Online Debiasing for Adaptively Collected High-Dimensional Data With Applications to Time Series Analysis

Yash Deshpande, Adel Javanmard, and Mohammad Mehrabi

Institute for Data, Systems and Society, Massachusetts Institute of Technology, Cambridge, MA; Data Sciences and Operations Department, University of Southern California, Los Angeles, CA

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, for example, 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 debias the LASSO estimate when the underlying parameter \(\theta_0\) has sparsity of order \(\sqrt{n}/\log p\). 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 the 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 improve the algorithms underlying the recommendations and influence the design of future phases. In commercial settings, wherein the data from the previous phase inform on the next phase, data collection is often adaptive in nature. For example, clinical trials are run often adaptive in nature. For example, clinical trials are run in phases, wherein the data from the 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 improve the algorithms underlying the recommendations and influence the design of future phases. In commercial settings, wherein the data from the previous phase inform on the next phase, data collection is often adaptive in nature. For example, clinical trials are run in phases, wherein the data from the 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 improve the algorithms underlying the recommendations and influence the design of future phases. In commercial settings, wherein the data from the previous phase inform on the next phase, data collection is often adaptive in nature. For example, clinical trials are run...
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 \( \theta_0 \) and (ii) regularized estimators beyond \( \hat{\theta}^L \), such as the LASSO. Such estimators are nonlinear, nonexplicit and, consequently it is difficult to characterize their distribution even with strong random design assumptions (Javanmard and Montanari 2014b). 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 \((y_1, x_1), \ldots, (y_n, x_n)\) one can compute the LASSO estimate \( \hat{\theta}^L = \hat{\theta}^L(y, X; \lambda_n) \)

\[
\hat{\theta}^L = \arg \min_{\theta} \left\{ \frac{1}{2n} \|y - X\theta\|^2 + \lambda_n \|\theta\|_1 \right\},
\]

If \( \theta_0 \) is sparse with at most \( s_0 \ll p \) nonzero entries and the design matrix \( X \) satisfies some technical conditions, the LASSO estimate, for an appropriate choice of \( \lambda_n \) has estimation error \( \|\hat{\theta}^L - \theta_0\|^2 \) of order \( \sigma^2 s_0 \log p / n \), with high probability (Bass and Michailidis 2015). 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 Equation (2), which disposes \( \hat{\theta}^L \) to have small \( \ell_1 \) norm. The second is the correlation induced between \( X \) and \( \varepsilon \) due to adaptive data collection. To address the first source, Zhang and Zhang (2014), Javanmard and Montanari (2014a), and Van de Geer et al. (2014) proposed the so-called debiased estimate of the form

\[
\hat{\theta}^{\text{off}} = \hat{\theta}^L + \frac{1}{n} MX^T(y - X\hat{\theta}^L),
\]

where \( M \) is chosen as an “approximate inverse” of the sample covariance \( \hat{\Sigma} = X^TX/n \). The intuition for this idea is the following decomposition that follows directly from Equations (1) and (3):²

\[
\hat{\theta}^{\text{off}} - \theta_0 = (I_p - M\hat{\Sigma})(\hat{\theta}^L - \theta_0) + \frac{1}{n} MX^T\varepsilon.
\]

When the data collection is nonadaptive, \( X \) and \( \varepsilon \) are independent and therefore, conditional on the design matrix \( X \), \( MX^T\varepsilon/n \) is distributed as \( N(0, \sigma^2 Q/n) \) where \( Q = M\hat{\Sigma}M^T \). Further, the bias in \( \hat{\theta}^{\text{off}} \) is isolated to the first term, which intuitively should be of smaller order than the second term, provided both \( \hat{\theta}^L - \theta_0 \) and \( M\hat{\Sigma} - I_p \) are small in an appropriate sense. This intuition suggests that, if the second term dominates the first term in \( \hat{\theta}^{\text{off}} \), we can produce confidence intervals for \( \theta_0 \) in the usual fashion using the debiased estimate \( \hat{\theta}^{\text{off}} \) (Javanmard and Montanari 2014a,b; Van de Geer et al. 2014). In the so-called random design setting—when the rows of \( X \) are drawn iid from a broad class of distributions—this approach to inference via the debiased estimate \( \hat{\theta}^{\text{off}} \) enjoys several optimality guarantees: the resulting confidence intervals have minimax optimal size (Javanmard 2014; Javanmard and Montanari 2014a; Cai and Guo 2017), and are semi-parametrically efficient (Van de Geer et al. 2014).

This line of argument breaks down when the samples are adaptively collected, as the debiased estimate \( \hat{\theta}^{\text{off}} \) still suffers the second source of bias. Indeed, this is exactly analogous to \( \hat{\theta}^L \) in low dimensions. Since \( M \) and \( \varepsilon \) are correlated, we can no longer assert that the term \( MX^T\varepsilon/n \) is unbiased. Indeed, characterizing its distribution can be quite difficult, given the intricate correlation between \( M \), \( \varepsilon \) and \( \theta_0 \) 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 article: (i) batched data collection and (ii) autoregressive time series.

### 1.1. Why Offline Debiasing Fails?

#### 1.1.1. Batched Data Collection

Consider an adaptive data collection setting where the experimenter (or analyst) collects data in two phases or batches. In the first phase, the experimenter collects an initial set of samples \((y_1, x_1), \ldots, (y_{n_1}, x_{n_1})\) of size \( n_1 < n \) where the responses follow Equation (1) and the covariates are iid from a distribution \( P_\varepsilon \). Following this, she computes an intermediate estimate \( \hat{\theta}^1 \) of \( \theta_0 \) and then collects additional samples \((y_{n_1+1}, x_{n_1+1}), \ldots, (y_{n_2}, x_{n_2})\) of size \( n_2 - n_1 \), where the covariates \( x_t \) are drawn independently from the law of \( x_t \), conditional on the event \( \{\langle x_t, \hat{\theta}^1 \rangle \geq c\} \), where \( c \) is a threshold, that may be data-dependent. This is a typical scenario where the response \( y_t \) represents an instantaneous reward that the experimenter wishes to maximize, as in multi-armed bandits (Lai and Robbins 1985; Bubeck and Cesa-Bianchi 2012). 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 (Zhou et al. 2008; Kim et al. 2011; Guidance 2018). The multi-armed bandit problem is a standard formalization of this tradeoff, and a variety of bandit algorithms are designed to operate in distinct phases of “explore–then exploit.” Data collected from one arm in such a bandit algorithm can be modeled within our setting. Beyond multi-armed bandit algorithms for response-adaptive trials, the FDA guidance document (Guidance 2018) outlines other motivations for adaptive designs, such as “enrichment,” where the trial is targeted to an adaptively chosen subpopulation. Such settings can be modeled within our framework by replacing \( \theta_1 \) with another “enrichment direction” vector, that can be computed from the data from the first batch.

With the full samples \((y_1, x_1), \ldots, (y_{n_2}, x_{n_2})\) 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 \( s_0 = 10 \) nonzero entries. We obtain the first batch \((y_1, x_1), \ldots, (y_{500}, x_{500})\) of observations with \( y_t = \langle x_t, \theta_0 \rangle + \varepsilon_t, x_t \sim N(0, \Sigma) \) and \( \varepsilon_t \sim N(0, 1) \) where the diagonal of \( \Sigma \) is all ones, \( \Sigma_{a,b} = 0.1, \) if \( |a - b| = 1 \) and \( \Sigma_{a,b} = 0 \) otherwise. Based on this data, we construct an intermediate estimator \( \hat{\theta}^1 \) 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 \( x_{501}, \ldots, x_{1000} \) independently from the law of \( x \sim (\Sigma_{x,x})^{-1} \hat{\Sigma}_{x,\hat{\theta}^1} \) and the corresponding outcomes \( y_{501}, \ldots, y_{1000} \) are generated according to Equation

²The notation \( \hat{\theta}^{\text{off}} \) 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 article.
(1). Unconditionally, $(x, \hat{\theta}^1) \sim N(0, (\hat{\theta}^1, \Sigma \hat{\theta}^1))$, so this choice of threshold corresponds to sampling covariates that correlate with $\hat{\theta}^1$ 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 $(y_1, x_1), \ldots, (y_{1000}, x_{1000})$, we compute the LASSO estimate $\hat{\theta}^2 = \hat{\theta}^2(y, x; \lambda)$ with $\lambda = 2.5\lambda_{\text{max}}(\Sigma)\sqrt{(\log p)/n}$. Offline debiasing yields the following prescription to debias the Lasso estimator: $\hat{\theta}^\text{off} = \hat{\theta}^1 + \frac{1}{\lambda} \Omega(\hat{\theta}^1)'X(y - X\hat{\theta}^1)$, where $\Omega(\hat{\theta})$ is the population precision matrix:

$$\Omega(\hat{\theta}^1)^{-1} = \frac{1}{2} E(xX^T) + \frac{1}{2} E \left[ xx^T \right] (x, \hat{\theta}^1) \geq \| \Sigma^{1/2} \hat{\theta}^1 \| .$$

We generate the dataset for 1000 Monte Carlo iterations and compute the offline debiased estimate $\hat{\theta}^\text{off}$ for each iteration. Figure 1 shows the histogram of the entries $\hat{\theta}^\text{off}$ on the support of $\theta_0$ for the two choices of $\hat{\theta}^1$. As we see $\hat{\theta}^\text{off}$ still has considerable bias, due to adaptivity in the data collection.

1.1.2. Autoregressive Time Series

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

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

where $A^{(\ell)} \in \mathbb{R}^{p \times p}$ are time invariant coefficients and $\zeta_t$ is the noise term satisfying $E(\zeta_t) = 0$ (zero-mean), $E(\zeta_t \zeta_t^T) = \Sigma_\zeta$ (stationary covariance), and $E(\zeta_t \zeta_{t-k}^T) = 0$ for $k > 0$ (no serial correlation). Given the data $z_1, \ldots, z_T$, the task of interest is to perform statistical inference on the model parameters, that is, coefficient matrices $A^{(1)}, \ldots, A^{(d)}$. 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 $z_{t-1}, z_{t-2}, \ldots, z_1$.

As in the batched data example, we will carry out a simple illustration. We generate data from a 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 $\zeta_t \sim \text{N}(0, \Sigma_\zeta)$, with $\Sigma_{ij} = 0.5^{2(|i-j|)}$ for $i, j \in [p]$. Note that this is a high-dimensional setting as the number of parameters $dp^2$ exceeds the sample size $(T - d)p$. As we will discuss in Section 3, the VAR model can be represented as a linear regression model with covariate vectors $x_t = (z_{t-d-1}, \ldots, z_{t})^T \in \mathbb{R}^p$ and noise term $\epsilon_t = \zeta_{d+1, t}$ if the focus is on estimating the $i$th rows of coefficient matrices.

We focus on the noise component of the offline debiased estimate, that is, $W^\text{off} = \frac{1}{n} M \sum_{t=1}^{T-d} x_t \epsilon_t$, with $M$ denoting the decorrelating matrix in the debiased estimate as per (3). The offline debiasing matrix $M$ has rows $(m_a)_{a \in [dp]}$, where $m_a$ is constructed from the following optimization problem:

$$\text{minimize } m^T \hat{\Sigma}_n m \quad \text{subject to } \| \hat{\Sigma}_n m - e_a \|_\infty \leq \mu_n,$$

(6)

with $\hat{\Sigma}_n = \frac{1}{n} \sum_{t=1}^n x_t x_t^T$. Theoretical analysis of Javanmard and Montanari (2014a) suggested the choice of $\mu_n \sim \sqrt{\log(dp)/n}$. In simulations we first solve a linear optimization to find the smallest $\mu_n$ that makes the optimization (6) feasible, and then construct $M$ using optimization (6) for that value of $\mu_n$.

To build some insight regarding optimization (6), let us invoke the bias-variance decomposition in Equation (4). For any fixed coordinate $a$, the corresponding entry of bias $\Delta_a$ is upper bounded by

$$|\Delta_a| \equiv |(e_a - m_a^T \hat{\Sigma}_n)(\hat{\theta}_a^1 - \theta_0a)| \leq |\hat{\Sigma}_n m_a - e_a| \cdot |\hat{\theta}_a^1 - \theta_0a|,$$

(7)

while the noise term $Z_a \equiv \frac{1}{n} m_a^T X^T$ is distributed as $N(0, \frac{1}{n} m_a^T \hat{\Sigma}_n m_a)$. Optimization (6) on the one hand tries to control $|\hat{\Sigma}_n m_a - e_a|$ and hence by Equation (7) controls the non-Gaussianity and bias of $\hat{\theta}^\text{off}$. On the other hand it minimizes the noise variance $m_a^T \hat{\Sigma}_n m_a$, which controls the length of confidence interval for $\theta_0a$, and the statistical power when testing for the null hypothesis $H_{0, a}: \theta_0a = 0$.

In Figure 2, we show the QQ-plot, PP-plot, and histogram of $W^\text{off}$ (corresponding to the entry (1, 1) of matrix $A_1$) for 1000 different realizations of the noise $\zeta$. As we observe, even the noise component $W^\text{off}$ is biased because the offline construction of $M$ depends on all features $x_t$ and hence endogenous noise $\zeta_t$. 

![Figure 1](image-url) Histograms of the offline debiased estimate $\hat{\theta}^\text{off}$ restricted to the support of $\theta_0$. The dashed line indicates the true coefficient size. Recall that the second batch is chosen based on an intermediate estimator $\hat{\theta}^1$ computed on the first batch. (Left) $\hat{\theta}^2$ is debiased LASSO on the first batch, (Right) $\hat{\theta}^1$ is ridge estimate on the first batch. As we observe the offline debiasing (even with access to the precision matrix $\Sigma$ of the random designs) has a significant bias and dose not admit a Gaussian distribution.
debiasing method fails for statistical inference purposes when the samples are correlated. Leading to valid statistical inference in this case.

Debiasing asymptotically removes the bias in the Lasso estimate, if the precision matrix is sufficiently sparse and prove that the offline study adapts settings where the population inverse covariance still works for some adaptive data settings under further structural assumptions.

Why the classical debiasing approach will not work in this case. Inference with adaptive collected samples and demonstrate why the classical debiasing approach will not work in this case.

2. Online Debiasing

We propose online debiased estimator \( \hat{\theta}^{on} = \hat{\theta}^{on} (y, X; (M_i)_{i \leq n}, \lambda) \) that takes the form

\[
\hat{\theta}^{on} = \hat{\theta}^l + \frac{1}{n} \sum_{i=1}^{n} M_i x_i (y_i - X_i^T \hat{\theta}^l).
\]

The term “online” comes from the first crucial constraint of predictability imposed on the sequence \((M_i)_{i \leq n}\).

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

With predictability, the data points \((y_i, X_i)\) are adapted to the filtration \((\mathcal{F}_i)_{i \leq n}\) and, moreover, the covariates \(x_i\) are predictable with respect to \(\mathcal{F}_i\). Intuitively, the \(\sigma\)-algebra \(\mathcal{F}_i\) 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 Equation (4) we can decompose \(\hat{\theta}^{on}\) into two components:

\[
\hat{\theta}^{on} = \theta_0 + \frac{1}{\sqrt{n}} (B_n (\hat{\theta}^l - \theta_0) + W_n)
\]

where \(B_n = \sqrt{n} (I_p - \frac{1}{n} \sum_i M_i x_i x_i^T)\), and

\[
W_n = \frac{1}{\sqrt{n}} \sum_i M_i x_i e_i.
\]

Predictability of \((M_i)_{i \leq n}\) ensures that \(W_n\) is unbiased and the bias in \(\hat{\theta}^{on}\) is contained entirely in the first term \(B_n (\hat{\theta}^l - \theta_0)\). Suppose that, analogous to online debiasing, we prove that the bias term \(B_n (\hat{\theta}^l - \theta_0)\) 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 \(\sqrt{n} W_n = \sum_i M_i X_i e_i\) is a martingale with respect to the filtration \(\mathcal{F}_i\), one might expect that \(W_n\) is asymptotically Gaussian. The following “stability” property, identified first by Lai and Wei (1982) in this context, is crucial to ensure that this intuition is correct.

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

For sums of independent random variables, guaranteeing that (i) the mean of summands vanishes and (ii) each individual summand is negligible in comparison to the sum (e.g., as in the classical Lindeberg condition) are necessary and sufficient for asymptotic normality. Martingales are, however, sums of dependent random variables and, therefore, (i) and (ii) are not sufficient for asymptotic normality. One needs, in general, a further condition on the predictable quadratic variation \(V_n\) (see Dvoretzky 1972; Hall and Heyde 2014). Notion of “stability” given in Definition 2.2 is a commonly used such condition.

An important contribution of our article is to develop online debiasing estimators \(\hat{\theta}^{on}\) whose underlying martingales are stable. The specifics of construction of predictable sequence \((M_i)_{i \leq n}\) and deriving the distributional characterization of the debiased estimator \(\hat{\theta}^{on}\) depend on the context of the problem at hand. In this article, we instantiate this idea in two concrete examples.
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 4.2). These results draw significantly on prior work in high-dimensional estimation (Basu and Michailidis 2015; Bühlmann and Van De Geer 2011).

2. Next, we propose constructions for the online debiasing sequence \( (M_i)_{i \leq n} \), 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 Javanmard and Montanari (2014a). The important change is the inclusion of an \( \ell_1 \) constraint in the program, which ensures stability of the underlying martingales, and allows the use of a martingale CLT theorem in our analysis.

3. We establish a distributional characterization of the resulting online debiased estimate \( \hat{\theta}^{\text{on}} \) (Theorems 3.8 and 4.10). Informally, this demonstrates that coordinates of \( \hat{\theta}^{\text{on}} \) are approximately Gaussian with a covariance computable from data.

### 2.1. Organization

In Section 5, we demonstrate how the online debiased estimate \( \hat{\theta}^{\text{on}} \) can be used to compute standard inferential primitives like confidence intervals and \( p \)-values. In Appendix D (supplementary material), we develop computationally efficient iterative descent methods to construct the online debiasing sequence \( (M_i)_{i \leq n} \). Due to space constraints, we deferred our numerical experiments to Appendix E (supplementary material), where we demonstrate the validity of the methodology on both synthetic and real data. In the interest of reproducibility, we made an R implementation of our algorithm publicly available at Deshpande (Javanmard). Proof of theorems and technical lemmas are given in Appendices F to I.

Our proposal of online debiasing approach builds on the insight in Deshpande et al. (2018), which has studied a similar problem for low-dimensional settings \((p < n)\). We provide a detailed discussion of 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.

### 2.2. Notation

Henceforth, we use the shorthand \([p] \equiv \{1, \ldots, p\}\) for an integer \( p \geq 1 \), 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 \( \ell_p \) norm of a vector \( v \), \( \|v\|_p = (\sum |v|^p)^{1/p} \) and \( \|v\|_0 \) for the number of nonzero elements of \( v \). We also denote by \( \supp(v) \), the support of \( v \), that is the positions of its nonzero entries. For a matrix \( A \), \( \|A\|_p \) represents its \( \ell_p \) 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 \( \ell_1 \) norm of its columns). In addition \( \phi(x) \) and \( \Phi(x) \) respectively denote the density and the distribution functions of standard normal variable. Also, we use the term with \textit{high probability} to imply that the probability converges to one as \( n \to \infty \).

### 3. Online Debiasing for High-Dimensional Time Series

Vector autoregressive (VAR) model is among the most popular models in time series analysis. These models are extensively used across science and engineering (see Holtz-Eakin, Newey, and Rosen 1988; Stock and Watson 2001; Fujita et al. 2007; Seth, Barrett, and Barnett 2015 for notable examples in macroeconomics, genomics and neuroscience). Given the data \( z_1, \ldots, z_T \), the fundamental task is to estimate the parameters of the VAR model, viz, the matrices \( A^{(1)}, \ldots, A^{(d)} \) in Equation (5). 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 \((T - dp)\) linear equations (Equation (5) with \( t = d + 1, \ldots, T \)). For the \( i \)-th coordinate of \( z_t \), Equation (5) reads

\[
 z_{t,i} = \sum_{\ell = 1}^{d} (z_{t-\ell,i} A^{(\ell)}_{i,1}) + \zeta_{t,i},
\]

where \( A^{(\ell)}_{i,1} \) denotes the \( i \)-th row of the matrix \( A^{(\ell)} \). This can be interpreted in the linear regression form, Equation (1), in dimension \( dp \) with \( \theta_0 \in \mathbb{R}^{dp} \), \( X \in \mathbb{R}^{(T-d) \times dp} \), \( y, \varepsilon \in \mathbb{R}^{T-d} \). We fix \( i \) and in order to lighten the notation, we hereafter remove the dependence on \( i \). We let \( \sigma^2 = \sum_{t,i} \varepsilon_t \) the variance of entries of \( \varepsilon \), and denote the rows of \( X \) by \( x_1, \ldots, x_n \in \mathbb{R}^{dp} \), 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 (Lai and Wei 1982; Shumway and Stoffer 2006). Basu and Michailidis (2015) considered 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 (Bühlmann and Van De Geer 2011, theor. 6.1), relying on proving a restricted eigenvalue property for the design \( X'X/n \). This result hinges on stationary properties of the model (5), which we summarize prior to stating the estimation result. We recall the following definition from Basu and Michailidis (2015).

**Definition 3.1** (Stability and invertibility of VAR(d) Process; Basu and Michailidis 2015). A VAR(d) process with an associated
reverse characteristic polynomial
\[ A(\gamma) = I - \sum_{\ell=1}^{d} A^{(\ell)} \gamma^{\ell}, \quad (12) \]
is called stable and invertible if \( \det(A(\gamma)) \neq 0 \) for all \( \gamma \in \mathbb{C} \) with \( |\gamma| = 1 \). Based on this characteristic polynomial, we also define the following spectral parameters:
\[
\mu_{\min}(A) = \min_{|\gamma|=1} \lambda_{\min}(A^*(\gamma)A(\gamma)) \\
\mu_{\max}(A) = \max_{|\gamma|=1} \lambda_{\max}(A^*(\gamma)A(\gamma)).
\]

**Theorem 3.2 (Estimation Bound).** Recall the relation \( y = X\theta_0 + \varepsilon \), where \( X, y, \theta_0 \) are given by Equation (11) and let \( \hat{\theta}^{1} \) be the Lasso estimator
\[
\hat{\theta}^{1} = \text{argmin}_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1 \right\}.
\]
Assume that \( |\text{supp}(\theta_0)| \leq s_0 \), and define
\[
\omega = \frac{d\lambda_{\max}(\Sigma_{\varepsilon})}{\lambda_{\min}(\Sigma_{\varepsilon})}, \quad \mu_{\max}(A), \quad \alpha = \frac{\lambda_{\min}(\Sigma_{\varepsilon})}{\mu_{\max}(A)}.
\]

There exists a universal constant \( C > 0 \), such that for any \( n \geq C\alpha^2s_0\log(dp) \) and \( \lambda_n = \lambda_0\sqrt{\log(dp)/n} \), with \( \lambda_0 \geq 4\lambda_{\max}(\Sigma_{\varepsilon})(1/\sqrt{\mu_{\max}(A)})/\mu_{\min}(A) \) the following happens. With probability at least \( 1 - (dp)^{-\delta} \), the estimate satisfies:
\[
\|\hat{\theta}^1 - \theta_0\|_1 \leq C\frac{\lambda_0}{\alpha} \sqrt{\frac{s_0^2\log(dp)}{n}}.
\]

In short, given the standardized setting where \( \lambda_0, \alpha \) are order one, the \( \ell_1 \) 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 (2015), and relies on establishing a now-standard restricted eigenvalue property for the design \( X'X/n \). 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 F (supplementary material) for the proof, as well as a discussion of the differences with the proof of Basu and Michailidis (2015).

### 3.1. Constructing the Online Debiasing 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 Javanmard and Montanari (2014a) and let the sequence \( \{M_i\}_{i\leq n} \) be defined by \( M_i = M \), if \( i < n/2 \) and \( M_i = 0 \), if \( i \geq n/2 \). 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 \( \sqrt{n}W_n = \sum_i M_iX_i\varepsilon_i \)) 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 \( E_0, \ldots, E_{K-1} \), with \( E_\ell \) of length \( r_\ell \), so that \( \sum_{\ell=0}^{K-1} r_\ell = n \). Over an episode \( \ell \), we keep the debiasing matrix \( M_\ell = M^{(\ell)} \) to be fixed over time points in the episode. Moreover, \( M^{(\ell)} \) is constructed using all the time points in previous episodes \( E_0, \ldots, E_{\ell-1} \) in the following way. Let \( n_\ell = r_0 + \cdots + r_{\ell-1} \), for \( \ell = 1, \ldots, K \); hence, \( n_K = n \). Define the sample covariance of the features in the first \( \ell \) episodes.
\[
\hat{\Sigma}^{(\ell)} = \frac{1}{n_\ell} \sum_{t \in E_\ell} x_t x_t^T
\]
The matrix \( M^{(\ell)} \) has rows \( \{m^{(\ell)}_a\}_{a \in [dp]} \) as the solution of the optimization:
\[
\begin{align*}
&\text{minimize } m^T \hat{\Sigma}^{(\ell)} m \\
&\text{subject to } \|\hat{\Sigma}^{(\ell)} m - e_a\|_\infty \leq \mu_\ell, \quad \|m\|_1 \leq L,
\end{align*}
\]
for appropriate values of \( \mu_\ell, L > 0 \). We then construct the online debiased estimator for coordinate \( a \) of \( \theta_0 \) as follows:
\[
\hat{\theta}^{on} = \hat{\theta}^1 + \frac{1}{n} \sum_{\ell=1}^{K-1} \sum_{t \in E_\ell} M^{(\ell)} x_t (y_t - (x_t, \hat{\theta}^1)) \quad (15)
\]

**Figure 3** shows a schematic illustration on this construction. As we discuss later, the tuning parameters \( \mu_\ell \) and \( L \) control the bias of \( \hat{\theta}^{on} \) (See bias-variance decomposition of \( \hat{\theta}^{on} \) given by Equation (20) and Lemma E.5 for rigorous claim.) The objective function \( m^T \hat{\Sigma}^{(\ell)} m \) directly impacts the variance of \( \hat{\theta}^{on} \) and hence controls the length of confidence intervals, or the statistical power when testing the null hypothesis \( H_{0,a} : \theta_{0,a} = 0 \). In other words, the tuning parameters \( \mu_\ell \) control the interpolation between bias and variance of \( \hat{\theta}^{on} \). We would like to choose \( \mu_\ell \) small enough to reduce the bias, but large enough so that the optimization (14) is still feasible. The following lemma shows that, with high probability, \( \mu_\ell \) of order \( \omega\sqrt{\log(dp)/n_\ell} \) is sufficient to make the optimization feasible. Recall that \( \omega \) is defined in Theorem 3.2.

**Lemma 3.3.** Let \( \Omega = \Sigma^{-1} = (E[X_iX_i^T])^{-1} \) be the precision matrix of the time series. There exists universal constants \( C, C' \) such that the following happens. Suppose that \( n_\ell \geq C_0^2 \log(dp) \). Then with probability \( 1 - (dp)^{-\delta} \),
\[
\max_{t,j} |\Omega \hat{\Sigma}^{(\ell)} - I(i = j)| \leq C' \sqrt{\frac{\log(dp)}{n_\ell}}.
\]

The proof of Lemma 3.3 is given in Appendix F.3 (supplementary material).

In simulations and in our developed software package Deshpande (Javanmard), we first solve a linear optimization to find the smallest \( \mu_\ell \) that makes the optimization (14) feasible, and then construct \( M^{(\ell)} \) using that value of \( \mu_\ell \).

### 3.1.1. Choice of Episode Length

In Section 3.2, we show that the constructed online debiased estimator \( \hat{\theta}^{on} \) is asymptotically unbiased and admits a normal distribution. To do that we provide a high probability bound on the bias of \( \hat{\theta}^{on} \) (see Lemma F.5). This bound is in terms of the batch sizes \( r_\ell \), from which we propose the following...
samples in the first episode. On the one side, we want to keep qualitatively. Note that the online debiasing does not debias the ℓ is asymptotically unbiased assuming μℓ becomes more effective as μℓ gets smaller and hence we would like to explain the role of batch sizes qualitatively. Note that the online debiasing does not debias the samples in the first episode. On the one side, we want to keep r₀ small enough to control the contribution of this episode to the bias of ̂θ. On the other hand, we want to have a large β> 1, the online debiased estimator ̂θ is biased. Here, we look into the noise component of the online debiased estimator given by Basu and Michailidis (2015)

\[ \Gamma_\ell(\ell) = \frac{1}{2\pi} \int_{-\pi}^{\pi} A^{-1}(e^{-j\theta}) \Sigma_\ell(A^{-1}(e^{-j\theta}))^* e^{i\ell\theta} d\theta, \]

where A(γ) is given in Equation (12). Figure 4 shows the heat maps of magnitudes of the elements of Σ and the precision matrix Ω = Σ⁻¹ for the on hand 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

\[ W^{on} = \frac{1}{\sqrt{n}} \sum_{\ell=1}^{K-1} M^{(\ell)} \sum_{i \in E_\ell} x_i, \]

However, as we will observe the constructed online debiased estimator empirically admits an unbiased normal distribution.

Revisiting the Numerical Example From Section 1.1.2. In Section 1.1.2, we considered a VAR(d) model with p = 15, d = 5, T = 60, and diagonal A(ℓ) matrices with value b = 0.15 on their diagonals. The covariance matrix Σ_ℓ of the noise terms ξ_ℓ is chosen as Σ_ℓ(i,j) = ρ^{|i-j|} with ρ = 0.5 and i, j ∈ [p]. The population covariance matrix of vector x_ℓ = (ζ_{l+1},...,ζ_{l+1}) is a dp by dp matrix Σ comprising d blocks of size p × p with Σ_(r-s) as block (r, s). The analytical formula to compute Γ_ℓ(ℓ) is given by Basu and Michailidis (2015)

Figure 3. Schematic for constructing the debiasing matrices M^{(ℓ)}. We divide time into K episodes E₀, . . . , E_{K-1}; in episode ℓ, M_ℓ is held constant at M^{(ℓ)}, which is a function of x_ℓ in all prior episodes.

Figure 4. Heat maps of magnitudes of elements of covariance matrix Σ ≡ E(x|x_0^T) (left plot), and precision matrix Ω = Σ⁻¹ (right plot). In this example, x_ℓ’s are generated from a VAR(d) model with covariance matrix of noise Σ_ℓ(i,j) = ρ^{|i-j|} with values d = 5, p = 15, T = 60, ρ = 0.5, and diagonal A(ℓ) matrices with b = 0.15 on diagonals.

For finite sample size, however, different choices of β may have different effect in removing the bias. In our software package Deshpande (Javanmard), we have left it as a user-defined input parameter. We refer to Appendix C (supplementary material) for further discussion on robustness of the inference outputs to the choice of β.

Before proceeding into the distributional characterization of the online debiased estimator for θ_0 (entries of coefficient matrices A^{(ℓ)}) we revisit the numerical example from Section 1.1.1 in which the (offline) debiased estimator of Javanmard and Montanari (2014a) does not display an unbiased normal distribution.
with $M^{(\ell)}$ constructed from the solutions to optimization (14) for $\ell = 1, \ldots, K - 1$. Also, recall that $\varepsilon = (\xi_{d+1,l}, \xi_{d+2,l}, \ldots, \xi_{T,l})$ by Equation (11).

In Figure 5, we show the QQ-plot, PP-plot and histogram of $W_n^0$ and $W_n^{\text{off}}$ (corresponding to the entry (1, 1) of matrix $A_1$) for 1000 different realizations of the noise $\xi_t$. As we observe, even the noise component $W_n^{\text{off}}$ is biased because the offline construction of $M$ depends on all features $x_t$ and hence on endogenous noise $\xi_t$. However, the online construction of decorrelating matrices $M^{(\ell)}$, makes the noise term a martingale and hence $W_n^{\text{on}}$ converges in distribution to a zero mean normal vector, allowing for a distributional characterization of the online debiased estimator.

### 3.2. Distributional Characterization of Online Debiasing

We start our analysis of the online debiased estimator $\hat{\theta}^{\text{on}}$ by considering a bias-variance decomposition. Using $y_t = \langle x_t, \theta_0 \rangle + \varepsilon_t$ in the definition (15)

$$\hat{\theta}^{\text{on}} - \theta_0 = \hat{\theta}^{\text{L}} - \theta_0 + \frac{1}{n} \sum_{\ell=1}^{K-1} \sum_{t \in E_\ell} M^{(\ell)} x_t x_t^T (\theta_0 - \hat{\theta}^{\text{L}}) + \frac{1}{n} \sum_{\ell=1}^{K-1} \sum_{t \in E_\ell} M^{(\ell)} x_t \varepsilon_t$$

$$= \left( I - \frac{1}{n} \sum_{\ell=1}^{K-1} \sum_{t \in E_\ell} M^{(\ell)} x_t x_t^T \right) \left( \hat{\theta}^{\text{L}} - \theta_0 \right) + \frac{1}{n} \sum_{\ell=1}^{K-1} \sum_{t \in E_\ell} M^{(\ell)} x_t \varepsilon_t , \tag{17}$$

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

$$B_n \equiv \sqrt{n} \left( I - \frac{1}{n} \sum_{\ell=1}^{K-1} \sum_{t \in E_\ell} r_{\ell} M^{(\ell)} R^{(\ell)} \right), \tag{18}$$

$$W_n \equiv \frac{1}{\sqrt{n}} \sum_{\ell=1}^{K-1} M^{(\ell)} \left( \sum_{t \in E_\ell} x_t \varepsilon_t \right), \tag{19}$$

we arrive at the following decomposition

$$\hat{\theta}^{\text{on}} = \theta_0 + \frac{1}{\sqrt{n}} \left( B_n (\hat{\theta}^{\text{L}} - \theta_0) + W_n \right). \tag{20}$$

Our first set of results concern the bias of $\hat{\theta}^{\text{on}}$, establishing that it is of smaller order compared to the bias of the LASSO estimate. The following theorem uses Lemma 3.3 to control the bias of the online debiased estimator.

#### Theorem 3.4. (Bias control) Consider the VAR($d$) model (5) and let $\hat{\theta}^{\text{on}}$ be the debiased estimator (15) where the decorrelating matrices $M^{(\ell)}$ are computed according to Equation (14), with $\mu_t = \varepsilon_t \omega \sqrt{\log(dp)/n}$ and $L \geq \|\Sigma_1\|_1$. Further assume that the base estimator is $\hat{\theta}^{\text{L}}$ computed with $\lambda = \lambda_0 \sqrt{\log(dp)/n}$ where $\lambda_0 \geq 4\lambda_{\max}(\Sigma_\varepsilon) (1 + \sqrt{\mu_{\max}(A)})/\mu_{\min}(A)$.

Then, under the sample size condition $n \geq C_0 \omega^2 s_0 \log(dp)$, we have

$$\sqrt{n}(\hat{\theta}^{\text{on}} - \theta_0) = W_n + \Delta_n, \tag{21}$$

where $E[W_n] = 0$ and

$$\mathbb{P} \left( \|\Delta_n\|_\infty \geq C_1 \frac{\lambda_0 (\omega + L \gamma) s_0 \log(dp)}{\alpha} \right) \leq (dp)^{-4}. \tag{22}$$

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

$$\|E(\hat{\theta}^{\text{on}} - \theta_0)\|_\infty \leq \frac{C_1 \lambda_0 (\omega + L \gamma) s_0 \log(dp)}{\alpha} + \frac{C_2 \|\theta_0\|_1}{(dp)^6}.$$ 

We refer to Appendix F.4 (supplementary material) for the proof of Theorem 3.4.

Note that the above theorem bounds the bias term $\Delta_n$ 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 $\lambda_{\min}(\Sigma_\varepsilon), \lambda_{\max}(\Sigma_\varepsilon), \mu_{\min}(A)$ and $\mu_{\max}(A)$ are bounded away from 0 and $\infty$, as $n, p \to \infty$.
2. With $\Omega = \Sigma^{-1} = (E[x_t x_t^T])^{-1}$ the precision matrix of the data points $[x_t]$, and $s_0$ the sparsity of $\theta_0 = (A_1^{(i)}, \ldots, A_d^{(i)})^T$, we assume that $\|\Omega\|_1 = o(\sqrt{n}/\log(dp))$.
Under Assumption 3.5, the spectral quantities \( \omega, \gamma, \alpha \) and therefore \( \lambda_0 \) 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 \((dp)^\epsilon\) for arbitrary large \( \epsilon > 0 \), by adjusting constant \( C_1 \) and the tail bound in Equation (22). 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 \( L_0(\log p)/n \). On the other hand, recall the filtration \( F_t \) generated by \( \{e_t, \ldots, e_1\} \) and rewrite (19) as \( W_n = \sum_{t=1}^n v_t e_t \), where \( v_t = M_{\phi(t)}(F_{t-1})/\sqrt{n} \) (Sample \( t \) belongs to episode \( \ell \)). We use Assumption 3.5 in Lemma 3.6 below, to show that for each coordinate \( a \in [dp] \), the conditional variance \( \sum_{t=1}^n \mathbb{E}(x_t^2 | F_{t-1}) \) is of order one. Hence \( || \Delta_n ||_\infty \) is asymptotically dominated by the noise variance when \( s_0 = o \left( \sqrt{\frac{n}{\log(dp)}} \right) \).

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 (see, e.g., Hall and Heyde 2014, corol. 3.2) and conclude that the unbiased component \( W_n \) admits a Gaussian limiting distribution.

Lemma 3.6. (Stability of martingale \( W_n \)) Let \( \hat{\theta}^{on} \) be the debiased estimator (15) with \( \mu_{\ell} = \tau \sqrt{\log (p)/n_{\ell}} \) and \( L = L_0 ||\Omega||_1 \), for an arbitrary constant \( L_0 \geq 1 \). Under Assumption 3.5, and for any fixed sequence of integers \( a(n) \in [dp] \), we have

\[
V_{n,a} = \frac{\sigma^2}{n} \sum_{\ell=1}^{K-1} \sum_{t \in E_\ell} (m_{\ell,a}^t x_t)^2 = \sigma^2 \Omega_{a,a} + o_P(1).
\]

In addition, we have

\[
\max \left\{ \frac{1}{\sqrt{n}} \langle (m_{\ell,a}^t x_t) e_t \rangle : \ell \in [K-1], t \in [n-1] \right\} = o_P(1).
\]

Proof of Lemma 3.6 is given in Appendix F.5 (supplementary material). With Lemma 3.6 in place, we apply a martingale central limit theorem (Hall and Heyde 2014, corol. 3.2) to obtain the following result.

Corollary 3.7. Consider the VAR(d) model (5) for time series and let \( \hat{\theta}^{on} \) be the debiased estimator (15) 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

\[
V_{n,a} = \frac{\sigma^2}{n} \sum_{\ell=1}^{K-1} \sum_{t \in E_\ell} (m_{\ell,a}^t x_t)^2.
\]

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

\[
\lim_{n \to \infty} \mathbb{P} \left\{ \frac{W_{n,a}}{\sqrt{V_{n,a}}} \leq x \right\} = \Phi(x),
\]

where \( \Phi \) is the standard Gaussian cdf.

For the task of statistical inference, Theorem 3.4 and Corollary 3.7 suggest to consider the scaled residual \( \sqrt{n}(\hat{\theta}^{on} - \theta_{0,a})/\sqrt{V_{n,a}} \) 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 F.6 (supplementary material). In its statement, we omit explicit constants that can be easily derived from Theorem 3.4.

Theorem 3.8. Consider the VAR(d) model (5) for time series and let \( \hat{\theta}^{on} \) be the debiased estimator (15) 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 \( s_0 = o \left( \sqrt{\frac{n}{\log(dp)}} \right) \), then the following holds true for any fixed sequence of integers \( a(n) \in [dp] \). For all \( x \in \mathbb{R} \), we have

\[
\lim_{n \to \infty} \mathbb{P} \left\{ \frac{\sqrt{n}(\hat{\theta}^{on} - \theta_{0,a})}{\sqrt{V_{n,a}}} \leq x \right\} = \Phi(x) = 0.
\]

Remark 3.9. Note that the conditional variance \( V_{n,a} \) involves the noise variance \( \sigma^2 \), which is unknown in practice, and needs to be estimated from data. In Appendix B (supplementary material), we give a proposal to estimate \( \sigma^2 \) form data.

4. Batched Data Collection

Recall the setting of adaptive data collection in batches from Section 1.1. In this setting, the samples naturally separate into two batches: the first \( n_1 \) data points where the covariates are iid from a distribution \( F_x \), and the second batch of \( n_2 \) data points, where the covariates are drawn independently from the law of \( x_1 \), conditional on the event \( \{x_1, \theta_1^\star \} \geq \zeta \), with \( \zeta \) potentially a data-dependent threshold. The following theorem is a version of (Bühlmann and Van De Geer 2011, theor. 6.1) and is proved in an analogous manner. It demonstrates that even with adaptive data collection consistent estimation using the LASSO is possible. Before proceeding to theorem statement, for the sake of reader’s convenience and to be self-contained let us recall the notion of compatibility.

Definition 4.1. For a symmetric matrix \( \Gamma \in \mathbb{R}^{p \times p} \) and a subset \( S \subseteq [p] \), we say that \( \Gamma \) is \( (\phi_0, S) \)-compatible if

\[
\min_{\theta \in \mathbb{R}^p} \left\{ \frac{|\langle \theta, \Gamma \theta \rangle|}{\|\theta_S\|^2} : \theta \in \mathbb{R}^p, \|\theta_S\|_1 \leq 3\|\theta_S\|_1 \right\} \geq \phi_0.
\]

Theorem 4.2 (Bühlmann and Van De Geer 2011, Theorem 6.1). Suppose that the true parameter \( \theta_0 \) is \( s_0 \)-sparse and the distribution \( F_x \) is such that with probability one the following two conditions hold: (i) the covariance \( \mathbb{E}(xx^\star) \) and \( \mathbb{E}(xx^\star|x, \theta_1^\star) \geq \zeta \) are \( \phi_0 \)-sub-Gaussian. Suppose that \( n \geq C_2 \frac{s_0^2 \zeta^2}{\phi_0^2} \log p \). Then, the LASSO estimate \( \hat{\theta}(y, x; \lambda_n) \) with \( \lambda_n = C_2 \kappa \sigma \sqrt{(\log p)/n} \) satisfies the following bound, with probability exceeding \( 1-p^{-3} \),

\[
||\hat{\theta} - \theta_0||_1 \leq C_1 \kappa \frac{\lambda_n}{\phi_0} \frac{\sqrt{\log p}}{n}.
\]
Remark 4.3. (Estimating the noise variance) For the correct estimation rate using the LASSO, Theorem 4.2 requires knowledge of the noise level \( \sigma \), which is used to calibrate the regularization \( \lambda_n \). Other estimators like the scaled LASSO (Sun and Zhang 2012) or the square-root LASSO (Belloni, Chernozhukov, and Wang 2011) allow to estimate \( \sigma \) consistently when it is unknown. This can be incorporated into the present setting, as done in Javanmard and Montanari (2014a). For simplicity, we focus on the case when the noise level is known. However, the results hold as far as a consistent estimate of \( \sigma \) is used. Formally, a consistent estimator refers to an estimate \( \hat{\sigma} = \hat{\sigma}(y, X) \) of the noise level satisfying, for any \( \varepsilon > 0 \),

\[
\lim_{n \to \infty} \sup_{\|h\|_2 \leq \delta} \mathbb{P} \left( \left| \frac{\hat{\sigma}}{\sigma} - 1 \right| \geq \varepsilon \right) = 0.
\]

Remark 4.4. At the expense of increasing the absolute constants in Theorem 4.2, the probability \( 1 - p^{-c} \) 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:

\[
\hat{\theta}^{\text{on}} = \hat{\theta} + \frac{1}{n} M^{(1)} X_1^T (y^{(1)} - X_1 \hat{\theta}) + \frac{1}{n} M^{(2)} X_2^T (y^{(2)} - X_2 \hat{\theta}),
\]

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 Equation (28) follows from the general recipe in Equation (8) by setting

- \( M_i = M^{(1)} \) for \( i = [n_1] \) and \( M_i = M^{(2)} \) for \( i = n_1 + 1, \ldots, n \).
- Filtrations \( \mathcal{F}_i \) constructed as follows. For \( i < n_1, y_1, \ldots, y_i, X_1, \ldots, X_i, \) and \( \epsilon_1, \ldots, \epsilon_i \) are measurable with respect to \( \mathcal{F}_i \).
- For \( i \geq n_1, y_1, \ldots, y_i, X_1, \ldots, X_n, \) and \( \epsilon_1, \ldots, \epsilon_i \) are measurable with respect to \( \mathcal{F}_i \).

Remark 4.5. (Requirements of design) Suppose that distribution \( \mathbb{P}_x \) and the intermediate estimate \( \hat{\theta} \), that is used in collecting the second batch, satisfy the following:

1. There exists a constant \( \Lambda_0 > 0 \) so that the eigenvalues of \( \mathbb{E}[xx^T'] \) and \( \mathbb{E}[xx^T|x, \hat{\theta}'] \geq \zeta \) are bounded below by \( \Lambda_0 \).
2. The laws of \( x \) and \( x|_{(x, \hat{\theta}') \geq \zeta} \) are \( \kappa \)-sub-Gaussian for a constant \( \kappa > 0 \).
3. The precision matrices \( \Omega = \mathbb{E}[xx^T|x, \hat{\theta}'] \geq \zeta \), satisfy \( \|\Omega\|_1 \vee \|\Omega^{(2)}(\hat{\theta}')\|_1 \leq L \).
4. The conditional covariance \( \Sigma^{(2)}(\theta') \) is \( \kappa \)-Lipschitz in its argument \( \theta' \), that is, \( \|\Sigma^{(2)}(\theta') - \Sigma^{(2)}(\theta)\|_\infty \leq K|\theta' - \theta|_1 \).

The first two conditions of Assumption 4.5 are for ensuring that the base LASSO estimator \( \hat{\theta} \) 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 4.5. We refer to Section H.4 for its proof.

Example 4.6. Let \( \mathbb{P}_x = \mathcal{N}(0, \Sigma) \) and \( \hat{\theta} \) be any vector such that \( \|\hat{\theta}\|_\infty \leq L \lambda_{\min}(\Sigma) \|\theta\|_2 / 2 \) and \( \|\Sigma^{-1}\|_1 \leq L \Sigma / 2 \). Then the distributions of \( x \) and \( x|_{(x, \hat{\theta}') \geq \zeta} \), with \( \zeta = \zeta(\hat{\theta}, \Sigma, \hat{\theta}')^{1/2} \) for a constant \( \zeta \geq 0 \) satisfy the conditions of Assumption 4.5 with \( \Lambda_0 = \lambda_{\min}(\Sigma), \quad \kappa = 3 \lambda_{\max}(\xi) |\xi' \vee \xi^{-1}|^{1/2} \).
\[ K = \sqrt{\bar{\sigma}(1 + \frac{\lambda_{\text{max}}(\Sigma)^{3/2}}{\lambda_{\text{min}}(\Sigma)^{1/2}})} = L \Sigma. \]

Under Assumption 4.5, we provide a nonasymptotic bound on the bias of the online debiased estimator \( \hat{\theta}_{\text{on}} \).

Theorem 4.7. (Nonasymptotic bound on bias) Under Assumption 4.5, there exist universal constants \( C_1, C_2, C_3 \) so that, when \( n \geq C_1 \kappa k_0^2 \log p/\phi_0^2 \) and \( n_1 \wedge n_2 \geq C_1 (\Lambda_0/\kappa^2 + k^2/\Lambda_0) \log p, \) we have that

\[ \sqrt{n} (\hat{\theta}_{\text{on}} - \theta_0) = W_n + \Delta_n, \]

where \( \mathbb{E}(W_n) = 0 \) and

\[ \mathbb{P}\left( \left\| \Delta_n \right\|_\infty \geq \frac{C_2 \kappa^2 \sigma_0 \log p}{\Lambda_0^{3/2} \sqrt{n}} \right) \leq p^{-3}. \]

Further we have

\[ \left\| \mathbb{E}(\hat{\theta}_{\text{on}} - \theta_0) \right\|_\infty \leq \frac{C_2 \kappa^2 \sigma_0 \log p}{\Lambda_0^{3/2} n} + \frac{C_3 \sigma_0 \kappa \| \theta_0 \|_1}{p^2}. \]  

The proof of Theorem 4.7 is given in Appendix H.2 (supplementary material). 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 \( \hat{\theta}^a \) whose bias is typically of order \( \kappa \sigma \sqrt{\log p/n}. \) 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 \( \sigma^2/n. \)

In order to establish asymptotic Gaussian behavior of the online debiased estimate \( \hat{\theta}_{\text{on}} \), we consider a specific asymptotic regime for the problem instances.

Assumption 4.8. (Asymptotic regime.) We consider problem instances indexed by the sample size \( n, \) where \( n, p, s_0 \) satisfy the following:

1. \( \lim inf_{n \to \infty} n_1/n_2 \geq c, \) 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

\[ \lim_{n \to \infty} \frac{1}{\Lambda_0} \sqrt{s_0} \left( L^2 K \vee \sqrt{\log p/\Lambda_0} \right) = 0. \]

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 4.9. Suppose that Assumption 4.5 holds and consider the asymptotic regime of Assumption 4.8. 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

\[ V_{n,a} = \sigma^2 \left( \frac{n_1}{n}(m_a^{(1)} + \lambda_{\text{max}}^{(1)} m_a^{(1)}) + \frac{n_2}{n}(m_a^{(2)} + \lambda_{\text{max}}^{(2)} m_a^{(2)}) \right). \]

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

\[ \lim_{n \to \infty} \mathbb{E} \left[ \varphi \left( \frac{W_{n,a}}{\sqrt{V_{n,a}}} \right) \right] = \mathbb{E}(\varphi(\xi)), \]

where \( \xi \sim N(0, 1). \)

The proof of Proposition 4.9 is deferred to Appendix H.3 (supplementary material). The combination of Theorem 4.7 and Proposition 4.9 immediately yields the following distributional characterization for \( \hat{\theta}_{\text{on}}. \)

Theorem 4.10. Under Assumptions 4.5 and 4.8, the conclusion of Proposition 4.9 holds with \( \sqrt{n} (\hat{\theta}_{\text{on}} - \theta_0) \) in place of \( W_n. \) In particular,

\[ \lim_{n \to \infty} \mathbb{P}\left( \sqrt{n} (\hat{\theta}_{\text{on}} - \theta_{0,a}) \leq x \right) = \Phi(x), \]

where \( V_{n,a} \) is defined as in Proposition 4.9.

To compare the sample size requirements made for \( \ell_1 \)-consistent estimation and those in Assumption 4.8, it is instructive to simplify to the case when \( \kappa, \phi_0, \Lambda_0 \) are of order one. Then \( \ell_1 \)-consistency (Theorem 4.2 in Appendix H, supplementary material) requires that \( n_1 \geq n_2 = \Omega(s_0^2 \log p) \), that is, at least one of the batches is larger than \( s_0^2 \log p. \) However, Theorem 4.10 makes the same assumption on \( n_1 \wedge n_2, \) or both batches exceed \( s_0 \log p \) in size. For online debiasing, this is the case of interest. Indeed if \( n_1 \ll n_2 \) (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 4.8, 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 4.10 via online debiasing.

4.1.1. Revisiting the Numerical Example from Section 1.1.1.

In the batched data example discussed in Section 1.1.1, 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 \( \hat{\theta}_{\text{on}} \) on the support of \( \theta_0 \) (blue) along with the corresponding histogram of entries of the debiased estimator \( \hat{\theta}_{\text{off}} \) (red). As we see for both choices of \( \hat{\theta}^1 \) (debiased LASSO and ridge estimate on the first batch), the online debiased estimator \( \hat{\theta}_{\text{on}} \) 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, nonadaptive batch

\[ \hat{\theta}_{\text{off},1} \equiv \hat{\theta}^1 (y^{(1)}, X_1) + \frac{1}{n} \Omega X_1^T (y^{(1)} - X_1 \hat{\theta}^1 (y^{(1)}, X_1)). \]

Figure 6 (right panel) shows the histogram of the entries of \( \hat{\theta}_{\text{off},1} \) restricted to the support of \( \theta_0, \) and the comparison with \( \hat{\theta}_{\text{on}}. \) As can be expected, both \( \hat{\theta}_{\text{off},1} \) and \( \hat{\theta}_{\text{on}} \) are appropriately matched.
centered around the true coefficient \( \theta \). However, as is common with sample-splitting, \( \hat{\theta}^{\text{off},1} \) displays a larger variance and correspondingly loses power in comparison with \( \hat{\theta}^{\text{on}} \) 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 Deshpande et al. (2018). We would like to discuss two points in comparing the contribution of Deshpande et al. (2018) with ours:

1. The method of Deshpande et al. (2018) is tailored to low-dimensional setting where the number of covariates \( p \) is less than the sample size \( n < p \). Specifically, denoting by \( \lambda_{\text{min}}(n) \) the minimum eigenvalue of \( X^T X \), Deshpande et al. (2018) considered a setting where \( \lambda_{\text{min}}(n) \rightarrow \infty \) almost surely. Note that for the batched data example, this amounts to \( \sqrt{n} - \sqrt{p} \rightarrow \infty \).

2. The work Deshpande et al. (2018) proposed 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 \( \hat{\theta}^{\text{OLS}} \) the least-square estimator, Deshpande et al. (2018) constructed a debiased estimator \( \hat{\theta}^d \) as follows:

\[
\hat{\theta}^d = \hat{\theta}^{\text{OLS}} + W_n (y - X \hat{\theta}^{\text{OLS}}),
\]

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

\[
w_n = \arg \min_{w \in \mathbb{R}^p} \| I - W_{n-1} X_{n-1} - w x_n^T \|^2 + \lambda \| w \|^2.
\]

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 \( w_l \) corresponds to \( M \lambda_l \) in our notation.

One can potentially think of using the ridge-type debiased estimator (41) in high-dimensional setting with using \( \hat{\theta}^d \) instead of \( \hat{\theta}^{\text{OLS}} \). 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 Deshpande et al. (2018). Considering the bias-variance decomposition of \( \hat{\theta}^d - \theta \) with \( b = (I - W_n X_n) (\hat{\theta}^{\text{OLS}} - \theta_0) \) and \( \nu = W_n e_n \), the above optimization aims at minimizing a weighted sum of the bias and the variance of \( \hat{\theta}^d \) in an online manner. The analysis of Deshpande et al. (2018) controls bias as follows:

\[
\| b \| \leq \| I - W_n X_n \|_F \| \hat{\theta}^{\text{OLS}} - \theta_0 \|_2
\]
\[
\leq \| I - W_n X_n \|_F \| \hat{\theta}^{\text{OLS}} - \theta_0 \|_2.
\]

However, in high-dimensional this bound is vacuous. Since \( W_n X_n \in \mathbb{R}^{p \times p} \) is of rank at most \( n < p \), \( I - W_n X_n \) has eigenvalue 1 with multiplicity at least \( p - n \). Therefore, \( \| I - W_n X_n \|_F \geq p - n \rightarrow \infty \) and \( \| I - W_n X_n \|_F \geq 1 \). Thus, even a refinement of Deshpande et al. (2018) would only yield an insufficient bias bound of the type

\[
\| b \|_2 \leq \| \hat{\theta}^d - \theta_0 \|_2 \approx \sigma \sqrt{\frac{\log p}{n}},
\]

which dominates the variance component \( \text{var}(\nu) = O(1/\sqrt{n}) \). 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 \( \| \text{bias} \|_F = O(n \log p / n) \) 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 (26) or (38) 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 VAR(\( d \)) model \( p_0 = dp \)).

5.1. Confidence Intervals

For fixed \( a \in [p_0] \) and significance level \( \alpha \in (0,1) \), let

\[
J_a(\alpha) \equiv [\hat{\theta}^{\text{on}}_a - \delta(\alpha,n), \hat{\theta}^{\text{on}}_a + \delta(\alpha,n)],
\]

\[
\delta(\alpha,n) \equiv \Phi^{-1}(1 - \alpha/2) \sqrt{V_{n,a}/n},
\]

Further, note that the length of confidence interval \( J_a(\alpha) \) is of order \( O(\sigma / \sqrt{n}) \) (using Lemma H.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 (Javanmard and Montanari 2014b; Javanmard 2014) 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 iid samples.

5.2. 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 \( \theta_0 \) denotes the model parameters, either for the batched data collection setting or the VAR(\( d \)) model (which encodes the entries \( A_{ij}^{(\ell)} \) in model (5)). 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 \( z_{t-i,j} \) on a covariate \( z_t \), 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:

\[
P_a = 2 \left( 1 - \Phi \left( \frac{\sqrt{n} \hat{\theta}_{0,a}^n}{\sqrt{V_{n,a}}} \right) \right) .
\]

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

\[
R(a) = \begin{cases} 1 & \text{if } P_a \leq \alpha \quad (\text{reject } H_0) , \\ 0 & \text{otherwise } (\text{fail to reject } H_0) . \end{cases}
\]

Employing the distributional characterizations (38) or (26), 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: \( \mathbb{P}_{\theta_{0,a} = 0}(P_a \leq u) = u \) for all \( u \in [0,1] \).

5.3. 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 Equation (23) for the VAR(\( d \)) model, that is:

\[
V_n \equiv \frac{\sigma^2}{n} \sum_{\ell=1}^{K-1} \sum_{i \in k_{\ell}} (M^{(\ell)} x_i)(M^{(\ell)} x_i)^\top .
\]

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

\[
\lim_{n \to \infty} \mathbb{P} \left( \sqrt{n} (V_{n,G}^{-1/2} (\hat{\theta}_{G}^n - \theta_G) \leq u \right) = 0 ,
\]

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 \( (a_1, \ldots, a_k) \leq (b_1, \ldots, b_k) \) indicates that \( a_i \leq b_i \) for \( i \in [k] \) and \( \Phi_k(u) = \Phi(u_1) \ldots \Phi_k(u_k) \).

Much in the same way as individual inference, we can use Lemma 5.1 for simultaneous inference on a group of parameters. Concretely, let \( S_{k,\alpha} \subseteq \mathbb{R}^k \) 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 \( I_G(\alpha) \subseteq \mathbb{R}^k \) as follows

\[
I_G(\alpha) = \hat{\theta}_{G}^n + \frac{1}{\sqrt{n}} (V_{n,G})^{1/2} S_{k,\alpha} .
\]

Then, using Lemma 5.1, we conclude that \( I_G(\alpha) \) is a valid confidence region, namely \( \lim_{n \to \infty} \mathbb{P}(\theta_{0,G} \in I_G(\alpha)) = 1 - \alpha \).

5.4. 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. In addition, through numerical studies 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).

The (offline) debiasing approach has been used as a tool to address a variety of problems from estimating average treatment effect and casual inference in high-dimension to precision matrix estimation, distributed multitask learning, and hierarchical testing. It has also been used for different statistical aims such as controlling FDR in high-dimensions (Javanmard and Javadi 2019), estimation of the prediction risk (Javanmard and Montanari 2018), inference on predictions (Cai and Guo 2017) and general functions of model parameters (Javanmard and Lee 2020). 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.

Funding

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

Supplementary Materials

The supplementary materials contain discussion of debiasing with sparse inverse covariance, robustness of online debiasing to episode growth rate, iterative schemes to implement online debiasing, further numerical experiments, and in addition proof of theorems and technical lemmas.
