional characterization of the resulting online debiased estimate $\widehat{\theta}^{\mathsf{on}}$ (Theorems 3.8 and 4.9). Informally, this demonstrates that coordinates of $\widehat{\theta}^{\mathsf{on}}$ are approximately Gaussian with a covariance computable from data.

In Section 5, we demonstrate how the online debiased estimate $\widehat{\theta}^{\mathsf{on}}$ can be used to compute standard inferential primitives like confidence intervals and p-values. Section 6 contains numerical experiments that demonstrate the validity our proposals on both synthetic and real data. In Section

7 we develop computationally efficient iterative descent methods to construct the online debiasing sequence $(M_i)_{i\le n}$. In the interest of reproducibility, we make an R implementation of our algorithm publicly available at `http://faculty.marshall.usc.edu/Adel-Javanmard/OnlineDebiasing`.

Our proposal of online debiasing approach builds on the insight in [DMST18], which has studied a similar problem for low-dimensional settings ($p < n$). We provide a detailed discussion of this this work in Section 4.1.1, highlighting the main distinctions and the inefficacy of that method for high-dimensional setting to further motivate our work and contributions.

Notation Henceforth, we use the shorthand $[p] \equiv \{1, \dots, p\}$ for an integer $p \ge 1$, and $a \wedge b \equiv \min(a, b)$, $a \vee b \equiv \max(a, b)$. We also indicate the matrices in upper case letters and use lower case letters for vectors and scalars. We write $\|v\|_p$ for the standard ℓ_p norm of a vector v, $\|v\|_p = (\sum_i |v_i|^p)^{1/p}$ and $\|v\|_0$ for the number of nonzero elements of v. We also denote by $\text{supp}(v)$, the support of v that is the positions of its nonzero entries. For a matrix A, $\|A\|_p$ represents its ℓ_p operator norm and $\|A\|_\infty = \max_{i,j} |A_{ij}|$ denotes the maximum absolute value of its entries. In particular, $\|A\|_1$ is the $\ell_1 - \ell_1$ norm of matrix A (the maximum ℓ_1 norm of its columns). For two matrices A, B, we use the shorthand $\langle A, B\rangle \equiv \text{trace}(A^{\mathsf{T}} B)$. In addition $\phi(x)$ and $\Phi(x)$ respectively represents the probability density function and the cumulative distribution function of standard normal variable. Also, we use the term *with high probability* to imply that the probability converges to one as $n \to \infty$.

3 Online debiasing for high-dimensional time series

The Gaussian *vector autoregressive model* of order d (or $\mathsf{VAR}(d)$ for short) [SS06], posits that data points z_t follow the dynamics:

$$z_t = \sum_{\ell=1}^{d} A^{(\ell)} z_{t-\ell} + \zeta_t, \tag{10}$$

where $A^{(\ell)} \in \mathbb{R}^{p\times p}$ and $\zeta_t \overset{\text{iid}}{\sim} \mathsf{N}(0, \Sigma_\zeta)$. VAR models are extensively used across science and engineering (see [FSGM+07, SW01, HENR88, SBB15] for notable examples in macroeconomics, genomics and neuroscience). Given the data $z_1, \dots, z_T$, the fundamental task is to estimate the parameters of the VAR model, viz. the matrices $A^{(1)}, \dots A^{(d)}$. The estimates of the parameters can be used in a variety of ways depending on the context: to detect or test for stationarity, forecast future data, or suggest causal links. Since each matrix is $p \times p$, this forms a putative total of dp^2 parameters, which we estimate from a total of $(T - d)p$ linear equations (Eq.(10) with $t = d + 1, \dots, T$). For the i^{th} coordinate of z_t, Eq.(10) reads

$$z_{t,i} = \sum_{\ell=1}^{d} \langle z_{t-\ell}, A_i^{(\ell)} \rangle + \zeta_{t,i}, \tag{11}$$

where $A_i^{(\ell)}$ denotes the i^{th} row of the matrix $A^{(\ell)}$. This can be interpreted in the linear regression form Eq.(1) in dimension dp with $\theta_0 \in \mathbb{R}^{dp}, X \in \mathbb{R}^{(T-d)\times dp}$, $y, \varepsilon \in \mathbb{R}^{T-d}$ identified as:

$$\begin{aligned}
\theta_0 &= (A_i^{(1)}, A_i^{(2)}, \dots, A_i^{(d)})^\mathsf{T}, \\
X &= \begin{bmatrix} z_d^\mathsf{T} & z_{d-1}^\mathsf{T} & \cdots & z_1^\mathsf{T} \\ z_{d+1}^\mathsf{T} & z_d^\mathsf{T} & \cdots & z_2^\mathsf{T} \\ \vdots & \vdots & \ddots & \vdots \\ z_{T-1}^\mathsf{T} & z_{T-2}^\mathsf{T} & \cdots & z_{T-d}^\mathsf{T} \end{bmatrix}, \\
y &= (z_{d+1,i}, z_{d+2,i}, \dots, z_{T,i}), \\
\varepsilon &= (\zeta_{d+1,i}, \zeta_{d+2,i}, \dots, \zeta_{T,i}). \qquad (12)
\end{aligned}$$

We omit the dependence on the coordinate i, and also denote the rows of X by $x_1, \dots, x_n \in \mathbb{R}^{dp}$, with $n = T - d$. Given sufficient data, or when T is large in comparison with dp, it is possible to estimate the parameters using least squares [SS06, LW82]. In [BM15], Basu and Michailidis consider the problem of estimating the parameters when number of time points T is small in comparison with the total number of parameters dp, with the proviso that the matrices $A^{(\ell)}$ are sparse. Their estimation results build on similar ideas as [BVDG11, Theorem 6.1], relying on proving a restricted eigenvalue property for the design $X^\mathsf{T} X/n$. This result hinges on stationary properties of the model (10), which we summarize prior to stating the estimation result.

Definition 3.1 (Stability and invertibility of $\mathsf{VAR}(d)$ Process [BM15])**.** *A* $\mathsf{VAR}(d)$ *process with an associated reverse characteristic polynomial*

$$\mathcal{A}(\gamma) = I - \sum_{\ell=1}^{d} A^{(\ell)} \gamma^\ell \,, \qquad (13)$$

is called stable and invertible if $\det(\mathcal{A}(\gamma)) \neq 0$ *for all* $\gamma \in \mathbb{C}$ *with* $|\gamma| = 1$*. Based on this characteristic polynomial, we also define the following spectral parameters:*

$$\begin{aligned}
\mu_{\min}(\mathcal{A}) &= \min_{|\gamma|=1} \lambda_{\min}(\mathcal{A}^*(\gamma)\mathcal{A}(\gamma)) \\
\mu_{\max}(\mathcal{A}) &= \max_{|\gamma|=1} \lambda_{\max}(\mathcal{A}^*(\gamma)\mathcal{A}(\gamma))
\end{aligned}$$

Theorem 3.2 (Estimation Bound)**.** *Recall the relation* $y = X\theta_0 + \varepsilon$*, where* X, y, θ_0 *are given by* (12) *and let* $\widehat{\theta}^\mathsf{L}$ *be the Lasso estimator*

$$\widehat{\theta}^\mathsf{L} = \arg\min_{\theta \in \mathbb{R}^{dp}} \Big\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1 \Big\} \,. \qquad (14)$$

Assume that $|\text{supp}(\theta_0)| \le s_0$*, and define*

$$\begin{aligned}
\omega &= \frac{d\lambda_{\max}(\Sigma_\zeta)}{\lambda_{\min}(\Sigma_\zeta)} \cdot \frac{\mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})} \\
\alpha &= \frac{\lambda_{\min}(\Sigma_\zeta)}{\mu_{\max}(\mathcal{A})}.
\end{aligned}$$

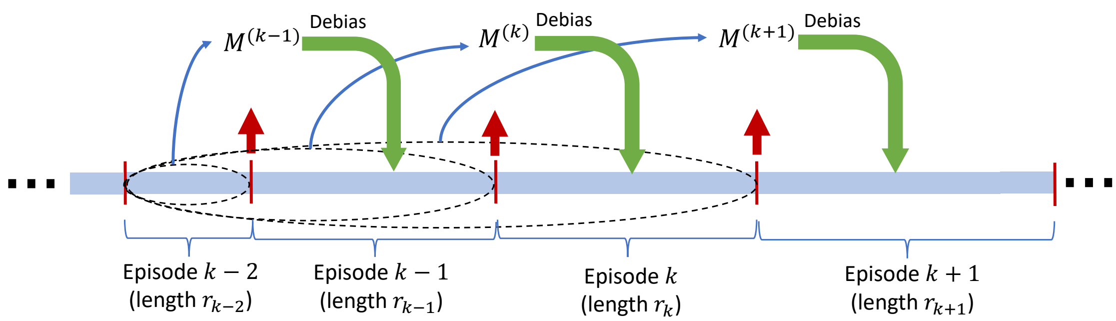


Figure 3: Schematic for constructing the debiasing matices $M^{(\ell)}$. We divide time into K episodes $E_0, \dots, E_{K-1}$; in episode ℓ, M_i is held constant at $M^{(\ell)}$, which is a function of x_t in *all prior* episodes.

There exists a universal constant $C > 0$, such that for any $n \geq C\alpha\omega^2 s_0 \log(dp)$ and $\lambda_n = \lambda_0\sqrt{\log(dp)/n}$, with $\lambda_0 \geq 4\lambda_{\max}(\Sigma_\zeta)(1 \vee \mu_{\max}(\mathcal{A}))/\mu_{\min}(\mathcal{A})$ the following happens. With probability at least $1-(dp)^{-6}$, the estimate satisfies:

$$\|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1 \leq C\frac{\lambda_0}{\alpha}\sqrt{\frac{s_0^2\log(dp)}{n}}.$$

In short, given the standardized setting where λ_0, α are order one, the ℓ_1 estimation error rate is of order $s_0\sqrt{\log(dp)/n}$, which is the same obtained in data without temporal dependence. Our proof is similar to that of Basu and Michailidis [BM15], and relies on establishing a now-standard restricted eigenvalue property for the design $X^{\mathsf{T}}X/n$. The spectral characteristics of the time series quantified in Definition 3.1 play an important part in establishing this. We refer the reader to Appendix A for the proof, as well as a discussion of the differences with the proof of [BM15].

3.1 Constructing the online debiased estimator

Our task now is to construct a predictable sequence of debiasing matrices $\{M_i\}_{i\leq n}$. One simple approach is the 'sample-splitting' approach: construct a generalized inverse M based on the first $n/2$ data points using, for example, the program of [JM14a] and let the sequence $\{M_i\}_{i\leq n}$ be defined by

$$M_i = \begin{cases} 0 & \text{if } i \leq n/2 \\ M & \text{if } n/2 < i \leq n. \end{cases}$$

It is easy to see that this is a valid predictable sequence. However, due to sample-splitting, it does not make an efficient use of the data and loses power. More importantly, it is not clear that the underlying martingale (the noise component of the debiased estimator $\sqrt{n}W_n = \sum_i M_i x_i \varepsilon_i$) will be stable in the sense of Definition 2.2. Our proposal generalizes sample-splitting via an episodic structure and, importantly, regularizes to ensure stability.

We partition the time indices $[n]$ into K episodes $E_0, \dots, E_{K-1}$, with E_ℓ of length r_ℓ, so that $\sum_{\ell=0}^{K-1} r_\ell = n$. Over an episode ℓ, we keep the debiasing matrix $M_i = M^{(\ell)}$ to be fixed over time points in the episode. Moreover, $M^{(\ell)}$ is constructed using all the time points in *previous* episodes $E_0, \dots, E_{\ell-1}$ in the following way. Let $n_\ell = r_0 + \ldots + r_{\ell-1}$, for $\ell = 1, \dots, K$; hence, $n_K = n$. Define the sample covariance of the features in the first ℓ episodes.

$$\widehat{\Sigma}^{(\ell)} = \frac{1}{n_\ell} \sum_{t \in E_0 \cup \ldots \cup E_{\ell-1}} x_t x_t^\mathsf{T} \,,$$

The matrix $M^{(\ell)}$ has rows $(m_a^\ell)_{a\in[dp]}$ as the solution of the optimization:

$$\begin{aligned} \text{minimize} \quad & m^\mathsf{T} \widehat{\Sigma}^{(\ell)} m \\ \text{subject to} \quad & \|\widehat{\Sigma}^{(\ell)} m - e_a\|_\infty \le \mu_\ell, \quad \|m\|_1 \le L \,, \end{aligned} \tag{15}$$

for appropriate values of $\mu_\ell, L > 0$. We then construct the online debiased estimator for coordinate a of θ_0 as follows:

$$\widehat{\theta}^{\mathsf{on}} = \widehat{\theta}^{\mathsf{L}} + \frac{1}{n} \sum_{\ell=1}^{K-1} \sum_{t\in E_\ell} M^{(\ell)} x_t (y_t - \langle x_t, \widehat{\theta}^{\mathsf{L}} \rangle) \,. \tag{16}$$

In Section 3.2, we show that the constructed online debiased estimator $\widehat{\theta}^{\mathsf{on}}$ is asymptotically unbiased and admits a normal distribution. To do that we provide a high probability bound on the bias of $\widehat{\theta}^{\mathsf{on}}$ (See Lemma A.5). This bound is in terms of the batch sizes r_ℓ, from which we propose the following guideline for choosing them: $r_0 \sim \sqrt{n}$ and $r_\ell \sim \beta^\ell$, for a constant $\beta > 1$, and $\ell \ge 1$.

Before proceeding into the distributional characterization of the online debiased estimator for θ_0 (entries of coefficient matrices $A^{(\ell)}$), we revisit the numerical example from Section 1.1 in which the (offline) debiased estimator of [JM14a] does not display an unbiased normal distribution. However, as we will observe the constructed online debiased estimator empirically admits an unbiased normal distribution.

Revisiting the numerical example from Section 1.1 In Section 1.1, we considered a $\mathsf{VAR}(d)$ model with $p = 15$, $d = 5$, $T = 60$, and diagonal $A^{(\ell)}$ matrices with value $b = 0.15$ on their diagonals. The covariance matrix Σ_ζ of the noise terms ζ_t is chosen as $\Sigma_\zeta(i,j) = \rho^{\mathbb{I}(i\neq j)}$ with $\rho = 0.5$ and $i,j \in [p]$. The population covariance matrix of vector $x_t = (z_{t+d-1}^\mathsf{T}, \ldots, z_t^\mathsf{T})^\mathsf{T}$ is a dp by dp matrix Σ consisting of d^2 blocks of size $p \times p$ with $\Gamma_z(r-s)$ as block (r,s). The analytical formula to compute $\Gamma_z(\ell)$ is given by [BM15]:

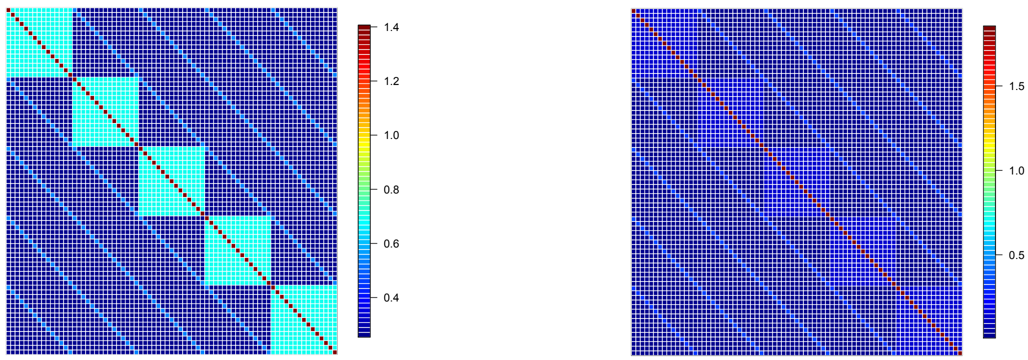
$$\Gamma_z(\ell) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathcal{A}^{-1}(e^{-j\theta}) \Sigma_\zeta (\mathcal{A}^{-1}(e^{-j\theta}))^* e^{j\ell\theta} d\theta \,,$$

where $\mathcal{A}(\gamma)$ is given in equation (13). Figure 4 shows the heat maps of magnitudes of the elements of Σ and the precision matrix $\Omega = \Sigma^{-1}$ for the on hand $\mathsf{VAR}(5)$ process. As evident from Figure 2, the noise component of offline debiased estimator is biased. Here, we look into the noise component of the online debiased estimator given by

$$W^{\text{on}} = \frac{1}{\sqrt{n}} \sum_{\ell=1}^{K-1} M^{(\ell)} \sum_{t\in E_\ell} x_t \varepsilon_t \,, \tag{17}$$

with $M^{(\ell)}$ constructed from the solutions to optimization (15) for $\ell = 1, \ldots, K-1$. Also, recall that $\varepsilon = (\zeta_{d+1,i}, \zeta_{d+2,i} \ldots, \zeta_{T,i})$ by equation (12).

In Figure 5, we show the QQ-plot, PP-plot and histogram of W_1^{on} and W_1^{off} (corresponding to the entry $(1,1)$ of matrix A_1) for 1000 different realizations of the noise ζ_t. As we observe, even the noise component W^{off} is biased because the offline construction of M depends on all features x_t and hence on endogenous noise ζ_t. However, the online construction of decorrelating matrices $M^{(\ell)}$, makes the noise term a martingale and hence W^{on} converges in distribution to a zero mean normal vector, allowing for a distributional characterization of the online debiased estimator.



(a) Heat map of Σ

(b) Heat map of Ω

Figure 4: Heat maps of magnitudes of elements of covariance matrix $\Sigma \equiv \mathbb{E}(x_i x_i^T)$ (left plot), and precision matrix $\Omega = \Sigma^{-1}$ (right plot). In this example. x_i's are generated from a $\mathsf{VAR}(d)$ model with covariance matrix of noise $\Sigma_\zeta(i,j) = \rho^{|i-j|}$ with values $d = 5$, $p = 15$, $T = 60$, $\rho = 0.5$, and diagonal $A^{(i)}$ matrices with $b = 0.15$ on diagonals.

3.2 Distributional characterization of online debiasing

We start our analysis of the online debiased estimator $\widehat{\theta}^{\mathsf{on}}$ by considering a bias-variance decomposition. Using $y_t = \langle x_t, \theta_0 \rangle + \varepsilon_t$ in the definition (16):

$$\begin{aligned}\widehat{\theta}^{\mathsf{on}} - \theta_0 &= \widehat{\theta}^{\mathsf{L}} - \theta_0 + \frac{1}{n}\sum_{\ell=1}^{K-1}\sum_{t\in E_\ell} M^{(\ell)} x_t x_t^{\mathsf{T}}(\theta_0 - \widehat{\theta}^{\mathsf{L}}) + \frac{1}{n}\sum_{\ell=1}^{K-1}\sum_{t\in E_\ell} M^{(\ell)} x_t \varepsilon_t \\ &= \Big(I - \frac{1}{n}\sum_{\ell=1}^{K-1}\sum_{t\in E_\ell} M^{(\ell)} x_t x_t^{\mathsf{T}}\Big)(\widehat{\theta}^{\mathsf{L}} - \theta_0) + \frac{1}{n}\sum_{\ell=1}^{K-1}\sum_{t\in E_\ell} M^{(\ell)} x_t \varepsilon_t \,. \end{aligned} \tag{18}$$

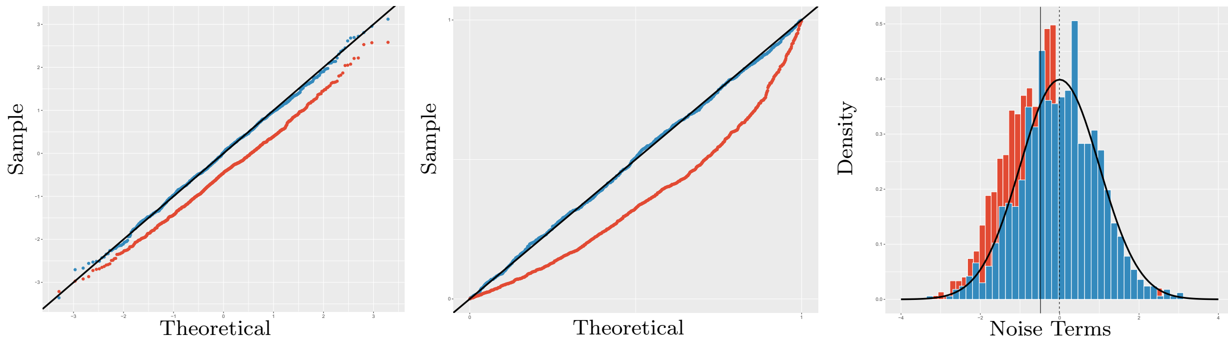
With the shorthand $R^{(\ell)} = (1/r_\ell)\sum_{t\in E_\ell} x_t x_t^{\mathsf{T}}$ for the sample covariance of features in episode ℓ and the bias B_n and variance term W_n below

$$B_n \equiv \sqrt{n}\Big(I - \frac{1}{n}\sum_{\ell=1}^{K-1} r_\ell M^{(\ell)} R^{(\ell)}\Big)\,, \tag{19}$$

$$W_n \equiv \frac{1}{\sqrt{n}}\sum_{\ell=1}^{K-1} M^{(\ell)}\Big(\sum_{t\in E_\ell} x_t \varepsilon_t\Big)\,, \tag{20}$$

we arrive at the following decomposition

$$\widehat{\theta}^{\mathsf{on}} = \theta_0 + \frac{1}{\sqrt{n}}\big(B_n(\widehat{\theta}^{\mathsf{L}} - \theta_0) + W_n\big)\,. \tag{21}$$



(a) (b) (c)

Figure 5: Plots 5a, 5b, and 5c show the QQ plots, PP plots, and the histogram of online debiased noise terms (blue) and offline debiased noise terms (red) over 1000 independent experiments, respectively and black curve/lines denote the ideal standard normal distribution. The solid and dash vertical lines in plot (c) indicate the location of the mean of offline and online debiased noise terms, respectively.

Our first set of results concern the bias of $\widehat{\theta}^{\sf on}$, establishing that this is asymptotically smaller than that of the LASSO estimate. The analysis of the bias focuses mostly on the term B_n, which in turn, is controlled by the parameter μ_ℓ in the optimization (15). We would like to choose μ_ℓ small enough to reduce the bias, but large enough so that the optimization (15) is still feasible. The following lemma shows that, with high probability, μ_ℓ of order $\omega\sqrt{\log(dp)/n_\ell}$ is sufficient to make the optimization feasible.

Lemma 3.3. *Let $\Omega = \Sigma^{-1} = (\mathbb{E}\{x_t x_t^{\sf T}\})^{-1}$ be the precision matrix of the time series. There exists universal constants C, C' such that the following happens. Suppose that $n_\ell \ge C\omega^2 \log(dp)$ where ω is defined in Theorem 3.2. Then with probability $1-(dp)^{-6}$:*

$$\max_{i,j} |\Omega\widehat{\Sigma}^{(\ell)} - \mathbb{I}(i=j)| \le C'\omega\sqrt{\frac{\log(dp)}{n_\ell}}.$$

The proof of Lemma 3.3 is given in Appendix A.3. The following theorem uses Lemma 3.3 to control the bias of the online debiased estimator.

Theorem 3.4. *(Bias control) Consider the $\mathsf{VAR}(d)$ model (10) and let $\widehat{\theta}^{\sf on}$ be the debiased estimator (16) where the decorrelating matrices $M^{(\ell)}$ are computed according to Eq.(15), with $\mu_\ell = c_1\omega\sqrt{(\log(dp)/n_\ell}$ and $L \ge \|\Omega\|_1$. Further assume that the base estimator is $\widehat{\theta}^{\sf L}$ computed with $\lambda = \lambda_0\sqrt{\log(dp)/n}$ where $\lambda_0 \ge 4\lambda_{\max}(\Sigma_\zeta)(1\vee\mu_{\max}(\mathcal{A}))/\mu_{\min}(\mathcal{A})$.*

Then, under the sample size condition $n \ge C\omega^2 s_0 \log(dp)$, we have

$$\sqrt{n}(\widehat{\theta}^{\sf on} - \theta_0) = W_n + \Delta_n, \tag{22}$$

where $\mathbb{E}\{W_n\} = 0$ and

$$\mathbb{P}\Big\{\|\Delta_n\|_\infty \ge C_1 \frac{\lambda_0(\omega + L\gamma)}{\alpha}\frac{s_0\log(dp)}{\sqrt{n}}\Big\} \le (dp)^{-4}, \tag{23}$$

The parameters ω, α are defined in Theorem 3.2, and $\gamma = d\lambda_{\max}(\Sigma_\zeta)/\mu_{\min}(\mathcal{A})$. Further, the bias satisfies

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

We refer to Appendix A.4 for the proof of Theorem 3.4.

Note that he above theorem bounds the bias term Δ_n for finite sample size n. To study these bounds in an asymptotic regime, we make the following assumption to simplify our presentation.

Assumption 3.5. *Suppose that*

1. *The parameters $\lambda_{\min}(\Sigma_\zeta)$, $\lambda_{\max}(\Sigma_\zeta)$, $\mu_{\min}(\mathcal{A})$ and $\mu_{\max}(\mathcal{A})$ are bounded away from 0 and ∞, as $n, p \to \infty$.*
2. *With $\Omega = \Sigma^{-1} = (\mathbb{E}\{x_t x_t^{\mathsf{T}}\})^{-1}$ the precision matrix of the data points $\{x_t\}$, and s_0 the sparsity of $\theta_0 = (A_i^{(1)}, \dots, A_i^{(d)})^{\mathsf{T}}$, we assume that $\|\Omega\|_1 = o(\sqrt{n}/\log(dp))$.*

Under Assumption 3.5 the spectral quantities ω, γ, α and (therefore) λ_0 are order one. We can also ignore the lower order term $\|\theta_0\|_1/(dp)^6$ in the high-dimensional regime. Indeed, the denominator $(dp)^6$ can be changed to $(dp)^c$ for arbitrary large $c > 0$, by adjusting constant C_1 and the tail bound in Eq.(23). Therefore, as far as $\|\theta_0\|_1$ grows polynomially at p, then this term vanishes asymptotically. The theorem, hence, shows that the bias of the online debiased estimator is of order $Ls_0(\log p)/n$. On the other hand, recall the filtration $\mathcal{F}_t$ generated by $\{\varepsilon_1, \dots, \varepsilon_t\}$ and rewrite (20) as $W_n = \sum_t v_t\varepsilon_t$, where $v_t = M^{(\ell)}x_t/\sqrt{n}$ (Sample t belongs to episode ℓ). We use Assumption 3.5 in Lemma 3.6 below, to show that for each coordinate $i \in [dp]$, the conditional variance $\sum_{t=1}^n \mathbb{E}(\varepsilon_t^2 v_{t,i}^2|\mathcal{F}_{t-1}) = (\sigma^2/n)\sum_{t=1}^n \langle m_a^\ell, z_t\rangle^2$ is of order one. Hence $\|\Delta_n\|_\infty$ is asymptotically dominated by the noise variance when $s_0 = o\left(\frac{\sqrt{n}}{L\log(dp)}\right)$.

Another virtue of Lemma 3.6 is that it shows the martingale sum W_n is stable in an appropriate sense. This is a key technical step that allows us to characterize the distribution of the noise term W_n by applying the martingale CLT (e.g., see [HH14, Corollary 3.2]) and conclude that the unbiased component W_n admits a Gaussian limiting distribution.

Lemma 3.6. *(Stability of martingale W_n) Let $\widehat{\theta}^{\mathsf{on}}$ be the debiased estimator (16) with $\mu_\ell = \tau\sqrt{(\log p)/n_\ell}$ and $L = L_0\|\Omega\|_1$, for an arbitrary constant $L_0 \ge 1$. Under Assumption 3.5, and for any fixed sequence of integers $a(n) \in [dp]$,*[3] *we have*

$$V_{n,a} \equiv \frac{\Sigma_{\zeta_{i,i}}}{n}\sum_{\ell=1}^{K-1}\sum_{t\in E_\ell}\langle m_a^\ell, x_t\rangle^2 = \Sigma_{\zeta_{i,i}} \cdot \Omega_{a,a} + o_P(1). \tag{24}$$

In addition, we have

$$\max\left\{\frac{1}{\sqrt{n}}|\langle m_a^\ell, x_t\rangle\varepsilon_t| : \ell \in [K-1],\, t \in [n-1]\right\} = o_P(1). \tag{25}$$

[3]We index the sequence with the sample size n that is diverging. Since we are in high-dimensional setting $p \ge n$ is also diverging.

We refer to Appendix A.5 for the proof of Lemma 3.6. With Lemma 3.6 in place, we can apply a martingale central limit theorem [HH14, Corollary 3.2] to obtain the following result.

Corollary 3.7. *Consider the* $\mathsf{VAR}(d)$ *model* (10) *for time series and let* $\widehat{\theta}^{\mathsf{on}}$ *be the debiased estimator* (16) *with* $\mu_\ell = C_1\omega\sqrt{(\log p)/n_\ell}$ *and* $L = L_0\|\Omega\|_1$*, for an arbitrary constant* $L_0 \geq 1$*. For any fixed sequence of integers* $a(n) \in [dp]$*, define the conditional variance* V_n *as*

$$V_{n,a} \equiv \frac{\Sigma_{\zeta_{i,i}}}{n} \sum_{\ell=1}^{K-1} \sum_{t \in E_\ell} \langle m_a^\ell, x_t \rangle^2 \,.$$

Under Assumption 3.5, for any fixed coordinate $a \in [dp]$*, and for all* $x \in \mathbb{R}$ *we have*

$$\lim_{n\to\infty} \mathbb{P}\Big\{\frac{W_{n,a}}{\sqrt{V_{n,a}}} \leq x\Big\} = \Phi(x)\,, \tag{26}$$

where Φ *is the standard Gaussian cdf.*

For the task of statistical inference, Theorem 3.4 and Corollary 3.7 suggest to consider the scaled residual $\sqrt{n}(\widehat{\theta}^{\mathsf{on}}_a - \theta_{0,a})/\sqrt{V_{n,a}}$ as the test statistics. Our next proposition characterizes its distribution. The proof is straightforward given the result of Theorem 3.4 and Corollary 3.7 and is deferred to Appendix A.6. In its statement we omit explicit constants that can be easily derived from Theorem 3.4.

Theorem 3.8. *Consider the* $\mathsf{VAR}(d)$ *model* (10) *for time series and let* $\widehat{\theta}^{\mathsf{on}}$ *be the debiased estimator* (16) *with* $\mu_\ell = C_1\omega\sqrt{(\log p)/n_\ell}$*,* $\lambda = \lambda_0\sqrt{\log(dp)/n}$*, and* $L = L_0\|\Omega\|_1$*, for an arbitrary constant* $L_0 \geq 1$*. Suppose that Assumption 3.5 holds and* $s_0 = o\left(\frac{\sqrt{n}}{\|\Omega\|_1 \log(dp)}\right)$*, then the following holds true for any fixed sequence of integers* $a(n) \in [dp]$*. For all* $x \in \mathbb{R}$*, we have*

$$\lim_{n\to\infty} \left|\mathbb{P}\left\{\frac{\sqrt{n}(\widehat{\theta}^{\mathsf{on}}_a - \theta_{0,a})}{\sqrt{V_{n,a}}} \leq x\right\} - \Phi(x)\right| = 0\,. \tag{27}$$

4 Batched data collection

Recall the stylized setting of adaptive data collection in batches from Section 1.1, where the samples naturally separate into two batches: the first n_1 data points where the covariates are i.i.d from a distribution $\mathbb{P}_x$, and the second batch of n_2 data points, where the covariates x_i are drawn independently from the law of x_1, conditional on the event $\{\langle x_1, \widehat{\theta}^1\rangle \geq \varsigma\}$, where ς is a potentially data-dependent threshold. The following theorem is a version of Theorem 6.1 in [BVDG11] and is proved in an analogous manner. It demonstrates that even with adaptive data collection consistent estimation using the LASSO is possible.

Theorem 4.1 ([BVDG11, Theorem 6.1])**.** *Suppose that the true parameter* θ_0 *is* s_0*-sparse and the distribution* $\mathbb{P}_x$ *is such that with probability one the following two conditions hold: (i) the covariance* $\mathbb{E}\{xx^\mathsf{T}\}$ *and* $\mathbb{E}\{xx^\mathsf{T}|\langle x, \widehat{\theta}^1\rangle \geq \varsigma\}$ *are* $(\phi_0, \mathrm{supp}(\theta_0))$*-compatible and (ii)* x *as well as* $x|_{\langle x,\widehat{\theta}^1\rangle\geq\varsigma}$ *are* κ*-subgaussian. Suppose that* $n \geq C_1(\kappa^4/\phi_0^2)s_0^2\log p$*. Then, the LASSO estimate* $\widehat{\theta}^{\mathsf{L}}(y, X; \lambda_n)$ *with* $\lambda_n = C_2\kappa\sigma\sqrt{(\log p)/n}$ *satisfies, with probability exceeding* $1 - p^{-3}$*,*

$$\|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1 \leq \frac{C' s_0 \lambda_n}{\phi_0} = \frac{C\kappa\sigma}{\phi_0} s_0 \sqrt{\frac{\log p}{n}}.$$

Remark 4.2. *(Estimating the noise variance) For the correct estimation rate using the LASSO, Theorem 4.1 requires knowledge of the noise level σ, which is used to calibrate the regularization λ_n. Other estimators like the scaled LASSO [SZ12] or the square-root LASSO [BCW11] allow to estimate σ consistently when it is unknown. This can be incorporated into the present setting, as done in [JM14a]. For simplicity, we focus on the case when the noise level is known. However, the results hold as far as a consistent estimate of σ is used. Formally, a consistent estimator refers to an estimate $\widehat{\sigma} = \widehat{\sigma}(y, X)$ of the noise level satisfying, for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} \sup_{\|\theta_0\|_0 \le s_0} \mathbb{P}\left(\left|\frac{\widehat{\sigma}}{\sigma} - 1\right| \ge \varepsilon\right) = 0\,. \tag{28}$$

Remark 4.3. *At the expense of increasing the absolute constants in Theorem 4.1, the probability $1 - p^{-3}$ can be made $1 - p^{-C}$ for any arbitrary constant $C > 1$.*

Let X_1 and X_2 denote the design matrices of the two batches and, similarly, $y^{(1)}$ and $y^{(2)}$ the two responses vectors. In this setting, we use an online debiased estimator as follows:

$$\widehat{\theta}^{\mathsf{on}} = \widehat{\theta}^{\mathsf{L}} + \frac{1}{n} M^{(1)} X_1^{\mathsf{T}} (y^{(1)} - X_1 \widehat{\theta}^{\mathsf{L}}) + \frac{1}{n} M^{(2)} X_2^{\mathsf{T}} (y^{(2)} - X_2 \widehat{\theta}^{\mathsf{L}}), \tag{29}$$

where we will construct $M^{(1)}$ as a function of X_1 and $M^{(2)}$ as a function of X_1 as well as X_2. The proposal in Eq.(29) follows from the general recipe in Eq.(8) by setting

- $M_i = M^{(1)}$ for $i = [n_1]$ and $M_i = M^{(2)}$ for $i = n_1 + 1, \dots, n$.
- Filtrations $\mathfrak{F}_i$ constructed as follows. 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 $\mathfrak{F}_i$. For $i \ge n_1$, $y_1, \dots, y_i$, $x_1, \dots, x_n$ and $\varepsilon_1, \dots \varepsilon_i$ are measurable with respect to $\mathfrak{F}_i$.

By construction, this choice satisfies the predictability condition, given by Definition 2.1.

Note that Eq.(29) nests an intuitive 'sample splitting' approach. Indeed, debiasing $\widehat{\theta}^{\mathsf{L}}$ using exactly one of the two batches is equivalent to setting one of $M^{(1)}$ or $M^{(2)}$ to 0. While sample splitting can be shown to work under appropriate conditions, our approach is more efficient with use of the data and gains power in comparison. We construct $M^{(1)}$ and $M^{(2)}$ using a modification of the program used in [JM14a]. Let $\widehat{\Sigma}^{(1)} = (1/n_1) X_1^{\mathsf{T}} X_1$ and $\widehat{\Sigma}^{(2)} = (1/n_2) X_2^{\mathsf{T}} X_2$ be the sample covariances of each batch; let $M^{(1)}$ have rows $(m_a^{(1)})_{1 \le a \le p}$ and similarly for $M^{(2)}$. Using parameters $\mu_\ell, L > 0$ that we set later, we choose $m_a^{(\ell)}$, the a^{th} row of $M^{(\ell)}$, as a solution to the program

$$\begin{aligned} &\text{minimize} \quad \langle m, \widehat{\Sigma}^{(\ell)} m \rangle \\ &\text{subject to} \quad \|\widehat{\Sigma}^{(\ell)} m - e_a\|_\infty \le \mu_\ell, \ \ \|m\|_1 \le L. \end{aligned} \tag{30}$$

Here e_a is the a^{th} basis vector: a vector which is one at the a^{th} coordinate and zero everywhere else.

The intuition for the program (30) is simple. The first constraint ensures that $\widehat{\Sigma}^{(\ell)} m$ is close, in ℓ_∞ sense to the e_a, the a^{th} basis vector and as we will see in Theorem 4.6 it controls the bias term Δ of $\widehat{\theta}^{\mathsf{on}}$. The objective is a multiple of the variance of the martingale term W in $\widehat{\theta}^{\mathsf{on}}$ (cf. Eq. (34)). We wish to minimize this as it directly affects the power of the test statistic or the length of valid confidence intervals constructed based on $\widehat{\theta}^{\mathsf{on}}$. The ℓ_1 constraint on m, which is missing in [JM14a], is crucial for our adaptive data setting. This constraint ensures that the value of the program $\langle m_a^{(\ell)}, \widehat{\Sigma}^{(\ell)} m_a^{(\ell)} \rangle$ is stable, and does not fluctuate much from sample to sample (this is formalized as the 'stability condition' in Lemmas C.8 and 3.6). It is this stability that ensures that the martingale part of the residual displays a central limit behavior.

Note that in the non-adaptive setting, inference can be performed conditional on design X, and fluctuation in $\langle m_a^{(\ell)}, \widehat{\Sigma}^{(\ell)} m_a^{(\ell)} \rangle$ is conditioned out. In the adaptive setting, this is not possible: one effectively cannot condition on the design without conditioning on the noise realization ε, and therefore we perform inference unconditionally on X.

4.1 Online debiasing: a distributional characterization

We begin the analysis of the online debiased estimator $\widehat{\theta}^{\mathsf{on}}$ by a decomposition that mimics the classical debiasing.

$$\widehat{\theta}^{\mathsf{on}} = \theta_0 + \frac{1}{\sqrt{n}} \big(B_n (\widehat{\theta}^{\mathsf{L}} - \theta_0) + W_n \big), \tag{31}$$

$$B_n = \sqrt{n} \Big(I_p - \frac{n_1}{n} M^{(1)} \widehat{\Sigma}^{(1)} - \frac{n_2}{n} M^{(2)} \widehat{\Sigma}^{(2)} \Big) \tag{32}$$

$$W_n = \frac{1}{\sqrt{n}} \sum_{i \le n_1} M^{(1)} x_i \varepsilon_i + \frac{1}{\sqrt{n}} \sum_{n_1 < i \le n} M^{(2)} x_i \varepsilon_i. \tag{33}$$

Assumption 4.4. *(Requirements of design) Suppose that the distribution $\mathbb{P}_x$ and the intermediate estimate $\widehat{\theta}^1$, that is used in collecting the second batch, satisfy the following:*

1. *There exists a constant $\Lambda_0 > 0$ so that the eigenvalues of $\mathbb{E}\{xx^{\mathsf{T}}\}$ and $\mathbb{E}\{xx^{\mathsf{T}} | \langle x, \widehat{\theta}^1 \rangle \ge \varsigma\}$ are bounded below by Λ_0.*
2. *The laws of x and $x|_{\langle x, \widehat{\theta}^1 \rangle \ge \varsigma}$ are κ-subgaussian for a constant $\kappa > 0$.*
3. *The precision matrices $\Omega = \mathbb{E}\{xx^{\mathsf{T}}\}^{-1}$ and $\Omega^{(2)}(\widehat{\theta}^1) = \mathbb{E}\{xx^{\mathsf{T}} | \langle x, \widehat{\theta}^1 \rangle \ge \varsigma\}^{-1}$ satisfy $\|\Omega\|_1 \vee \|\Omega^{(2)}(\widehat{\theta}^1)\|_1 \le L$.*
4. *The conditional covariance $\Sigma^{(2)}(\theta) = \mathbb{E}\{xx^{\mathsf{T}} | \langle x, \theta \rangle \ge \varsigma\}$ is K-Lipschitz in its argument θ, i.e. $\|\Sigma^{(2)}(\theta') - \Sigma^{(2)}(\theta)\|_\infty \le K \|\theta - \theta'\|_1$.*

The first two conditions of Assumption 4.4 are for ensuring that the base LASSO estimator $\widehat{\theta}^{\mathsf{L}}$ has small estimation error. In addition, our debiasing makes use of the third and fourth constraints on the precision matrices of the sampling distributions. In the above, we will typically allow $L = L_n$ to diverge with n.

In the following Example we show that Gaussian random designs satisfy all the conditions of Assumption 4.4. We refer to Section C.4 for its proof.

Example 4.5. *Let $\mathbb{P}_x = \mathsf{N}(0, \Sigma)$ and $\widehat{\theta}$ be any vector such that $\|\widehat{\theta}\|_1 \|\widehat{\theta}\|_\infty \le L_\Sigma \lambda_{\min}(\Sigma) \|\widehat{\theta}\|/2$ and $\|\Sigma^{-1}\|_1 \le L_\Sigma/2$. Then the distributions of x and $x|_{\langle x, \widehat{\theta} \rangle \ge \varsigma}$, with $\varsigma = \bar{\varsigma} \langle \widehat{\theta}, \Sigma \widehat{\theta} \rangle^{1/2}$ for a constant $\bar{\varsigma} \ge 0$ satisfy the conditions of Assumption 4.4 with*

$$\Lambda_0 = \lambda_{\min}(\Sigma), \quad \kappa = 3\lambda_{\max}^{1/2}(\Sigma)(\bar{\varsigma} \vee \bar{\varsigma}^{-1}), \quad K = \sqrt{8}(1 + \bar{\varsigma}^2) \frac{\lambda_{\max}(\Sigma)^{3/2}}{\lambda_{\min}(\Sigma)^{1/2}}, \quad L = L_\Sigma.$$

Under Assumption 4.4 we provide a *non-asymptotic* bound on the bias of the online debiased estimator $\widehat{\theta}^{\mathsf{on}}$.

Theorem 4.6. *(Non-asymptotic bound on bias) Under Assumption 4.4, there exists universal constants C_1, C_2, C_3 so that, when $n \geq C_1 \kappa^4 s_0^2 \log p/\phi_0^2$ and $n_1 \wedge n_2 \geq C_1(\Lambda_0/\kappa^2 + \kappa^2/\Lambda_0)\log p$, we have that*

$$\sqrt{n}(\widehat{\theta}^{\mathsf{on}} - \theta_0) = W_n + \Delta_n, \tag{34}$$

where $\mathbb{E}\{W_n\} = 0$ and

$$\mathbb{P}\Big\{\|\Delta_n\|_\infty \geq \frac{C_2\kappa^2}{{\Lambda_0}^{3/2}}\frac{\sigma s_0 \log p}{\sqrt{n}}\Big\} \leq p^{-3}. \tag{35}$$

Further we have

$$\|\mathbb{E}\{\widehat{\theta}^{\mathsf{on}} - \theta_0\}\|_\infty \leq \frac{C_2\kappa^2}{{\Lambda_0}^{3/2}}\frac{\sigma s_0 \log p}{n} + \frac{C_3\|\theta_0\|_1}{p^2}\,. \tag{36}$$

The proof of Theorem 4.6 is given in Appendix C.2. Note that, in the high-dimensional setting of $n \ll p$, the term $\|\theta_0\|_1/p^2$ will be of lower order as compared to $s_0 \log p/n$. Therefore, when the parameters $\Lambda_0, \sigma, \kappa$ are of order one, the theorem shows that the bias of the online debiased estimator is of order $s_0 \log p/n$, This may be compared with the LASSO estimator $\widehat{\theta}^{\mathsf{L}}$ whose bias is typically of order $\lambda \asymp \sigma\sqrt{\log p/n}$. In particular, in the regime when $s_0 = o(\sqrt{n/\log p})$, this bias is asymptotically dominated by the variance, which is of order $\sigma/\sqrt{n}$.

In order to establish asymptotic Gaussian behavior of the online debiased estimate $\widehat{\theta}^{\mathsf{on}}$, we consider a specific asymptotic regime for the problem instances.

Assumption 4.7. *(Asymptotic regime) We consider problem instances indexed by the sample size n, where n, p, s_0 satisfy the following:*

1. *$\liminf_{n\to\infty} \frac{n_1 \wedge n_2}{n} \geq c$, for a positive universal constant $c \in (0, 1]$. In other words, both batches contain at least a fixed fraction of data points.*
2. *The parameters satisfy:*

$$\lim_{n\to\infty} \frac{1}{\Lambda_0} s_0 \sqrt{\frac{\log p}{n}}\left(L^2 K \vee \sqrt{\frac{\log p}{\Lambda_0}}\right) = 0\,. \tag{37}$$

The following proposition establishes that in the asymptotic regime, the unbiased component W_n has a Gaussian limiting distribution. The key underlying technical idea is to ensure that the martingale sum in W_n is stable in an appropriate sense.

Proposition 4.8. *Suppose that Assumption 4.4 holds and consider the asymptotic regime of Assumption 4.7. Let $a = a(n) \in [p]$ be a fixed sequence of coordinates. Define the conditional variance $V_{n,a}$ of the a^{th} coordinate as*

$$V_{n,a} = \sigma^2\Big(\frac{n_1}{n}\langle m_a^{(1)}, \widehat{\Sigma}^{(1)} m_a^{(1)}\rangle + \frac{n_2}{n}\langle m_a^{(2)}, \widehat{\Sigma}^{(2)} m_a^{(2)}\rangle\Big)\,. \tag{38}$$

Then, for any bounded continuous $\varphi : \mathbb{R} \to \mathbb{R}$

$$\lim_{n\to\infty} \mathbb{E}\Big\{\varphi\Big(\frac{W_{n,a}}{\sqrt{V_{n,a}}}\Big)\Big\} = \mathbb{E}\{\varphi(\xi)\},$$

where $\xi \sim \mathsf{N}(0,1)$*. The same holds for* φ *being a step function* $\varphi(z) = \mathbb{I}(z \le x)$ *for any* $x \in \mathbb{R}$*. In particular,*

$$\lim_{n\to\infty} \mathbb{P}\Big\{\frac{W_{n,a}}{\sqrt{V_{n,a}}} \le x\Big\} = \Phi(x),$$

where Φ *is the standard Gaussian cdf.*

The proof of Proposition 4.8 is deferred to Appendix C.3. The combination of Theorem 4.6 and Proposition 4.8 immediately yields the following distributional characterization for $\widehat{\theta}^{\mathsf{on}}$.

Theorem 4.9. *Under Assumptions 4.4 and 4.7, the conclusion of Proposition 4.8 holds with* $\sqrt{n}(\widehat{\theta}^{\mathsf{on}}_a - \theta_0)$ *in place of* W_n*. In particular,*

$$\lim_{n\to\infty} \mathbb{P}\Big\{\sqrt{\frac{n}{V_{n,a}}}(\widehat{\theta}^{\mathsf{on}}_a - \theta_{0,a}) \le x\Big\} = \Phi(x), \tag{39}$$

where $V_{n,a}$ *is defined as in Proposition 4.8.*

To compare the sample size requirements made for ℓ_1-consistent estimation and those in Assumption 4.7, it is instructive to simplify to the case when $\kappa, \phi_0, \Lambda_0$ are of order one. Then ℓ_1-consistency (Theorem 4.1 in Appendix C) requires that $n_1 \vee n_2 = \Omega(s_0^2 \log p)$, i.e. at least one of the batches is larger than $s_0^2 \log p$. However, Theorem 4.9 makes the same assumption on $n_1 \wedge n_2$, or both batches exceed $s_0^2 \log p$ in size. For online debiasing, this is the case of interest. Indeed if $n_1 \gg n_2$ (or vice versa), we can apply offline debiasing to the larger batch to obtain a debiased estimate. Conversely, when n_1 and n_2 are comparable as in Assumption 4.7, this 'sample-splitting' approach leads to loss of power corresponding to a constant factor reduction in the sample size. This is the setting addressed in Theorem 4.9 via online debiasing.

4.1.1 Revisiting the numerical example from Section 1.1.

In the batched data example discussed in Section 1.1, we observed that the classical offline debiasing fails in providing unbiased estimate of the true parameters. Here, we will repeat the same experiment and numerically characterize the distribution of the proposed online debiased estimator.

Figure 6 (left panel) shows the histogram of the entries of online debiased estimator $\widehat{\theta}^{\mathsf{on}}$ on the support of θ_0 (blue) along with the corresponding histogram of entries of the debiased estimator $\widehat{\theta}^{\mathsf{off}}$ (red). As we see for both choices of $\widehat{\theta}^1$ (debiased LASSO and ridge estimate on the first batch), the online debiased estimator $\widehat{\theta}^{\mathsf{on}}$ is appropriately centered around the true coefficients.

One can also split samples in the following way. Since the second batch of data was adaptively collected while the first batch was not, we can compute a debiased estimate using only the first, non-adaptive batch:

$$\widehat{\theta}^{\mathsf{off},1} \equiv \widehat{\theta}^{\mathsf{L}}(y^{(1)}, X_1) + \frac{1}{n}\Omega X_1^{\mathsf{T}}(y^{(1)} - X_1\widehat{\theta}^{\mathsf{L}}(y^{(1)}, X_1)). \tag{40}$$

Figure 6 (right panel) shows the histogram of the entries of $\widehat{\theta}^{\mathsf{off},1}$ restricted to the support of θ_0, and the comparison with $\widehat{\theta}^{\mathsf{on}}$. As can be expected, both $\widehat{\theta}^{\mathsf{off},1}$ and $\widehat{\theta}^{\mathsf{on}}$ are appropriately centered around the true coefficient 1. However, as is common with sample-splitting, $\widehat{\theta}^{\mathsf{off},1}$ displays a larger variance and correspondingly loses power in comparison with $\widehat{\theta}^{\mathsf{on}}$ since it uses only half of the data. The power loss becomes even more pronounced when there are more than two phases of data collection, or if the phases are particularly imbalanced.

Comparison with ridge-type debiasing approach of [DMST18]. This work studies a similar problem, namely performing statistical inference using adaptively collected data using a debiasing approach. To compare with our setting, there are two important points to note:

1. The method of [DMST18] is tailored to low-dimensional setting where the number of covariates p is less than the sample size $(p < n)$. More specifically, denoting by $\lambda_{\min}(n)$ the minimum eigenvalue of $X^{\mathsf{T}}X$, [DMST18] considers a setting where $\lambda_{\min}(n) \to \infty$ almost surely. Note that for the batched data example, this amounts to $\sqrt{n} - \sqrt{p} \to \infty$.

2. The work [DMST18] proposes a different method of debiasing which albeit being valid in low-dimensional setting it comes with fundamental challenges to be generalized to high-dimensional setting. Letting $\widehat{\theta}^{\mathsf{OLS}}$ the least square estimator, [DMST18] constructs a debiased estimator $\widehat{\theta}^{\mathsf{d}}$ as follows:

$$\widehat{\theta}^{\mathsf{d}} = \widehat{\theta}^{\mathsf{OLS}} + W_n(y - X\widehat{\theta}^{\mathsf{OLS}})\,, \tag{41}$$

   where the matrix W_n is constructed recursively as $W_n = [W_{n-1}|w_n]$ and $X_n = [X_{n-1}|x_n]$ with

$$w_n = \arg\min_{w\in\mathbb{R}^p} \|I - W_{n-1}X_{n-1} - wx_n^{\mathsf{T}}\|_F^2 + \lambda\|w\|_2^2\,. \tag{42}$$

   Therefore, the decorrelating matrix W_n is constructed in an online way as it is a predictable sequence according to Definition 2.1. Note that w_i corresponds to $M_i x_i$ in our notation.

One can potentially think of using the ridge-type debiased estimator (42) in high-dimensional setting with using $\widehat{\theta}^{\mathsf{L}}$ instead of $\widehat{\theta}^{\mathsf{OLS}}$. In Figure 6, we include the histogram of such estimate (gray histogram under the name "ridgeOnline"). As we see the corresponding histogram is biased and deviates from a normal distribution which implies that this approach does not extend to high-dimensional setting.

Some intuition for this may be seen by following the argument of [DMST18]. Considering the bias-variance decomposition of $\widehat{\theta}^{\mathsf{d}} - \theta_0 = \mathsf{b} + \mathsf{v}$ with $\mathsf{b} = (I - W_nX_n)(\widehat{\theta}^{\mathsf{OLS}} - \theta_0)$ and $\mathsf{v} = W_n\varepsilon_n$, the above optimization aims at minimizing a weighted sum of the bias and the variance of $\widehat{\theta}^{\mathsf{d}}$ in an online manner. The analysis of [DMST18] controls bias as follows

$$\|\mathsf{b}\| \le \|I - W_nX_n\|_{\mathrm{op}}\, \|\widehat{\theta}^{\mathsf{OLS}} - \theta_0\|_2 \le \|I - W_nX_n\|_F\, \|\widehat{\theta}^{\mathsf{OLS}} - \theta_0\|_2\,.$$

However, in high-dimension this bound is vacuous. Since $W_nX_n \in \mathbb{R}^{p\times p}$ is of rank at most $n < p$, $I - W_nX_n$ has eigenvalue 1 with multiplicity at least $p - n$. Therefore $\|I - W_nX_n\|_F \ge p - n \to \infty$ and $\|I - W_nX_n\|_{\mathrm{op}} \ge 1$. Thus, even a refinement of [DMST18] would only yield an insufficient bias bound of the type

$$\|\mathsf{b}\|_2 \le \|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_2 \approx \sigma\sqrt{\frac{s_0 \log p}{n}}\,,$$

which dominants the variance component $\mathrm{Var}(\mathsf{v}) = O(1/\sqrt{n})$. Our scheme of online debiasing overcomes this obstacle by adapting to the geometry of the high-dimensional regime. In particular, it yields the bias bound of order $\|\mathbb{E}\{\widehat{\theta}^{\mathsf{on}} - \theta_0\}\|_\infty = O(s_0(\log p)/n)$ which is dominated by the noise term, provided that $s_0 = o(\sqrt{n}/\log p)$.

5 Statistical inference

An immediate use of distributional characterizations (27) or (39) is to construct confidence intervals and also provide valid p-values for hypothesis testing regarding the model coefficients. Throughout, we make the sparsity assumption $s_0 = o(\sqrt{n}/\log p_0)$, with p_0 the number of model parameters (for the batched data collection setting $p_0 = p$, and for the $\mathsf{VAR}(d)$ model $p_0 = dp$).

Confidence intervals: For fixed coordinate $a \in [p_0]$ and significance level $\alpha \in (0,1)$, we let

$$\begin{aligned} J_a(\alpha) &\equiv [\widehat{\theta}^{\mathsf{on}}_a - \delta(\alpha,n), \widehat{\theta}^{\mathsf{on}}_a + \delta(\alpha,n)]\,, \\ \delta(\alpha,n) &\equiv \Phi^{-1}(1-\alpha/2)\sqrt{V_{n,a}/n}\,, \end{aligned} \tag{43}$$

where $V_{n,a}$ is defined by Equation (24) for the $\mathsf{VAR}(d)$ model and by Equation (38) for the batched data collection setting.

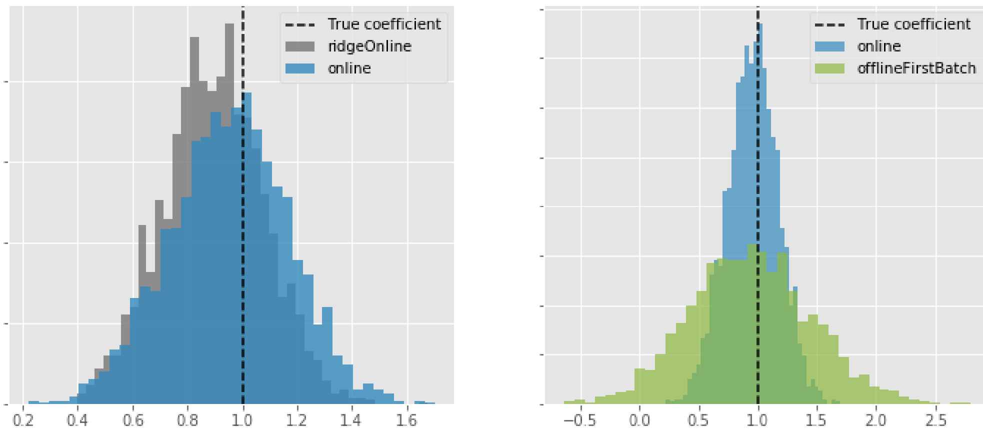
As a result of Proposition 3.8, the confidence interval $J_a(\alpha)$ is asymptotically valid because

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}(\theta_{0,a} \in J_a(\alpha)) &= \lim_{n\to\infty} \mathbb{P}\Big\{\frac{\sqrt{n}(\widehat{\theta}^{\mathsf{on}}_a - \theta_{0,a})}{\sqrt{V_{n,a}}} \le \Phi^{-1}(1-\alpha/2)\Big\} \\ &\quad - \lim_{n\to\infty} \mathbb{P}\Big\{\frac{\sqrt{n}(\widehat{\theta}^{\mathsf{on}}_a - \theta_{0,a})}{\sqrt{V_{n,a}}} \le \Phi^{-1}(1-\alpha/2)\Big\} \\ &= \Phi(\Phi^{-1}(1-\alpha/2)) - \Phi(-\Phi^{-1}(1-\alpha/2)) = 1-\alpha\,. \end{aligned} \tag{44}$$

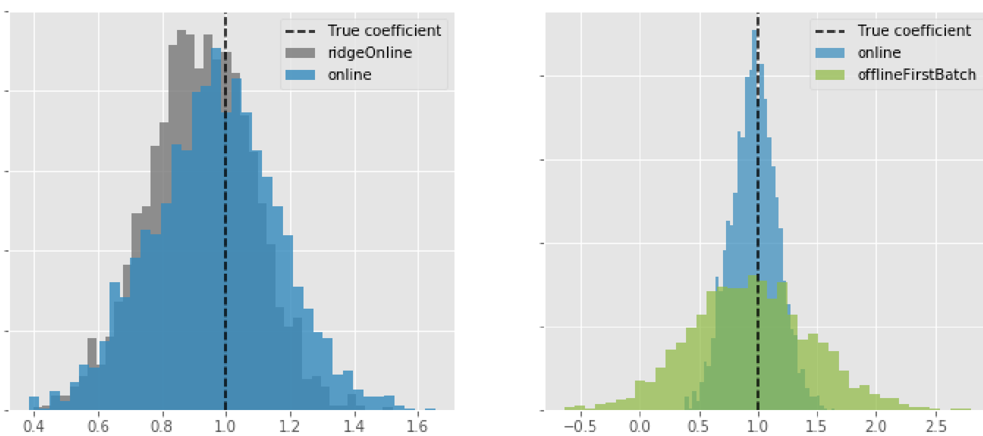
Further, note that the length of confidence interval $J_a(\alpha)$ is of order $O(\sigma/\sqrt{n})$ (using Lemma C.8 for the batched data collection setting and Lemma 3.6 for the time series). It is worth noting that this is the minimax optimal rate [JM14b, Jav14] and is of the same order of the length of confidence intervals obtained by the least-square estimator for the classical regime $n > p$ with i.i.d samples.

Hypothesis testing: Another consequence of Proposition 3.8 is that it allows for testing hypothesis of form $H_0 : \theta_{0,a} = 0$ versus the alternative $H_A : \theta_{0,a} \neq 0$ and provide valid p-values. Recall that θ_0 denotes the model parameters, either for the batched data collection setting or the $\mathsf{VAR}(d)$ model (which encodes the entries $A^{(\ell)}_{i,j}$ in model (10)). Such testing mechanism is of crucial importance in practice as it allows to diagnose the significantly relevant covariates to the outcome. In case of time series, it translates to understanding the effect of a covariate $z_{t-\ell,j}$ on a covariate $z_{t,i}$, and to provide valid statistical measures (p-values) for such associations. We construct two-sided p-values for testing H_0, using our test statistic as follows:

$$P_a = 2\left(1 - \Phi\left(\frac{\sqrt{n}|\widehat{\theta}^{\mathsf{on}}_a|}{\sqrt{V_{n,a}}}\right)\right)\,. \tag{45}$$



(a) with $\widehat{\theta}^1$ the debiased LASSO on first batch



(b) with $\widehat{\theta}^1$ the ridge estimate on first batch

Figure 6: (Left) Histograms of the online debiased estimate $\widehat{\theta}^{\mathsf{on}}$ and the ridge debiased estimator [DMST18], restricted to the support of θ_0. (Right) Histograms of the offline debiased estimate *only using the first batch*, $\widehat{\theta}^{\mathsf{off},1}$ given by (40) and the online debiased estimate $\widehat{\theta}^{\mathsf{on}}$. The dashed line indicates the true coefficient size. Offline debiasing $\widehat{\theta}^{\mathsf{off},1}$ using only the first batch works well (green histograms called offlineFirstBatch), but then loses power in comparison. Online debiasing is cognizant of the adaptivity and debiases without losing power even in the presence of adaptivity.

Our testing (rejection) rule given the p-value P_a is:

$$R(a) = \begin{cases} 1 & \text{if } P_a \le \alpha \quad \text{(reject } H_0) \,, \\ 0 & \text{otherwise} \quad \text{(fail to reject } H_0) \,. \end{cases} \tag{46}$$

Employing the distributional characterizations (39) or (27), it is easy to verify that the constructed p-value P_a is valid in the sense that under the null hypothesis it admits a uniform distribution: $\mathbb{P}_{\theta_{0,a}=0}(P_a \le u) = u$ for all $u \in [0, 1]$.

Group inference In many applications, one may want to do inference for a group of model parameters, $\theta_{0,G} \equiv (\theta_{0,a})_{a\in G}$ simultaneously, rather than the individual inference. This is the case particularly, when the model covariates are highly correlated with each other or they are likely to affect the outcome (in time series application, the future covariate vectors) jointly.

To address group inference, we focus on the time series setting. The setting of batched data collection can be handled in a similar way. We first state a simple generalization of Proposition 3.8 to a group of coordinates with finite size as $n, p \to \infty$. The proof is very similar to the proof of Proposition 3.8 and is omitted.

Lemma 5.1. *Let $G = G(n)$ be a sequence of sets $G(n) \subset [dp]$ with $|G(n)| = k$ fixed as $n, p \to \infty$. Also, let the conditional variance $V_n \in \mathbb{R}^{dp\times dp}$ be defined by (24) for the* $\mathsf{VAR}(d)$ *model, that is:*

$$V_n \equiv \frac{\sigma^2}{n} \sum_{\ell=1}^{K-1} \sum_{t\in E_\ell} (M^{(\ell)} x_t)(M^{(\ell)} x_t)^{\mathsf{T}} \,. \tag{47}$$

Under the assumptions of Proposition 3.8, for all $u = (u_1, \dots, u_k) \in \mathbb{R}^k$ we have

$$\lim_{n\to\infty} \left| \mathbb{P}\left\{ \sqrt{n}(V_{n,G})^{-1/2}(\widehat{\theta}^{\mathsf{on}}_G - \theta_{0,G}) \le u \right\} - \Phi_k(u) \right| = 0 \,, \tag{48}$$

where $V_{n,G} \in \mathbb{R}^{k\times k}$ is the submatrix obtained by restricting V_n to the rows and columns in G. Here $(a_1, \dots, a_k) \le (b_1, \dots, b_k)$ indicates that $a_i \le b_i$ for $i \in [k]$ and $\Phi_k(u) = \Phi(u_1) \dots \Phi_k(u)$.

Much in the same way as individual inference, we can use Lemma 5.1 for simultaneous inference on a group of parameters. Concretely, let $\mathcal{S}_{k,\alpha} \subseteq \mathbb{R}^k$ be any Borel set with k-dimensional Gaussian measure at least $1 - \alpha$. Then for a group $G \subset [dp]$, with size $|G| = k$, we construct the confidence set $J_G(\alpha) \subseteq \mathbb{R}^k$ as follows

$$J_G(\alpha) \equiv \widehat{\theta}^{\mathsf{on}}_G + \frac{1}{\sqrt{n}} (V_{n,R})^{1/2} \mathcal{S}_{k,\alpha} \,. \tag{49}$$

Then, using Lemma 5.1 (along the same lines in deriving (44)), we conclude that $J_G(\alpha)$ is a valid confidence region, namely

$$\lim_{n\to\infty} \mathbb{P}(\theta_{0,G} \in J_G(\alpha)) = 1 - \alpha \,. \tag{50}$$

6 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 available at http://faculty.marshall.usc.edu/Adel-Javanmard/OnlineDebiasing.

Consider the $\mathsf{VAR}(d)$ time series model (10). 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

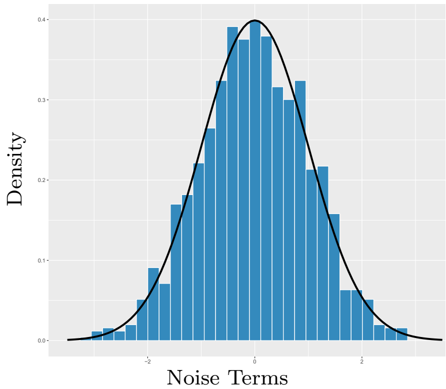
$$b \cdot \text{Bern}(q) \cdot \text{Unif}(\{+1, -1\}) + \mathcal{N}(0, 1/p^2)\,.$$

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

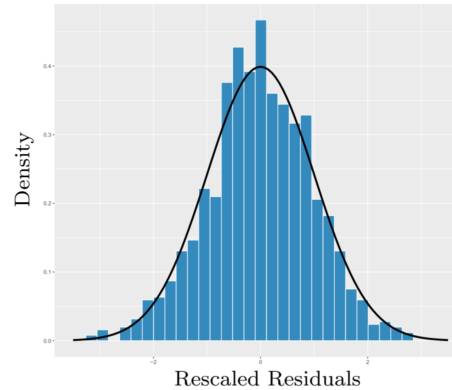
$$W_n \equiv \frac{1}{\sqrt{n}} \sum_{\ell=0}^{K-1} M^{(\ell)} \Big(\sum_{t \in E_\ell} x_t \varepsilon_t \Big)\,, \tag{51}$$

the rescaled residual $T_n \in \mathbb{R}^{dp}$ with $T_{n,a} = \sqrt{\frac{n}{V_{n,a}}}(\widehat{\theta}_a^{\mathsf{on}} - \theta_{0,a})$, and $V_{n,a}$ given by Equation (24) and $\sigma = 1$. Left and right plots of Figure 7 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.

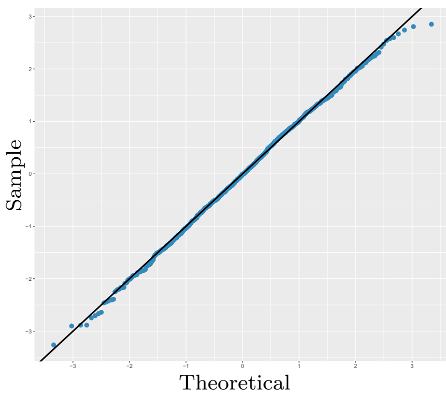
True and False Positive Rates. Consider the linear time-series model (10) with $A^{(i)}$ matrices having entries drawn independently from the distribution $b \cdot \text{Bern}(q) \cdot \text{Unif}(\{+1, -1\})$ and noise terms be gaussian with covariance matrix Σ_ζ. In this example, we evaluate the performance of our proposed online debiasing method for constructing confidence intervals and hypothesis testing as discussed in Section 5. We 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 1 and 2 summarize the results for various configurations of the $\text{Var}(d)$ processes and significance level $\alpha = 0.05$. Table 1 corresponds to the cases where noise covariance has the structure $\Sigma_\zeta(i,j) = 0.1^{|i-j|}$ and Table 2 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.



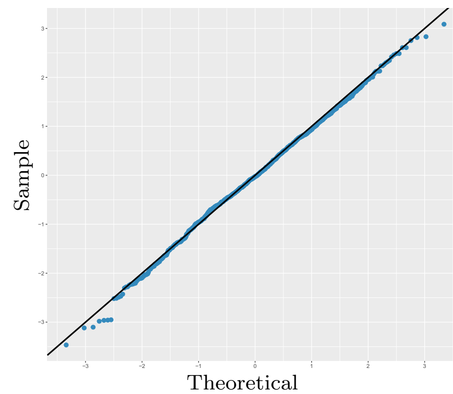
(a) Histogram of Noise Terms W_n



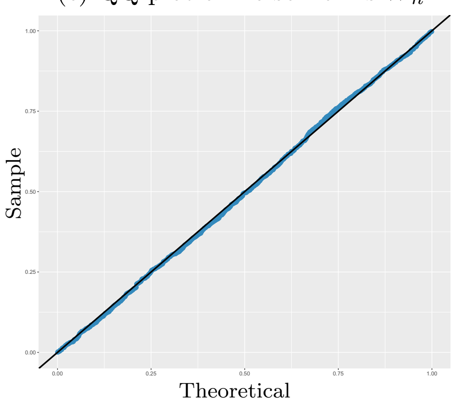
(b) Histogram of Residuals T_n



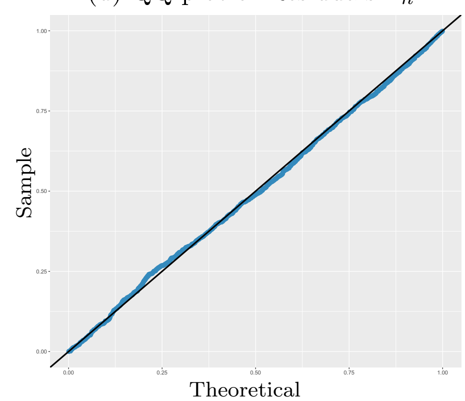
(c) QQ plot of Noise Terms W_n



(d) QQ plot of Residuals T_n



(e) PP plot of Noise Terms W_n



(f) PP plot of Residuals T_n

Figure 7: A simple example of an online debiased Var(3) process with dimension $p = 20$ and $T = 50$ sample data points. Plots 7a, 7c, 7e 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 (10). Plots 7b, 7d, 7f 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.

Table 1: 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 $\alpha = 0.05$

d \ Parameters	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 2: 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 $\alpha = 0.05$

d \ Parameters	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

6.1 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 [GWC16] where a sparse VAR model is fit to data and also in [WBBM17] 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[4] over 71 weeks, so for each week t, we have information

[4]Bottled Juices, Cereals, Cheeses, Cookies, Crackers, Canned Soup, Front-end-Candies, Frozen Juices, Soft Drinks,

$x_t \in \mathbb{R}^{33}$ for sales, prices and promotions of the 11 categories. For thorough explanation on calculating sales, prices and promotions, we refer to [SPHD04] and [GWC16]. 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. (45)). We refer to Appendix E for the reports of the p-values. By running the Benjamini–Yekutieli procedure [BY01] (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 [GWC16], 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.

7 Implementation and extensions

7.1 Iterative schemes to implement online debiasing

The online debiased estimator (16) involves the decorrelating matrices $M^{(\ell)}$, whose rows $(m_a^\ell)_{a\in[dp]}$ are constructed by the optimization (15). For the sake of computational efficiently, it is useful to work with a Lagrangian equivalent version of this optimization. Consider the following optimization

$$\text{minimize}_{\|m\|_1\le L} \quad \frac{1}{2} m^\mathsf{T}\widehat{\Sigma}^{(\ell)} m - \langle m, e_a\rangle + \mu_\ell \|m\|_1 \,, \tag{52}$$

with μ_ℓ and L taking the same values as in Optimization (15).

The next result, from [Jav14, Chapter 5] is on the connection between the solutions of the unconstrained problem (52) and (15). For the reader's convenience, the proof is also given in Appendix B.1.

Lemma 7.1. *A solution of optimization (52) is also a solution of the optimization problem (15). Also, if problem (15) is feasible then problem (52) has bounded solution.*

Using the above lemma, we can instead work with the Lagrangian version (52) for constructing the decorrelating vector m_a^ℓ.

Here, we propose to solve optimization problem (52) 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.

7.1.1 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 (52) by separating m_j from the other

Snack Crackers and Canned Tuna

coordinates:

$$\frac{1}{2}\widehat{\Sigma}^{(\ell)}_{j,j} m_j^2 + \sum_{r,s\neq j} \widehat{\Sigma}^{(\ell)}_{r,s}\, m_r m_s - m_a + \mu_\ell \|m_{\sim j}\|_1 + \mu_\ell |m_j|\,, \tag{53}$$

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 (53) with respect to m_j gives

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

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

$$m_j = \frac{1}{\widehat{\Sigma}^{(\ell)}_{j,j}} \eta\Big(-\widehat{\Sigma}^{(\ell)}_{j,\sim j} m_{\sim j} + \mathbb{I}(a=j); \mu_\ell\Big)\,, \tag{54}$$

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

$$\eta(z,\mu) = \begin{cases} z-\mu & \text{if } z > \mu\,,\\ 0 & \text{if } -\mu \le z \le \mu\,,\\ z+\mu & \text{otherwise}\,. \end{cases} \tag{55}$$

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 (52)). 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 [DSSSC08].

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.

7.1.2 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 (52) 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:

$$m(h+1) = \arg\min_{\|m\|_1\le L} \Big\{\mathcal{L}(m(h)) + \langle \nabla\mathcal{L}(m(h)), m - m(h)\rangle + \frac{\eta}{2}\|m - m(h)\|_2^2 + \mu_\ell \|m\|_1\Big\}\,. \tag{56}$$

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 (56) are guaranteed to converge to the global minimum of (52).

Plugging for $\mathcal{L}(m)$ and dropping the constant term $\mathcal{L}(m(h))$, update (56) reads as

$$\begin{aligned} m(h+1) &= \arg \min_{\|m\|_1 \le L} \left\{ \langle \widehat{\Sigma}^{(\ell)} m(h) - e_a, m - m(h) \rangle + \frac{\eta}{2} \|m - m(h)\|_2^2 + \mu_\ell \|m\|_1 \right\} \\ &= \arg \min_{\|m\|_1 \le L} \left\{ \frac{\eta}{2} \Big(m - m(h) + \frac{1}{\eta} (\widehat{\Sigma}^{(\ell)} m(h) - e_a) \Big)^2 + \mu_\ell \|m\|_1 \right\}. \end{aligned} \tag{57}$$

To compute the update (57), 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 (55). 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.

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

7.2 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 (5) we write

$$\begin{aligned} \widehat{\theta}^{\mathsf{off}} &= \theta_0 + (I - M\widehat{\Sigma})(\widehat{\theta}^{\mathsf{L}} - \theta_0) + \frac{1}{n} M X^\mathsf{T} \varepsilon \\ &= \theta_0 + (I - M\widehat{\Sigma})(\widehat{\theta}^{\mathsf{L}} - \theta_0) + \frac{1}{n}(M - \Omega) X^\mathsf{T} \varepsilon + \frac{1}{n} \Omega X^\mathsf{T} \varepsilon \,, \end{aligned} \tag{58}$$

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

$$\begin{aligned} \sqrt{n}(\widehat{\theta}^{\mathsf{off}} - \theta_0) &= \Delta_1 + \Delta_2 + \frac{1}{\sqrt{n}} \Omega X^\mathsf{T} \varepsilon \,, \\ \Delta_1 &= \sqrt{n}(I - M\widehat{\Sigma})(\widehat{\theta}^{\mathsf{L}} - \theta_0) \,, \\ \Delta_2 &= \frac{1}{\sqrt{n}}(M - \Omega) X^\mathsf{T} \varepsilon \,. \end{aligned} \tag{59}$$

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

$$\|\Delta_2\|_\infty \le \frac{1}{\sqrt{n}}\|M-\Omega\|_1\|X^{\mathsf{T}}\varepsilon\|_\infty\,, \tag{60}$$

where $\|M-\Omega\|_1$ denotes the $\ell_1 - \ell_1$ norm of $M-\Omega$ (the maximum ℓ_1 norm of its columns). By using [BM15, 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

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

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

$$m_a \in \arg\min_{m\in\mathbb{R}^{dp}} \quad \frac{1}{2} m^{\mathsf{T}}\widehat{\Sigma}m - \langle m, e_a\rangle + \mu\|m\|_1\,. \tag{61}$$

Lemma 7.2. *Consider the decorrelating vectors m_a, $a \in [dp]$, given by optimization (61) 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 α and ω are defined in Proposition A.4.

The proof of Lemma 7.2 is deferred to Section B.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 7.3. *Consider the $\mathsf{VAR}(d)$ model (10) for time series and let $\widehat{\theta}^{\mathsf{off}}$ be the (offline) debiased estimator (4), with the decorrelating matrix $M = (m_1,\dots,m_{dp})^{\mathsf{T}} \in \mathbb{R}^{dp\times dp}$ constructed as in (61), 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

$$\lim_{n\to\infty} \sup_{\|\theta_0\|_0 \le s_0} \left| \mathbb{P}\left\{ \frac{\sqrt{n}(\widehat{\theta}^{\mathsf{off}}_a - \theta_{0,a})}{\sqrt{V_{n,a}}} \le x \right\} - \Phi(x) \right| = 0\,, \tag{62}$$

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

We refer to Section B.3 for the proof of Theorem 7.3.

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 8a 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 8b, 8c, and 8d demonstrate normality of the rescaled residuals of the offline debiased estimator built by decorrelating matrix M with rows coming from optimization described in (61).

After this paper was posted, we learned of simultaneous work (an updated version of [BDMP17]) 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 7.2 and that paper: 1) [BDMP17] considers decorrelating matrix M constructed by an optimization of form (15), using the entire sample covariance $\widehat{\Sigma}^{(K)}$, while we work with the Lagrangian equivalent (61). 2) [BDMP17] considers $\mathsf{VAR}(1)$ model, while we work with $\mathsf{VAR}(d)$ models. 3) [BDMP17] 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 7.3.

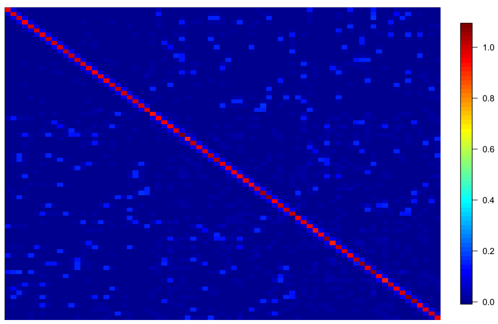
7.3 Concluding remarks

In this work we devised the 'online debiasing' approach for the high-dimensional regression and showed that it asymptotically admits an unbiased Gaussian distribution, even when the samples are collected adaptively. Also through numerical examples we demonstrated that the (offline) debiased estimator suffers from the bias induced by the correlation in the samples and cannot be used for valid statistical inference in these settings (unless the precision matrix is sufficiently sparse).

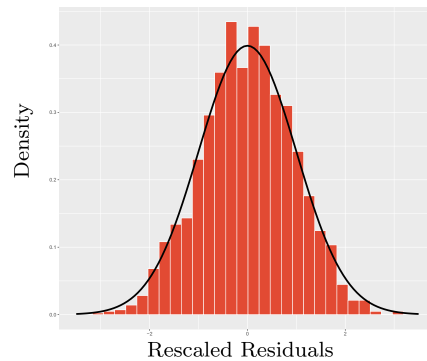
Since its proposal, the (offline) debiasing approach has been used as a tool to address a variety of problems such as estimating average treatment effect and casual inference in high-dimension [AIW16], precision matrix estimation [JvdG17], distributed multitask learning, and studying neuronal functional network dynamics [SML+18], hierarchical testing [GRBC19], to name a few. It has also been used for different statistical aims such as controlling FDR in high-dimensions [JJ+19], estimation of the prediction risk [JM18], inference on predictions [CG17, JL17] and explained variance [CG18, JL17], and testing more general hypotheses regarding the model parameters, like testing membership in a convex cone, testing the parameter strength, and testing arbitrary functions of the parameters [JL17]. We anticipate that the online debiasing approach and analysis can be used to tackle similar problems under adaptive data collection. We leave this for future work.

Acknowledgements

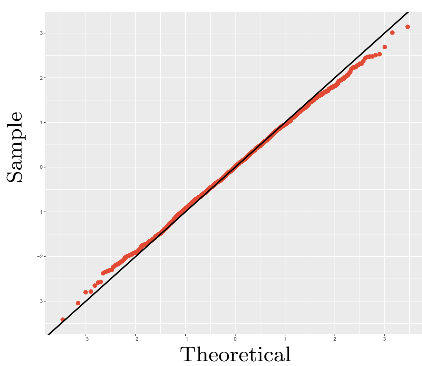
A. Javanmard was partially supported by an Outlier Research in Business (iORB) grant from the USC Marshall School of Business, a Google Faculty Research Award and the NSF CAREER Award DMS-1844481.



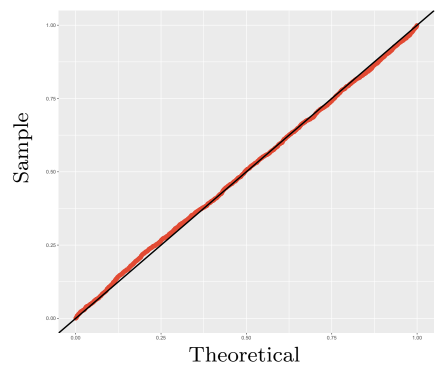
(a) Heat map of magnitudes of entries of $\Omega = \mathbb{E}(x_i x_i^T)^{-1}$



(b) Histogram of Rescaled Residuals



(c) QQ plot of Rescaled Residuals



(d) PP plot of Rescaled Residuals

Figure 8: 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 8b, 8c, and 8d) validates the successful performance of offline debiasing estimator under sparsity of precision matrix Ω (figure 8a) as we discussed in theorem 7.3.

References

[AIW16] Susan Athey, Guido W Imbens, and Stefan Wager, *Approximate residual balancing: De-biased inference of average treatment effects in high dimensions*, arXiv preprint arXiv:1604.07125 (2016). 31

[BB15] Hamsa Bastani and Mohsen Bayati, *Online decision-making with high-dimensional covariates*, Available at SSRN 2661896 (2015). 3, 4

[BCB+12] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al., *Regret analysis of stochastic and non-stochastic multi-armed bandit problems*, Foundations and Trends® in Machine Learning **5** (2012), no. 1, 1–122. 4

[BCW11] Alexandre Belloni, Victor Chernozhukov, and Lie Wang, *Square-root lasso: pivotal recovery of sparse signals via conic programming*, Biometrika **98** (2011), no. 4, 791–806. 16

[BDMP17] Sumanta Basu, Sreyoshi Das, George Michailidis, and Amiyatosh K Purnanandam, *A system-wide approach to measure connectivity in the financial sector*, Available at SSRN 2816137 (2017). 31

[BM12] M. Bayati and A. Montanari, *The LASSO risk for gaussian matrices*, IEEE Trans. on Inform. Theory **58** (2012), 1997–2017. 2

[BM15] Sumanta Basu and George Michailidis, *Regularized estimation in sparse high-dimensional time series models*, The Annals of Statistics **43** (2015), no. 4, 1535–1567. 3, 7, 9, 10, 11, 30, 37, 38, 39, 48

[BVDG11] Peter Bühlmann and Sara Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media, 2011. 7, 9, 15, 39, 51

[BY01] Yoav Benjamini and Daniel Yekutieli, *The control of the false discovery rate in multiple testing under dependency*, Annals of statistics (2001), 1165–1188. 27

[CG17] T Tony Cai and Zijian Guo, *Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity*, The Annals of statistics **45** (2017), no. 2, 615–646. 3, 31

[CG18] ______, *Semi-supervised inference for explained variance in high-dimensional linear regression and its applications*, arXiv preprint arXiv:1806.06179 (2018). 31

[DM12] Yash Deshpande and Andrea Montanari, *Linear bandits in high dimension and recommendation systems*, Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on, IEEE, 2012, pp. 1750–1754. 4

[DMST18] Yash Deshpande, Lester Mackey, Vasilis Syrgkanis, and Matt Taddy, *Accurate inference for adaptive linear models*, International Conference on Machine Learning, 2018, pp. 1202–1211. 8, 20, 22

[DSSSC08] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra, *Efficient projections onto the l 1-ball for learning in high dimensions*, Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 272–279. 28

[FSGM+07] André Fujita, Joao R Sato, Humberto M Garay-Malpartida, Rui Yamaguchi, Satoru Miyano, Mari C Sogayar, and Carlos E Ferreira, *Modeling gene expression regulatory networks with the sparse vector autoregressive model*, BMC systems biology **1** (2007), no. 1, 39. 8

[GRBC19] Zijian Guo, Claude Renaux, Peter Bühlmann, and T Tony Cai, *Group inference in high dimensions with applications to hierarchical testing*, arXiv preprint arXiv:1909.01503 (2019). 31

[GWC16] Sarah Gelper, Ines Wilms, and Christophe Croux, *Identifying demand effects in a large network of product categories*, Journal of Retailing **92** (2016), no. 1, 25–39. 26, 27

[HENR88] Douglas Holtz-Eakin, Whitney Newey, and Harvey S Rosen, *Estimating vector autoregressions with panel data*, Econometrica: Journal of the Econometric Society (1988), 1371–1395. 8

[HH14] Peter Hall and Christopher C Heyde, *Martingale limit theory and its application*, Academic press, 2014. 14, 15, 48, 57

[HTW15] Trevor Hastie, Robert Tibshirani, and Martin Wainwright, *Statistical learning with sparsity: the lasso and generalizations*, Chapman and Hall/CRC, 2015. 51

[Jav14] Adel Javanmard, *Inference and estimation in high-dimensional data analysis*, Ph.D. thesis, PhD Thesis, Stanford University, 2014. 3, 21, 27

[JJ+19] Adel Javanmard, Hamid Javadi, et al., *False discovery rate control via debiased lasso*, Electronic Journal of Statistics **13** (2019), no. 1, 1212–1253. 31

[JL17] Adel Javanmard and Jason D Lee, *A flexible framework for hypothesis testing in high-dimensions*, arXiv preprint arXiv:1704.07971 (2017). 31

[JM14a] Adel Javanmard and Andrea Montanari, *Confidence intervals and hypothesis testing for high-dimensional regression.*, Journal of Machine Learning Research **15** (2014), no. 1, 2869–2909. 3, 7, 10, 11, 16, 17

[JM14b] ———, *Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory*, IEEE Transactions on Information Theory **60** (2014), no. 10, 6522–6554. 2, 3, 21

[JM18] ———, *Debiasing the lasso: Optimal sample size for gaussian designs*, The Annals of Statistics **46** (2018), no. 6A, 2593–2622. 31

[JvdG17] Jana Janková and Sara van de Geer, *Honest confidence regions and optimality in high-dimensional precision matrix estimation*, Test **26** (2017), no. 1, 143–162. 31

[KHW⁺11] Edward S Kim, Roy S Herbst, Ignacio I Wistuba, J Jack Lee, George R Blumenschein, Anne Tsao, David J Stewart, Marshall E Hicks, Jeremy Erasmus, Sanjay Gupta, et al., *The battle trial: personalizing therapy for lung cancer*, Cancer discovery **1** (2011), no. 1, 44–53. 4

[LR85] Tze Leung Lai and Herbert Robbins, *Asymptotically efficient adaptive allocation rules*, Advances in applied mathematics **6** (1985), no. 1, 4–22. 4

[LW82] Tze Leung Lai and Ching Zong Wei, *Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems*, The Annals of Statistics (1982), 154–166. 2, 7, 9

[NXTZ17] Xinkun Nie, Tian Xiaoying, Jonathan Taylor, and James Zou, *Why adaptively collected data have negative bias and how to correct for it.* 2

[PRC⁺16] Vianney Perchet, Philippe Rigollet, Sylvain Chassang, Erik Snowberg, et al., *Batched bandit problems*, The Annals of Statistics **44** (2016), no. 2, 660–681. 4

[RT10] Paat Rusmevichientong and John N Tsitsiklis, *Linearly parameterized bandits*, Mathematics of Operations Research **35** (2010), no. 2, 395–411. 4

[SBB15] Anil K Seth, Adam B Barrett, and Lionel Barnett, *Granger causality analysis in neuroscience and neuroimaging*, Journal of Neuroscience **35** (2015), no. 8, 3293–3297. 8

[SML⁺18] Alireza Sheikhattar, Sina Miran, Ji Liu, Jonathan B Fritz, Shihab A Shamma, Patrick O Kanold, and Behtash Babadi, *Extracting neuronal functional network dynamics via adaptive granger causality analysis*, Proceedings of the National Academy of Sciences **115** (2018), no. 17, E3869–E3878. 31

[SPHD04] Shuba Srinivasan, Koen Pauwels, Dominique M Hanssens, and Marnik G Dekimpe, *Do promotions benefit manufacturers, retailers, or both?*, Management Science **50** (2004), no. 5, 617–629. 27

[SRR19] Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo, *On the bias, risk and consistency of sample means in multi-armed bandits*, arXiv preprint arXiv:1902.00746 (2019). 2

[SS06] Robert H Shumway and David S Stoffer, *Time series analysis and its applications: with r examples*, Springer Science & Business Media, 2006. 8, 9

[SW01] James H Stock and Mark W Watson, *Vector autoregressions*, Journal of Economic perspectives **15** (2001), no. 4, 101–115. 8

[SZ12] Tingni Sun and Cun-Hui Zhang, *Scaled sparse linear regression*, Biometrika **99** (2012), no. 4, 879–898. 16

[Tib96] R. Tibshirani, *Regression shrinkage and selection with the Lasso*, J. Royal. Statist. Soc B **58** (1996), 267–288. 2

[VBW15] Sofia Villar, Jack Bowden, and James 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** (2015), no. 2, 199. 2

[VdGBR+14] Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, Ruben Dezeure, et al., *On asymptotically optimal confidence regions and tests for high-dimensional models*, The Annals of Statistics **42** (2014), no. 3, 1166–1202. 3

[Ver12] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, Compressed Sensing: Theory and Applications (Y.C. Eldar and G. Kutyniok, eds.), Cambridge University Press, 2012, pp. 210–268. 61, 62

[WBBM17] Ines Wilms, Sumanta Basu, Jacob Bien, and David S Matteson, *Interpretable vector autoregressions with exogenous time series*, arXiv preprint arXiv:1711.03623 (2017). 26

[XQL13] Min Xu, Tao Qin, and Tie-Yan Liu, *Estimation bias in multi-armed bandit algorithms for search advertising*, Advances in Neural Information Processing Systems, 2013, pp. 2400–2408. 2

[ZLK+08] Xian Zhou, Suyu Liu, Edward S Kim, Roy S Herbst, and J Jack Lee, *Bayesian adaptive design for targeted therapy development in lung cancer—a step toward personalized medicine*, Clinical Trials **5** (2008), no. 3, 181–193. 4

[ZZ11] C.-H. Zhang and S.S. Zhang, *Confidence Intervals for Low-Dimensional Parameters in High-Dimensional Linear Models*, arXiv:1110.2563, 2011. 3

A Proofs of Section 3

A.1 Technical preliminaries

Recall the definition of the regression design from Eqs.(12) in the time series case:

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

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

$$M(f_x) \equiv \sup_{r\in[-\pi,\pi]} \sigma_{\max}(f_x(r))\,. \tag{63}$$

Likewise, the minimum eigenvalue of the spectrum is defined as $m(f_x) \equiv \inf_{r\in[-\pi,\pi]} \sigma_{\min}(f_x(r))$, 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 (10) 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 (12)). Then (10) can be written as

$$x_t = \widetilde{A} x_{t-1} + \tilde{\zeta}_t\,, \tag{64}$$

with

$$\widetilde{A} = \left(\begin{array}{cccc|c} A_1 & A_2 & \dots & A_{d-1} & A_d \\ \hline & & I_{(d-1)p} & & 0 \end{array}\right), \qquad \tilde{\zeta}_t = \begin{pmatrix} \zeta_{t+d-1} \\ 0 \end{pmatrix}. \tag{65}$$

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 A.1 ([BM15]). *We have:*

$$
\begin{aligned}
\frac{1}{2\pi}\lambda_{\max}(\Sigma) \le M(f_x) &\le \frac{\lambda_{\max}(\Sigma_\zeta)}{\mu_{\min}(\tilde{\mathcal{A}})}, \\
\lambda_{\min}(\Sigma) &\ge \frac{\lambda_{\min}(\Sigma_\zeta)}{\mu_{\max}(\mathcal{A})}.
\end{aligned} \tag{66}
$$

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

Lemma A.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

$$v^\mathsf{T}\Gamma_z(\ell)v = \sum_{1\le r,s\le d} v_r^\mathsf{T}\Gamma_x(\ell + r - s)v_s\,. \tag{67}$$

Since the spectral density $f_z(\theta)$ is the Fourier transform of the autocorrelation function, we have by Equation (67),

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

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 $M(f_x) \le dM(f_z)$. Now, by using [BM15, Equation (4.1)] for the process z_t, with reverse characteristic polynomial $\mathcal{A}$, we obtain

$$\lambda_{\max}(\Sigma) \le 2\pi M(f_x) \le 2\pi dM(f_z) \le \frac{d\lambda_{\max}(\Sigma_\zeta)}{\mu_{\min}(\mathcal{A})}\,. \tag{68}$$

□

The following proposition is a straightforward consequence of the spectral bounds above and [BM15, Proposition 2.4].

Proposition A.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 $\eta \ge 0$,*

$$\mathbb{P}\left(|u^\mathsf{T}(\widehat{\Sigma}^{(\ell)} - \Sigma)v| > \frac{d\lambda_{\max}(\Sigma_\zeta)}{\mu_{\min}(\mathcal{A})}\eta\right) \le 6\exp\left(-cn_\ell \min\{\eta^2, \eta\}\right). \tag{69}$$

A.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 A.4. *Let $\{z_1, \dots, z_T\}$ be generated according to the (stable) $\mathsf{VAR}(d)$ process (10) and let $n = T - d$. Then there exist constants $c \in (0, 1)$ and $C > 1$ such that for all $n \ge C\omega^2 \log(dp)$, with probability at least $1 - \exp(-cn/\omega^2)$, satisfies*

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

Here, α, ω and τ are given by:

$$\begin{aligned}
\omega &= \frac{d\lambda_{\max}(\Sigma_\zeta)\mu_{\max}(\mathcal{A})}{\lambda_{\min}(\Sigma_\zeta)\mu_{\min}(\mathcal{A})}\,,\\
\alpha &= \frac{\lambda_{\min}(\Sigma_\zeta)}{2\mu_{\max}(\mathcal{A})}\,,\\
\tau &= \omega^2\sqrt{\frac{\log(dp)}{n}}\,.
\end{aligned} \tag{70}$$

Given Proposition A.4, the estimation result of Theorem 3.2 is standard (see [BVDG11]). Proposition A.4 can be proved analogous to [BM15, Proposition 4.2], with the following considerations and minor modifications:

1. [BM15] 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 [BM15]). 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 A.4.

2. [BM15, Proposition 4.2] assumes $n \ge Ck\max\{\omega^2, 1\}\log(dp)$, with $k = \sum_{\ell=1}^d \|\mathrm{vec}(A^{(\ell)})\|_0$, the total number of nonzero entries of matrices A_ℓ and then it is later used to get $\tau \le 1/(Ck)$. 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 A.2 in lieu of Lemma A.1.

A.3 Proof of Lemma 3.3

The idea is to use Proposition A.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 A.1 to bound $\lambda_{\min}(\Sigma)$ from below. Combining this with Proposition A.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.

A.4 Proof of Theorem 3.4

Starting from the decomposition (21), 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 (21), 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 A.5. *Suppose that the Optimization problem* (15) *is feasible for all* $i \in [dp]$. *Let* ω *and* γ *be:*

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

Then, with probability at least $1 - (dp)^{-8}$

$$\|B_n\|_\infty \le \frac{r_0}{\sqrt{n}} + C(\omega + L\gamma)\sqrt{\frac{\log(dp)}{n}} \sum_{\ell=1}^{K-1} \Big(\frac{r_\ell}{\sqrt{n_\ell}} + \sqrt{r_\ell}\Big). \tag{71}$$

The bound provided in Lemma A.5 holds for general batch sizes $r_0, \dots, r_{K-1}$. We choose the batch lengths as $r_\ell = \beta^\ell$ for some $\beta > 1$ and $\ell = 1, \dots, K-1$. We also let $r_0 = \sqrt{n}$ and choose r_{K-1} so that the total lengths of batches add up to n (that is $r_0 + r_1 + \dots + r_{K-1} = n$). Therefore, $K = O(\log_\beta(n))$. Following this choice, bound (71) simplifies to:

$$\|B_n\|_\infty \le C_\beta(\omega + \gamma L)\sqrt{\log(dp)}\,, \tag{72}$$

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

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

$$\begin{aligned}\|\Delta_n\|_\infty &\le C_\beta(\omega + L\gamma)\sqrt{\log(dp)} \cdot \Big(\frac{s_0\lambda_n}{\alpha}\Big)\\ &\le C_\beta \frac{\lambda_0(\omega + L\gamma)}{\alpha}\frac{s_0\log(dp)}{\sqrt{n}}.\end{aligned} \tag{73}$$

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.(73). 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.(73). For the second, we use Lemma D.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 A.5:

Proof of Lemma A.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

$$B_{n,a} = \sqrt{n} e_a - \frac{1}{\sqrt{n}} \sum_{\ell=1}^{K-1} r_\ell R^{(\ell)} m_a^\ell = \frac{r_0}{\sqrt{n}} e_a + \sum_{\ell=1}^{K-1} \frac{r_\ell}{\sqrt{n}} \Big(e_a - R^{(\ell)} m_a^\ell \Big) , \tag{74}$$

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 (15), we have

$$\|\widehat{\Sigma}^{(\ell)} m_a^\ell - e_a\|_\infty \le \mu_\ell \,, \quad \ell = 0, \dots, K-1 \,. \tag{75}$$

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 third 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 $1 - 2K(dp)^{-9}$,

$$\begin{aligned}
\|B_{n,a}\|_\infty &\le \frac{r_0}{\sqrt{n}} + \frac{C}{\sqrt{n}} \sum_{\ell=0}^{K-2} r_\ell \left(\omega \sqrt{\frac{\log(dp)}{n_\ell}} + L\gamma \sqrt{\frac{\log(dp)}{n_\ell}} + L\gamma \sqrt{\frac{\log(dp)}{r_\ell}} \right) \\
&\le \frac{r_0}{\sqrt{n}} + C(\omega + L\gamma) \sqrt{\frac{\log(dp)}{n}} \sum_{\ell=0}^{K-2} \Big(\frac{r_\ell}{\sqrt{n_\ell}} + \sqrt{r_\ell} \Big).
\end{aligned}$$

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. □

A.5 Proof of Lemma 3.6

We start by proving Claim (24). 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 $V_{n,a}$. Using the shorthand $\sigma^2 = \Sigma_{\zeta_{i,i}}$ Therefore, we decompose the difference as follows:

$$
\begin{aligned}
V_{n,a} - \Omega_{a,a} &= \frac{\sigma^2}{n} \sum_{\ell=1}^{K-1} \sum_{t\in E_\ell} \Big[\langle m_a^\ell, x_t\rangle^2 - \Omega_{a,a}\Big] - \frac{r_0\sigma^2}{n}\Omega_{a,a} \\
&= \frac{\sigma^2}{n} \sum_{\ell=1}^{K-1} \sum_{t\in E_\ell} \Big[\langle m_a^\ell, x_t\rangle^2 - \langle m_a, \mathbb{E}\{x_t x_t^{\mathsf{T}}\} m_a\rangle\Big] - \frac{r_0\sigma^2}{n}\Omega_{a,a} \\
&= \frac{\sigma^2}{n} \sum_{\ell=1}^{K-1} \sum_{t\in E_\ell} [\langle m_a^\ell, x_t\rangle^2 - \langle m_a, x_t\rangle^2] \\
&\quad + \frac{1}{n}\sum_{t=0}^{n-1} \langle m_a, (x_t x_t^{\mathsf{T}} - \mathbb{E}\{x_t x_t^{\mathsf{T}}\}) m_a\rangle - \frac{r_0\sigma^2}{n}\Omega_{a,a}\,.
\end{aligned} \tag{76}
$$

We treat each of these three terms separately. Write

$$
\begin{aligned}
\left|\frac{1}{n} \sum_{\ell=1}^{K-1} \sum_{t\in E_\ell} [\langle m_a^\ell, x_t\rangle^2 - \langle m_a, x_t\rangle^2]\right| &= \frac{1}{n}\left| \sum_{\ell=1}^{K-1} \sum_{t\in E_\ell} [\langle m_a^\ell - m_a, x_t\rangle \langle m_a^\ell + m_a, x_t\rangle]\right| \\
&\le \frac{1}{n}\left\| \sum_{\ell=1}^{K-1} \sum_{t\in E_\ell} \langle m_a^\ell - m_a, x_t\rangle x_t\right\|_\infty \|m_a^\ell + m_a\|_1 \\
&\le \frac{2L}{n}\left\| \sum_{\ell=1}^{K-1} \sum_{t\in E_\ell} \langle m_a^\ell - m_a, x_t\rangle x_t\right\|_\infty .
\end{aligned} \tag{77}
$$

To bound the last quantity, note that

$$
\begin{aligned}
\frac{1}{n}\left\| \sum_{\ell=1}^{K-1} \sum_{t\in E_\ell} \langle m_a^\ell - m_a, x_t\rangle x_t\right\|_\infty &\le \left\| e_a - \frac{1}{n}\sum_{\ell=1}^{K-1} \sum_{t\in E_\ell} \langle m_a^\ell, x_t\rangle x_t\right\|_\infty \\
&\quad + \left\| e_a - \frac{1}{n}\sum_{\ell=1}^{K-1} \sum_{t\in E_\ell} \langle m_a, x_t\rangle x_t\right\|_\infty \\
&= \left\| e_a - \frac{1}{n}\sum_{\ell=1}^{K-1} r_\ell R^{(\ell)} m_a^\ell\right\|_\infty + \left\| e_a - \widehat{\Sigma}^{(K)} m_a\right\|_\infty \\
&= \frac{1}{\sqrt{n}}\|B_{n,a}\|_\infty + \left\| e_a - \widehat{\Sigma}^{(K)} m_a\right\|_\infty \\
&\le CL\gamma\sqrt{\frac{\log(dp)}{n}} + C\omega\sqrt{\frac{\log(dp)}{n}} \le C(L\gamma+\omega)\sqrt{\frac{\log(dp)}{n}}\,,
\end{aligned} \tag{78}
$$

for some constant C. The last inequality follows from the positive events of Lemma A.5 and Lemma 3.3. Combining Equations (77) and (78), we obtain

$$\left|\frac{1}{n}\sum_{\ell=1}^{K-1}\sum_{t\in E_\ell}[\langle m_a^\ell, x_t\rangle^2 - \langle m_a, x_t\rangle^2]\right| \le CL(\omega + L\gamma)\sqrt{\frac{\log(dp)}{n}}. \tag{79}$$

For the second term in (76), we can use Proposition A.3 with $v = u = m_a/\|m_a\|, \eta = C\sqrt{\log(dp)/n}$ to obtain

$$\begin{aligned}\left|\frac{1}{n}\sum_{t=0}^{n-1}\langle m_a, (x_t x_t^\mathsf{T} - \mathbb{E}\{x_t x_t^\mathsf{T}\})m_a\rangle\right| &= \left|\langle m_a, (\widehat{\Sigma}^{(K-1)} - \Sigma)m_a\rangle\right| \\ &\le \frac{Cd\lambda_{\max}(\Sigma_\zeta)}{\mu_{\min}(\mathcal{A})}\|m_a\|^2\sqrt{\frac{\log(dp)}{n}} \\ &\le \frac{Cd\lambda_{\max}(\Sigma_\zeta)}{\mu_{\min}(\mathcal{A})\lambda_{\min}(\Sigma)^2}\sqrt{\frac{\log(dp)}{n}} \qquad (80)\\ &\le \frac{C\omega}{\alpha}\sqrt{\frac{\log(dp)}{n}}, \qquad (81)\end{aligned}$$

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 (79) and (81) in Equation (76) we get the Claim (24).

We next prove Claim (25). 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

$$\mathbb{P}\left(\|X\|_\infty < \sqrt{2\lambda_{\max}(\Sigma)\log(dpn)}\right) \ge 1 - (pdn)^{-2}, \tag{82}$$

Putting these bounds together we get

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

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

A.6 Proof of Proposition 3.8

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

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

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

Invoking the decomposition (22) we have

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

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

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

By taking the limit and using Equation (22), we get

$$\lim_{n\to\infty} \sup_{\|\theta_0\|_0 \le s_0} \mathbb{P}\Big\{ \frac{\sqrt{n}(\widehat{\theta}^{\sf on}_a - \theta_{0,a})}{\sqrt{V_{n,a}}} \le x \Big\} \le \Phi(x+\varepsilon) + \lim_{n\to\infty} \sup_{\|\theta_0\|_0 \le s_0} \mathbb{P}\Big\{ \frac{|\Delta_a|}{\sqrt{V_{n,a}}} \ge \varepsilon \Big\} \tag{84}$$

We show that the limit on the right hand side vanishes for any $\varepsilon > 0$. By virtue of Lemma 3.6 (Equation (24)), we have

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}\Big\{ \frac{|\Delta_a|}{\sqrt{V_{n,a}}} \ge \varepsilon \Big\} &\le \lim_{n\to\infty} \mathbb{P}\Big\{ \frac{|\Delta_a|}{\sigma\sqrt{\Omega_{a,a}}} \ge \varepsilon \Big\} \\ &\le \lim_{n\to\infty} \mathbb{P}\Big\{ |\Delta_a| \ge \varepsilon\sigma\sqrt{\Omega_{a,a}} \Big\} \\ &\le \lim_{n\to\infty} (dp)^{-4} = 0\,. \end{aligned} \tag{85}$$

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 (23) of Theorem 3.4.

Using (85) in bound (84) and then taking the limit $\varepsilon \to 0$, we obtain (83).

B Proofs of Section 7

B.1 Proof of Lemma 7.1

Rewrite the optimization problem (15) as follows:

$$\begin{aligned} &\text{minimize} \quad m^{\sf T}\widehat{\Sigma}^{(\ell)} m \\ &\text{subject to} \quad \langle z, \widehat{\Sigma}^{(\ell)} m - e_a\rangle \le \mu_\ell, \quad \|m\|_1 \le L, \quad \|z\|_1 = 1\,, \end{aligned} \tag{86}$$

The Lagrangian is given by

$$\mathcal{L}(m,z,\lambda) = m^{\mathsf{T}}\widehat{\Sigma}^{(\ell)}m + \lambda(\langle z, \widehat{\Sigma}^{(\ell)}m - e_a\rangle - \mu_\ell), \quad \|z\|_1 = 1, \quad \|m\|_1 \le L\,, \tag{87}$$

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

$$\begin{aligned} \text{maximize} \quad & -\frac{\lambda^2}{4} z^{\mathsf{T}}\widehat{\Sigma}^{(\ell)}z - \lambda\langle z, e_a\rangle - \lambda\mu_\ell \\ \text{subject to} \quad & \frac{\lambda}{2} \le L, \quad \|z\|_1 = 1\,, \end{aligned} \tag{88}$$

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

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\beta^{\mathsf{T}}\widehat{\Sigma}^{(\ell)}\beta - \langle \beta, e_a\rangle + \mu_\ell\|\beta\|_1 \\ \text{subject to} \quad & \|\beta\|_1 \le L\,, \end{aligned} \tag{89}$$

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 (89) is the dual to problem (86), we have that if (86) is feasible then the problem (89) is bounded.

B.2 Proof of Lemma 7.2

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

$$\alpha = \frac{\lambda_{\min}(\Sigma_\zeta)}{2\mu_{\max}(\mathcal{A})}\,, \qquad \tau = C\omega^2\sqrt{\frac{\log(dp)}{n}}\,, \tag{90}$$

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 (61). 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

$$\frac{1}{2}\nu^{\mathsf{T}}\widehat{\Sigma}\nu \le \langle \nu, e_a - \widehat{\Sigma}m_a^*\rangle + \mu_n(\|m_a^*\|_1 - \|m_a^* + \nu\|_1)\,. \tag{91}$$

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 $S = \mathrm{supp}(\Omega e_a)$ 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 (91) is upper bounded by

$$(\|\nu_S\|_1 + \|\nu_{S^c}\|_1)\frac{\mu_n}{2} + \mu_n\left(\|\nu_S\|_1 - \|\nu_{S^c}\|_1\right) = \frac{3}{2}\mu_n\|\nu_S\|_1 - \frac{1}{2}\mu_n\|\nu_{S^c}\|_1\,, \tag{92}$$

Given that $\widehat{\Sigma} \succeq 0$, the left hand side of (91) is non-negative, which implies that $\|\nu_{S^c}\|_1 \le 3\|\nu_S\|_1$ and hence

$$\|\nu\|_1 \le 4\|\nu_S\|_1 \le 4\sqrt{s_\Omega}\|\nu_S\|_2 \le 4\sqrt{s_\Omega}\|\nu\|_2\,. \tag{93}$$

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

$$\nu^{\mathsf{T}}\widehat{\Sigma}\nu \ge \alpha\|\nu\|_2^2 - \alpha\tau\|\nu\|_1^2 \ge \alpha(1-16s_\Omega\tau)\|\nu\|_2^2 \ge \frac{\alpha}{2}\|\nu\|_2^2\,, \tag{94}$$

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

Putting (91), (92) and (94) 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 93, we get

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

which completes the proof.

B.3 Proof of Theorem 7.3

Continuing from the decomposition (59) we have

$$\sqrt{n}(\widehat{\theta}^{\mathsf{off}} - \theta_0) = \Delta_1 + \Delta_2 + Z\,, \tag{95}$$

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 (61) (which is shown to be its dual in Lemma (7.1)) has bounded solution. Hence, its solution should satisfy the KKT condition which reads as

$$\widehat{\Sigma}m_a - e_a + \mu\,\text{sign}(m_a) = 0\,, \tag{96}$$

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

$$\|\Delta_1\|_\infty \le C\sqrt{n}\mu s_0\sqrt{\frac{\log p}{n}} = O_P\Big(s_0\frac{\log(dp)}{\sqrt{n}}\Big)\,. \tag{97}$$

We next bound the bias term Δ_2. By virtue of [BM15, 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 7.2 gives us the following bound

$$\|\Delta_2\|_\infty \le \left(\max_{i\in[dp]} \|(M-\Omega)e_i)\|\right)\left(\frac{1}{\sqrt{n}}\|X^{\mathsf{T}}\varepsilon\|_\infty\right) = O_P\Big(s_\Omega \frac{\log(dp)}{\sqrt{n}}\Big)\,. \tag{98}$$

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 B.1.

Lemma B.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 (61), the following holds. For any fixed sequence of integers $a(n) \in [dp]$, we have*

$$m_a^{\mathsf{T}}\widehat{\Sigma}m_a = \Omega_{a,a} + o_P(1/\sqrt{\log(dp)})\,. \tag{99}$$

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

$$\lim_{n\to\infty}\sup_{\|\theta_0\|_0\le s_0} \mathbb{P}\left\{\frac{\sqrt{n}(\widehat{\theta}^{\mathsf{off}}_a - \theta_{0,a})}{\sqrt{V_{n,a}}} \le u\right\} \le \Phi(u)\,. \tag{100}$$

A similar lower bound can be proved analogously. By the decomposition (95) 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 [HH14, Corollary 3.2], we have that $\widetilde{Z}_a \to \mathsf{N}(0,1)$ in distribution. In other words,

$$\lim_{n\to\infty}\mathbb{P}\{\widetilde{Z}_a u\} = \Phi(u)\,. \tag{101}$$

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

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

Now by taking the limit of both sides and using (101) and Lemma B.1, we obtain

$$\limsup_{n\to\infty} \sup_{\|\theta_0\|_0\le s_0} \mathbb{P}\left\{\frac{\sqrt{n}(\widehat{\theta}^{\sf off}_a - \theta_{0,a})}{\sqrt{V_{n,a}}} \le u\right\} \le \\ \Phi(u+2\delta+\delta|u|) + \limsup_{n\to\infty} \sup_{\|\theta_0\|_0\le s_0} \mathbb{P}\left\{\frac{\Delta_a}{\sqrt{V_{n,a}}} \ge \delta\right\}. \tag{102}$$

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 B.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.

B.3.1 Proof of Lemma B.1

By invoking bound (66) on minimum eigenvalue of the population covariance, we have

$$\lambda_{\min}(\Sigma) \ge \frac{\lambda_{\min}(\Sigma_\zeta)}{\mu_{\max}(\mathcal{A})}, \tag{103}$$

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 7.2, we obtain

$$\|m_a\|_1 \le \|\Omega e_a\| + \|m_a - \Omega e_a\|_1 \le \frac{s_\Omega}{\lambda_{\min}(\Sigma)} + \frac{192\tau}{\alpha} s_\Omega \sqrt{\frac{\log(dp)}{n}} = O(s_\Omega)\,. \tag{104}$$

We also have

$$\|m_a - \Omega e_a\|_\infty \le \|m_a - \Omega e_a\|_1 = O\Big(s_\Omega\sqrt{\frac{\log(dp)}{n}}\Big)\,. \tag{105}$$

In addition, by the KKT condition (96) we have

$$\|\widehat{\Sigma} m_a - e_a\|_\infty \le \mu\,. \tag{106}$$

Combining bounds (104), (105) and (106), we have

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

which completes the proof.

C Proofs of Section 4

C.1 Consistency results for LASSO under adaptively collected samples

Theorem 4.1 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 4.1 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.

C.1.1 Proof of Theorem 4.1

The important technical step is to prove that, under the conditions specified in Theorem 4.1, 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 C.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 C.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 D.4 we have that $\|x_{i,a}x_{i,b}\|_{\psi_1} \leq 2\|x_i\|_{\psi_2}^2 \leq 2\kappa^2$ almost surely. Then using the tail inequality Lemma D.5 we have for any $\varepsilon \leq 2e\kappa^2$

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

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

Lemma C.3 ([BVDG11, 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 C.1.

Proof of Proposition C.1. Combining Lemmas C.2 and C.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 C.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 D.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 D.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. [HTW15] and is given below for the reader's convenience.

Proof of Theorem 4.1. Throughout we condition on the intersection of good events in Proposition C.1 and Lemma C.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

$$
\begin{aligned}
\frac{1}{2}\langle \nu, \widehat{\Sigma}\nu\rangle &\le \frac{1}{n}\langle X^{\mathsf{T}}\varepsilon, \nu\rangle + \lambda_n(\|\theta_0\|_1 - \|\widehat{\theta}^{\mathsf{L}}\|_1) \\
&\le \frac{1}{n}\|\nu\|_1\|X^{\mathsf{T}}\varepsilon\|_\infty + \lambda_n(\|\theta_0\|_1 - \|\widehat{\theta}^{\mathsf{L}}\|_1) \\
&\le \lambda_n\Big\{\frac{1}{2}\|\nu\|_1 + \|\theta_0\|_1 - \|\widehat{\theta}^{\mathsf{L}}\|_1\Big\}. \qquad (107)
\end{aligned}
$$

First we show that the error vector ν satisfies $\|\nu_{S_0^c}\|_1 \le 3\|\nu_{S_0}\|_1$, where $S_0 \equiv \text{supp}(\theta_0)$. 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.(107) 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.

□

C.2 Bias control: Proof of Theorem 4.6

Recall the decomposition (31) 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 4.1, 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 C.5. *Denote by $\Omega = (\mathbb{E}\{xx^{\mathsf{T}}\})^{-1}$ and $\Omega^{(2)}(\widehat{\theta}) = (\mathbb{E}\{xx^{\mathsf{T}}|\langle x, \widehat{\theta}\rangle \ge \varsigma\})^{-1}$ be the population precision matrices for the first and second batches. Suppose that $n_1 \wedge n_2 \ge 2\Lambda_0/\kappa^2 \log p$. Then, with probability at least $1 - p^{-4}$*

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

In particular, with the same probability, the online debiasing program (29) is feasible with $\mu_\ell = 15\kappa^2{\Lambda_0}^{-1}\sqrt{(\log p)/n_\ell} < 1/2$.

It follows from the lemma, Theorem 4.1 and the previous display that, with probability at least $1 - 2p^{-3}$

$$\begin{aligned}\|\Delta_n\|_\infty &\le \|B_n\|_\infty \|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1 \\ &\le 15\kappa\Lambda_0^{-1/2}\sqrt{n}\Big(\frac{n_1}{n}\sqrt{\frac{\log p}{n_1}} + \frac{n_2}{n}\sqrt{\frac{\log p}{n_2}}\Big)\cdot 120\kappa\sigma\phi_0^{-1}s_0\sqrt{\frac{\log p}{n}}, \\ &\le 2000\frac{\kappa^2\sigma}{\sqrt{\Lambda_0}\phi_0}\frac{s_0\log p}{n}(\sqrt{n_1} + \sqrt{n_2}) \\ &\le 4000\frac{\kappa^2\sigma}{\sqrt{\Lambda_0}\phi_0}\frac{s_0\log p}{\sqrt{n}}. \end{aligned} \tag{108}$$

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 C.4 and G_2 denote the good event of Theorem 4.1 as below:

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

On the intersection $G = G_1 \cap G_2$, Δ_n satisfies the bound (108). 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 D.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 C.5.

Proof of Lemma C.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 D.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 D.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 $\varepsilon \leq 12\kappa^2/\Lambda_0$.

□

C.3 Central limit asymptotics: proofs of Proposition 4.8 and Theorem 4.9

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:

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

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

Lemma C.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 $\{m : \|\Sigma m - v\|_\infty \le \mu, \|m\|_1 \le L\}$. We prove that $|f(\Sigma) - f(\Sigma')| \le L^2 \|\Sigma - \Sigma'\|_\infty$.

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

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

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

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

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

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

On the other hand, the value is given by

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

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 $\|\Sigma^{-1}\|_1 \le L$. The value is then $\langle e_a, \Sigma^{-1} e_a \rangle \le \lambda_{\max}(\Sigma^{-1})$ which gives the upper bound. □

Lemma C.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 4.4 and 4.7*

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

Proof. Using Lemma C.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 C.2 the first and third term vanish in probability. It is straightforward to apply Theorem 4.1 to the intermediate estimate $\widehat{\theta}^1$; indeed Assumption 4.7 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 4.7. ☐

Lemma C.9. *Under Assumptions 4.4 and 4.7, 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 4.8 and Theorem 4.9.

Proof of Proposition 4.8. 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^{th} coordinate $W_{n,a}$) is adapted to the filtration $\mathfrak{F}_i$. We can now apply the martingale central limit theorem [HH14, Corollary 3.1] to $W_{n,a}$ to obtain the result. From Lemmas C.7 and C.8 we know that $V_{n,a}$ is bounded away from 0, asymptotically. The stability and conditional Lindeberg conditions of [HH14, Corollary 3.1] are verified by Lemmas C.8 and C.9. □

Proof of Theorem 4.9. This is a straightforward corollary of the bias bound of 4.6 and Proposition 4.8. We will show that:

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

The reverse inequality follows using the same argument.

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

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

Therefore,

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

By Proposition 4.8 the first term converges to $\Phi(x+\delta)$. To see that the second term vanishes, observe first that Lemma C.7 and Lemma C.8, imply that $V_{n,a}$ is bounded away from 0 in probability. Using this:

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

by applying Theorem 4.6 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. □

C.4 Proofs for Gaussian designs

In this Section we prove that Gaussian designs of Example 4.5 satisfy the requirements of Theorem 4.1 and Theorem 4.6.

The following distributional identity will be important.

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

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

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 C.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 C.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 C.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 C.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 C.10:

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

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:

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

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, $\mathrm{d}\log \mathbb{E}\{e^{\lambda\xi_1'}\}/\mathrm{d}\lambda \le \lambda + C$, 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 $\log \mathbb{E}\{e^{\lambda\xi_1'}\} \le \lambda^2/2$, 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:

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

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

□

For Example 4.5, it remains only to show the constraint on the approximate sparsity of the inverse covariance. We show this in the following

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

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

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

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

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 C.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 C.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

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

Finally (c) holds since

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

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

D Technical preliminaries

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

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

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

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

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

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

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

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

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

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

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 [Ver12] is a Bernstein-type tail inequality for sub-exponential random variables.

Lemma D.5 ([Ver12, 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 $\varepsilon \ge 0$:*

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

We also use a martingale generalization of [Ver12, Proposition 5.16], whose proof is we omit.

Lemma D.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 $\varepsilon \ge 0$:*

$$\mathbb{P}\Big\{\Big|\frac{1}{n}\sum_{i=1}^n X_i\Big| \ge \varepsilon\Big\} \le 2\exp\Big\{-\frac{n\varepsilon}{6eK}\min\Big(\frac{\varepsilon}{eK}, 1\Big)\Big\} \tag{110}$$

The following is a rough bound on the LASSO error.

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

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

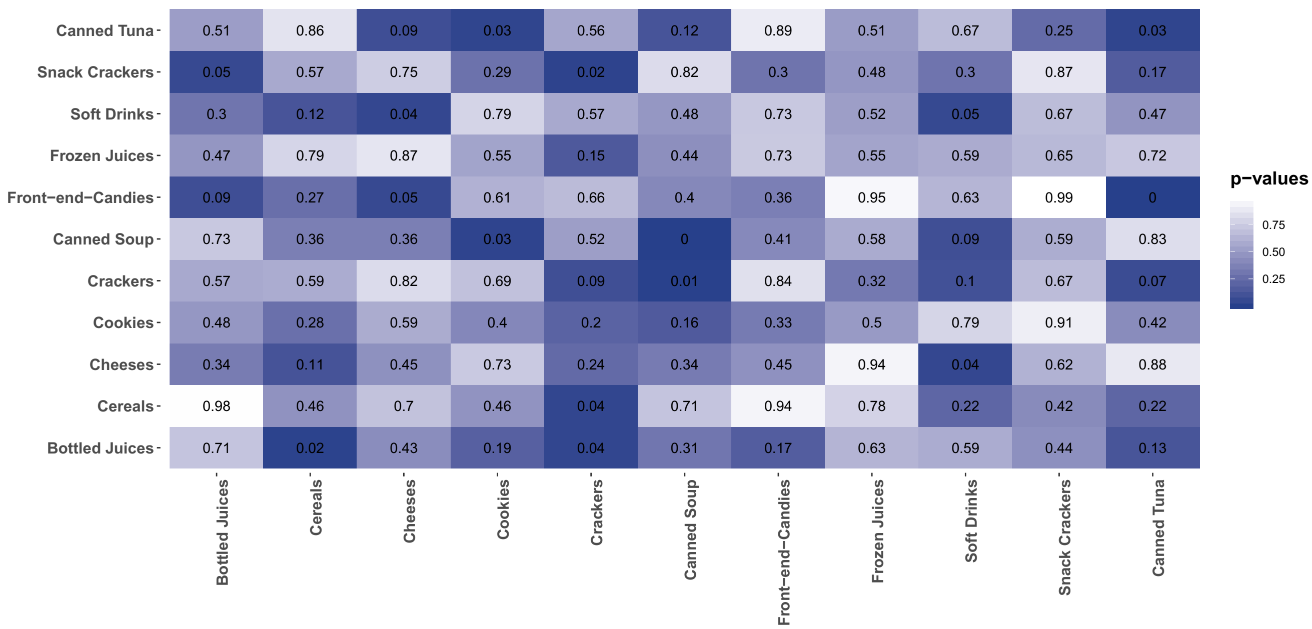
We now use triangle inequality and the bound above to get the claim:

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

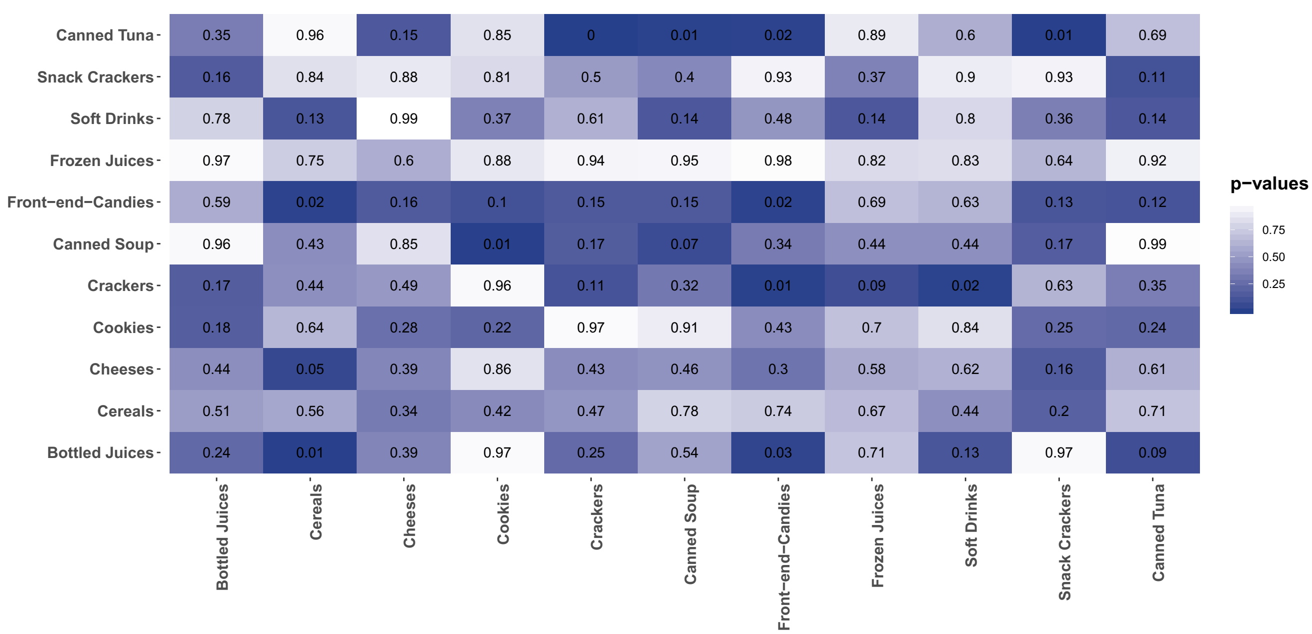
□

E Simulation results for the Dominick's data set

In this section we report the p-values obtained by the online debiasing for the cross-category effects. Figures 9, 10, 11 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.

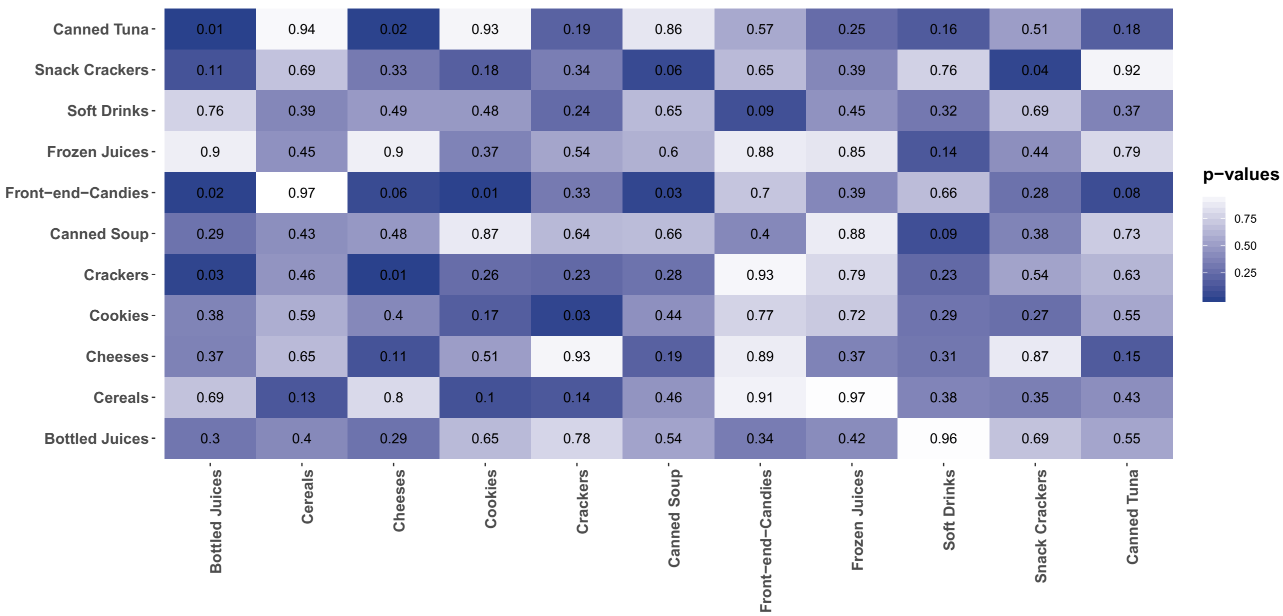


(a) **1-Week** effect of **sales** of $x-$axis categories on **sales** of $y-$axis categories

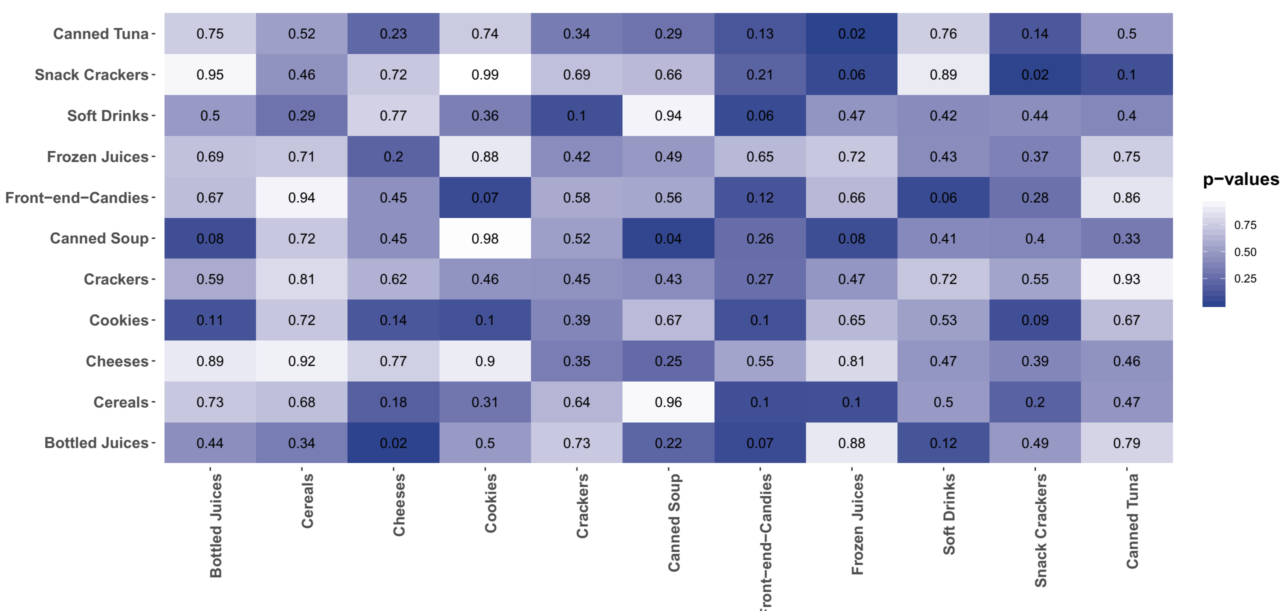


(b) **1-Week** effect of **prices** of $x-$axis categories on **sales** of $y-$axis categories

Figure 9: Figures 9a, and 9b 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.

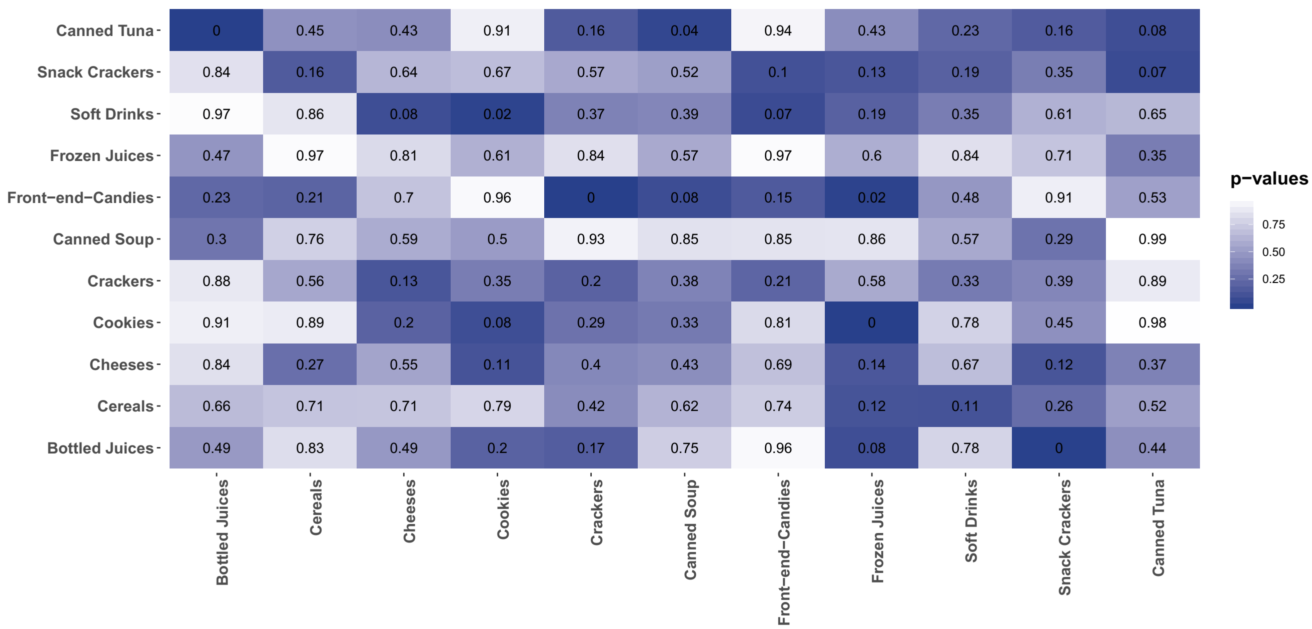


(a) **1-Week** effect of **promotions** of $x-$axis categories on **sales** of $y-$axis categories

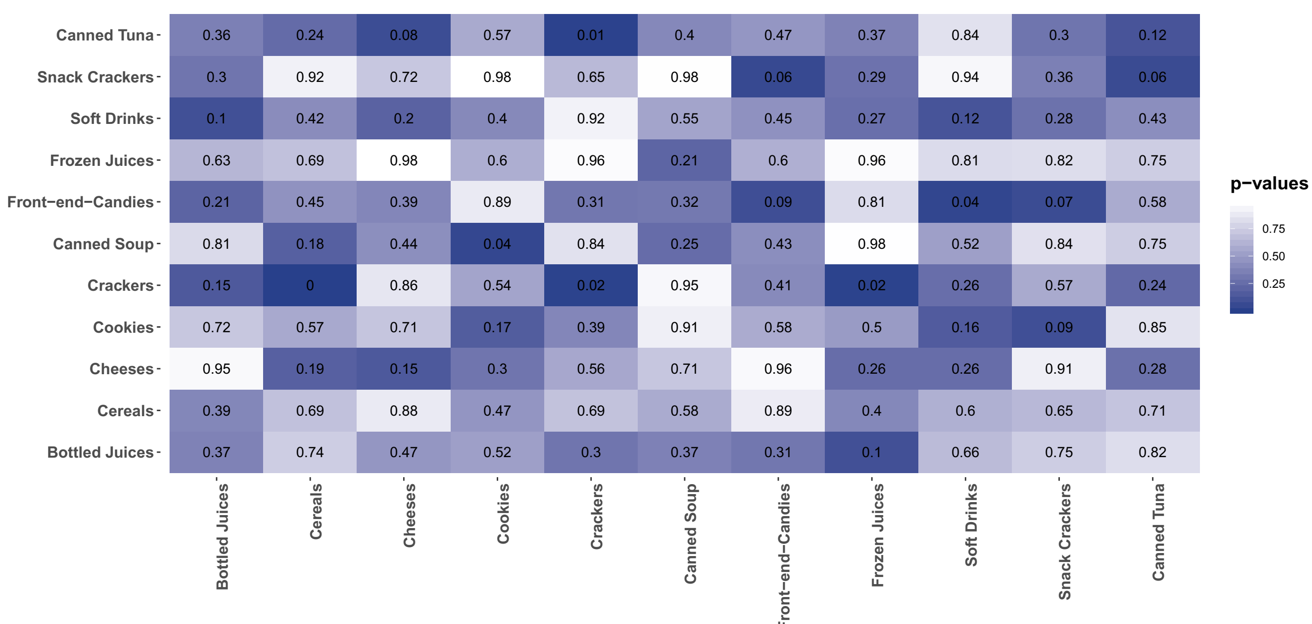


(b) **2-Week** effect of **promotions** of $x-$axis categories on **sales** of $y-$axis categories

Figure 10: Figures 10a, and 10b 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 11: Figures 11a, and 11b 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.