Accurate Inference in Adaptive Linear Models

Yash Deshpande[*] Lester Mackey[†] Vasilis Syrgkanis[‡] Matt Taddy[§]

December 25, 2017

Abstract

Estimators computed from adaptively collected data do not behave like their non-adaptive brethren. Rather, the sequential dependence of the collection policy can lead to severe distributional biases that persist even in the infinite data limit. We develop a general method – $\boldsymbol{W}$*-decorrelation* – for transforming the bias of adaptive linear regression estimators into variance. The method uses only coarse-grained information about the data collection policy and does not need access to propensity scores or exact knowledge of the policy. We bound the finite-sample bias and variance of the $\boldsymbol{W}$-estimator and develop asymptotically correct confidence intervals based on a novel martingale central limit theorem. We then demonstrate the empirical benefits of the generic $\boldsymbol{W}$-decorrelation procedure in two different adaptive data settings: the multi-armed bandit and the autoregressive time series settings.

1 Introduction

Randomized experiments have played a pivotal role in advancing our understanding in many fields of science and engineering. Throughout, we will assume that data collected is in the form of n samples $(y_i, \boldsymbol{x}_i)_{i\leq n}$. Here y_i are the outcomes and $\boldsymbol{x}_i \in \mathbb{R}^p$ is a vector of features or covariates associated with the samples. In the standard linear model, the outcomes y_i and covariates $\boldsymbol{x}_i$ are related as through a parameter β as:

$$y_i = \langle \boldsymbol{x}_i, \beta \rangle + \varepsilon_i. \tag{1}$$

In this model, the 'noise' term ε_i represents inherent variation in the sample, or the variation that is not captured in the model. Parametric models of the type (1) are a fundamental building block in many regression and classification settings. A further common, and often critical, assumption is that the covariates $\boldsymbol{x}_i$ are independent of the outcomes $(y_j)_{j\neq i}$ and the inherent variation $(\varepsilon_j)_{j\in[n]}$. This paper is motivated from experiments where the sample $(y_i, \boldsymbol{x}_i)_{i\leq n}$ is not completely randomized but rather *adaptive* chosen. By adaptive, we mean that the choice of a new data point $(y_i, \boldsymbol{x}_i)$ is guided from inferences on past data. Consider the following sequential paradigms:

1. Multi-armed bandits: This class of sequential decision making problems captures the classical 'exploration versus exploitation' tradeoff. At each time i, the experimenter chooses an 'action' $\boldsymbol{x}_i$ from a set of available actions $\mathcal{X}$ and accrues a reward $R(y_i)$ where $(y_i, \boldsymbol{x}_i)$ follow the model (1). Here the experimenter must balance the conflicting goals of learning about the underlying model (i.e., β) for better future rewards, while still accruing reward in the current time step.

2. Active learning: In such settings, acquiring labels y_i is costly, and the experimenter must learn with as few outcomes as possible. At time i, based on prior data $(y_j, \boldsymbol{x}_j)_{j\leq i-1}$ the experimenter chooses a new data point $\boldsymbol{x}_i$ to label based on its value in the learning problem.

[*] Department of Mathematics, Massachusetts Institute of Technology and Microsoft Research

[†] Microsoft Research

[‡] Microsoft Research

[§] Chicago Booth and Microsoft Research

3. Time series analysis: Here, the data points $(y_i, \boldsymbol{x}_i)$ are naturally ordered in time, with $(y_i)_{i\leq n}$ denoting a time series and the covariates $\boldsymbol{x}_i$ can include observations from the past time points.

Here, time induces a natural sequential dependence across the samples. In the first two instances, the actions or policy of the experimenter are responsible for creating such dependence. In the case of time series data, this dependence is endogenous, and a consequence of the modeling. A common feature, however, is that the choice of the design or sequence $(\boldsymbol{x}_i)_{i\leq n}$ is typically not made for inference on the model after the data collection is completed. This does not, of course, imply that accurate estimates on the parameters β cannot be made from the data. Indeed, it is often the case that the sample is informative enough to extract consistent estimators of the underlying parameters. Indeed, this is often crucial to the success of the experimenter's policy. For instance, notions such as 'regret' in sequential decision-making or the risk in active learning are intimately connected with the accurate estimation of the underlying parameters [Castro and Nowak, 2008, Audibert and Bubeck, 2009, Bubeck et al., 2012, Rusmevichientong and Tsitsiklis, 2010]. Our motivation is the natural follow-up question of accurate *ex poste* inference in the standard statistical sense:

> Can adaptive data be used to compute accurate confidence regions and p-values?

As we will see, the key challenge is that even in the simple linear model of (1), the distribution of classical estimators can differ from the predicted central limit behavior of non-adaptive designs. In this context we make the following contributions:

- **Decorrelated estimators:** We present a general method to decorrelate arbitrary estimators $\widehat{\beta}(\boldsymbol{y}, \boldsymbol{X}_n)$ constructed from the data. This construction admits a simple decomposition into a 'bias' and 'variance' term. In comparison with competing methods, like propensity weighting, our proposal requires little explicit information about the data-collection policy.

- **Bias and variance control:** Under a natural exploration condition on the data collection policy, we establish that the bias and variance can be controlled at nearly optimal levels. In the multi-armed bandit setting, we prove this under an especially weak averaged exploration condition.

- **Asymptotic normality and inference:** We establish a martingale central limit theorem under a moment stability assumption. Applied to our decorrelated estimators, this allows us to construct confidence intervals and conduct hypothesis tests in the usual fashion.

- **Validation:** We demonstrate the usefulness of the decorrelating construction in two different scenarios: multi-armed bandits (MAB) and autoregressive (AR) time series. We observe that our decorrelated estimators retain expected central limit behavior in regimes where the standard estimators do not, thereby facilitating accurate inference.

The rest of the paper is organized with our main results in Section 2, discussion of related work in Section 3, and experiments in Section 4.

2 Main results: $\boldsymbol{W}$-decorrelation and inference

We focus on the linear model and assume that the data pairs $(y_i, \boldsymbol{x}_i)$ satisfy:

$$y_i = \langle \boldsymbol{x}_i, \beta \rangle + \varepsilon_i, \tag{2}$$

where ε_i are independent and identically distributed random variables with $\mathbb{E}\{\varepsilon_i\} = 0$, $\mathbb{E}\{\varepsilon_i^2\} = \sigma^2$ and bounded third moment. We assume that the samples are ordered naturally in time and let $\{\mathcal{F}_i\}_{i\geq 0}$ denote the filtration representing increasing information in the sample. Fornally, we let data points $(y_i, \boldsymbol{x}_i)$ be adapted to this filtration, i.e. $(y_i, \boldsymbol{x}_i)$ are measurable with respect to $\mathcal{F}_j$ for all $j \geq i$.

Our goal in this paper is to use the available data to construct *ex poste* confidence intervals and p-values for individual parameters, i.e. entries of β. A natural starting point is to consider is the standard least squares estimate:

$$\widehat{\beta}_{\mathsf{OLS}} = (\boldsymbol{X}_n^{\mathsf{T}} \boldsymbol{X}_n)^{-1} \boldsymbol{X}_n^{\mathsf{T}} \boldsymbol{y}_n,$$

where $\boldsymbol{X}_n = [\boldsymbol{x}_1^{\mathsf{T}}, \dots \boldsymbol{x}_n^{\mathsf{T}}] \in \mathbb{R}^{n \times p}$ and $\boldsymbol{y}_n = [y_1, \dots y_n] \in \mathbb{R}^n$. When the data collection not adaptive, classical results imply that the standard least squares estimate $\widehat{\beta}_{\mathsf{OLS}}$ is distributed asymptotically as $\mathsf{N}(\beta, \sigma^2(\boldsymbol{X}_n^{\mathsf{T}} \boldsymbol{X}_n)^{-1})$, where $\mathsf{N}(\mu, \Sigma)$ denotes the Gaussian distribution with mean μ and covariance Σ. Lai and Wei [1982] extend these results to the current scenario:

Theorem 1 (Theorems 1, 3 [Lai and Wei, 1982])**.** *Let $\lambda_{\min}(n)$ ($\lambda_{\max}(n)$) denote the minimum (resp. maximum) eigenvalue of $\boldsymbol{X}_n^{\mathsf{T}} \boldsymbol{X}_n$. Under the model* (2)*, assume that* (*i*) *ε_i have finite third moment and* (*ii*) *almost surely, $\lambda_{\min}(n) \to \infty$ with $\lambda_{\min} = \Omega(\log \lambda_{\max})$ and* (*iii*) *$\log \lambda_{\max} = o(n)$. Then the following limits hold almost surely:*

$$\left\| \widehat{\beta}_{\mathsf{OLS}} - \beta \right\|_2^2 \leq C \frac{\sigma^2 p \log \lambda_{\max}}{\lambda_{\min}}$$
$$\left| \frac{1}{n\sigma^2} \left\| \boldsymbol{y}_n - \boldsymbol{X}_n \widehat{\beta}_{\mathsf{OLS}} \right\|_2^2 - 1 \right| \leq C(p) \frac{1 + \log \lambda_{\max}}{n}.$$

Further assume the following stability condition: there exists a sequence of non random matrices $\boldsymbol{A}_n$ such that (*iii*) *$\boldsymbol{A}_n^{-1} (\boldsymbol{x_{n-i}}^{\mathsf{T}} \boldsymbol{X}_n)^{1/2} \to \mathrm{I}_p$ and* (*iv*) *$\max_i \left\| \boldsymbol{A}_n^{-1} \boldsymbol{x}_i \right\| \to 0$ in probability. Then,*

$$(\boldsymbol{X}_n^{\mathsf{T}} \boldsymbol{X}_n)^{1/2} (\widehat{\beta}_{\mathsf{OLS}} - \beta) \overset{\mathrm{d}}{\Rightarrow} \mathsf{N}(0, \sigma^2 \mathrm{I}_p).$$

At first blush, this allows to construct confidence regions in the usual way. More precisely, the result implies that $\widehat{\sigma}^2 = \|\boldsymbol{y_n} - \boldsymbol{X}_n \widehat{\beta}_{\mathsf{OLS}}\|_2^2 / n$ is a consistent estimate of the noise variance. Therefore, the interval $[\widehat{\beta}_{\mathsf{OLS},1} - 1.96 \widehat{\sigma} (\boldsymbol{X}_n^{\mathsf{T}} \boldsymbol{X}_n)_{11}^{-1}, \widehat{\beta}_{\mathsf{OLS},1} + 1.96 \widehat{\sigma} (\boldsymbol{X}_n^{\mathsf{T}} \boldsymbol{X}_n)_{11}^{-1}]$ is a 95% two-sided confidence interval for the first coordinate β_1. Indeed, this result is sufficient for a variety of scenarios with weak dependence across samples, such as when the $(y_i, \boldsymbol{x}_i)$ form a Markov chain that mixes rapidly. However, while the assumptions for consistency are minimal, the additional stability assumption required for asymptotic normality poses some challenges. In particular:

1. The stability condition can provably fail to hold for scenarios where the dependence across samples is non-negligible. This is not a weakness of Theorem 1: indeed, in such cases, the central limit theorem can fail to hold for the OLS estimator [Lai and Wei, 1982, Lai and Siegmund, 1983].

2. The rate of convergence to the asymptotic central limit theorem depends on the *quantitative rate* of the stability condition. In other words, variability in the inverse covariance $\boldsymbol{X}_n^{\mathsf{T}} \boldsymbol{X}_n$ can cause deviations from normality of OLS estimator [Dvoretzky, 1972]. In finite samples, this can manifest itself in the bias of the OLS estimator as well as in higher moments.

An example of this phenomenon is the standard multi-armed bandit problem [Lai and Robbins, 1985]. At each time point $i \leq n$, the experimenter (or data collecting policy) chooses an arm $k \in \{1, 2, \dots, p\}$ and observes a reward y_i with mean β_k. With $\beta \in \mathbb{R}^p$ denoting the mean rewards, this falls within the scope of model 2, where the vectors $\boldsymbol{x}_i$ takes the value $\boldsymbol{e}_k$ (the k^{th} basis vector), if the k^{th} arm or option is chosen at time i.[1] Other stochastic bandit problems with covariates such as contextual or linear bandits [Rusmevichientong and Tsitsiklis, 2010, Li et al., 2010, Deshpande and Montanari, 2012] can also be incorporated fairly naturally into our framework. For the purposes of this paper, however, we restrict ourselves to the simple case of multi-armed bandits without covariates. In this setting, ordinary least squares estimates correspond to

[1] Strictly speaking, the model 2 assumes that the errors have the same variance, which need not be true for the multi-armed bandit as discussed. We focus on the homoscedastic case where the errors have the same variance in this paper.

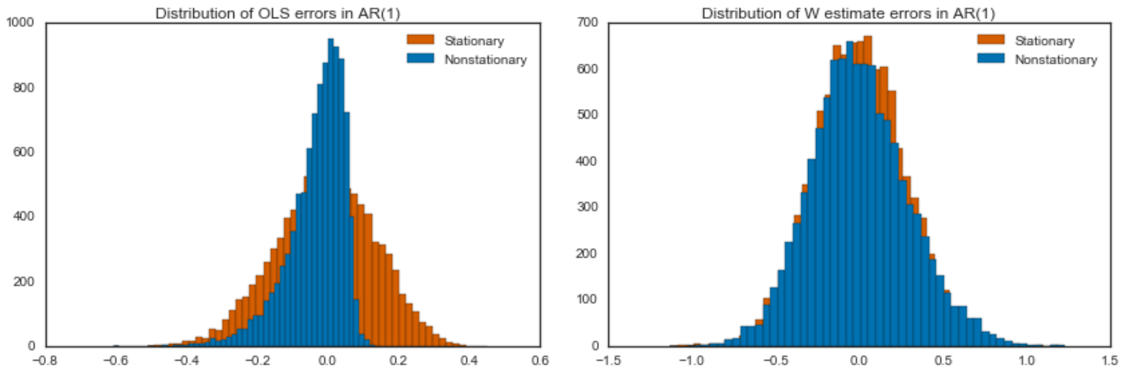

Figure 1: The distribution of errors for (left) the OLS estimator for stationary and (nearly) nonstationary AR(1) time series and (right) error distribution for both models after decorrelation. $n = 50$, $\varepsilon_i \sim \mathsf{N}(0, 1)$.

computing sample means for each arm. The stability condition of Theorem 1 requires that $N_k(a)$, the number of times a specific arm $k \in [p]$ is sampled is asymptotically deterministic as n grows large. This is true for certain regret-optimal algorithms [Russo, 2016, Garivier and Cappé, 2011]. Indeed, for such algorithms, as the sample size n grows large, the suboptimal arm is sampled $N_k(n) \sim C_k(\beta) \log n$ for a constant $C_k(\beta)$ that depends on β and the distribution of noise ε_i. However, in finite samples, the dependence on $C_k(\beta)$ and the slow convergence rate of $1/\sqrt{\log n}$ lead to significant deviation from the expected central limit behavior.

Villar et al. [2015] studied a variety of multi-armed bandit algorithms in the context of clinical trials. They empirically demonstrate that sample mean estimates from data collected using many standard multi-armed bandit algorithms are biased. Recently, Nie et al. [2017] proved that this bias is negative for Thompson sampling and UCB. The presence of bias in sample means demonstrates that standard methods for inference, as advocated by Theorem 1, can be misleading when the same data is now used for inference. As a pertinent example, testing the hypotheses "the mean reward of arm 1 exceeds that of 2" based on classical theory can be significantly affected by adaptive data collection.

The papers [Villar et al., 2015, Nie et al., 2017] focus on the finite sample effect of the data collection policy on the bias and suggest methods to reduce the bias. It is not hard to find examples where higher moments or tails of the distribution can be influenced by the data collecting policy. A simple, yet striking, example is the standard autoregressive model (AR) for time series data. In its simplest form, the AR model has one covariate, i.e. $p = 1$ with $\boldsymbol{x}_i = y_{i-1}$. In this case:

$$y_i = \beta y_{i-1} + \varepsilon_i. \tag{3}$$

Here the least squares estimate is given by $\widehat{\beta}_{\mathsf{OLS}} = \sum_{i \le n-1} y_{i+1} y_i / \sum_{i \le n-1} y_{i-1}^2$. When $|\beta|$ is bounded away from 1, the series is asymptotically stationary and the OLS estimate has Gaussian tails. On the other hand, when $\beta - 1$ is on the order of $1/n$ the limiting distribution of the least squares estimate is non-Gaussian and dependent on the gap $\beta - 1$ (see Chan and Wei [1987] for a description in terms of the standard Weiner process). An histogram for the OLS errors in two cases: (i) stationary with $\beta = 0.02$ and (ii) (nearly) nonstationary with $\beta = 0.9$ is shown on the left in Figure 1 where the large β example case is clearly non-Gaussian.

On the other hand, *using the same data* our decorrelating procedure is able to obtain estimates admitting Gaussian limit distributions, as evidenced in the right panel of Figure 1. We show a similar phenomenon in the MAB setting where our decorrelating procedure corrects for the unstable behavior of the OLS estimator (see Section 4 for details on the empirics). Delegating discussion of further related work to 3, we now describe this procedure and its motivation.

2.1 Removing the effects of adaptivity: $\boldsymbol{W}$-decorrelation

We propose to decorrelate the OLS estimator by constructing:

$$\widehat{\beta}^d = \widehat{\beta}_{\mathsf{OLS}} + \boldsymbol{W}_n(y - \boldsymbol{X}_n\widehat{\beta}_{\mathsf{OLS}}), \tag{4}$$

for a specific choice of a 'decorrelating' or 'whitening' matrix $\boldsymbol{W}_n \in \mathbb{R}^{p\times n}$. This is inspired from analogous constructions for high-dimensional linear regression by Zhang and Zhang [2014], Javanmard and Montanari [2014b,a], Van de Geer et al. [2014]. As we will see, these ideas can be useful also in the present regime where we keep p fixed and $n \gtrsim p$. By rearranging:

$$\begin{aligned}\widehat{\beta}^d - \beta &= (\mathrm{I}_p - \boldsymbol{W}_n\boldsymbol{X}_n)(\widehat{\beta}_{\mathsf{OLS}} - \beta) + \boldsymbol{W}_n\boldsymbol{\varepsilon}_n \\ &\equiv \mathsf{b} + \mathsf{v}. \end{aligned} \tag{5}$$

We interpret b as a 'bias' and v as a 'variance'. This is aided by the following critical constraint on the construction of the whitening matrix $\boldsymbol{W}_n$:

Definition 1 (Well-adaptedness of $\boldsymbol{W}_n$)**.** *Without loss of generality, we assume that ε_i are adapted to $\mathcal{F}_i$. Let $\mathcal{G}_i \subset \mathcal{F}_i$ be a filtration such that $\boldsymbol{x}_i$ are adapted w.r.t. $\mathcal{G}_i$ and ε_i is independent of $\mathcal{G}_i$. We say that $\boldsymbol{W_n}$ is well-adapted if the columns of $\boldsymbol{W}_n$ are adapted to $\mathcal{G}_i$, i.e. the i^{th} column $\boldsymbol{w}_i$ is measurable with respect to $\mathcal{G}_i$.*

With this in hand, we have the following simple lemma.

Lemma 2. *Assuming $\boldsymbol{W}_n$ is well-adapted, as in Definition 1:*

$$\begin{aligned}\big\|\beta - \mathbb{E}\{\widehat{\beta}^d\}\big\| &\le \mathbb{E}\big\{\|\mathrm{I}_p - \boldsymbol{W}_n\boldsymbol{X}_n\|_2\|\widehat{\beta}_{\mathsf{OLS}} - \beta\|_2\big\}, \\ \mathrm{Var}(\mathsf{v}) &= \sigma^2\mathbb{E}\{\boldsymbol{W}_n\boldsymbol{W}_n^{\mathsf{T}}\}.\end{aligned}$$

A concrete proposal is to trade-off the bias, controlled by the size of $\mathrm{I}_p - \boldsymbol{W}_n\boldsymbol{X}_n$, with the the variance which appears through $\boldsymbol{W}_n\boldsymbol{W}_n^{\mathsf{T}}$. This leads to the following optimization problem:

$$\boldsymbol{W}_n = \arg\min_{\boldsymbol{W}} \big\|\mathrm{I}_p - \boldsymbol{W}\boldsymbol{X}_n\big\|_F^2 + \lambda\mathsf{Tr}\big(\boldsymbol{W}\boldsymbol{W}^{\mathsf{T}}\big). \tag{6}$$

Solving the above in closed form yields ridge estimators for β, and by continuity, also the standard least squares estimator. Departing from Zhang and Zhang [2014], Javanmard and Montanari [2014a], we solve the above in an *online* fashion in order to obtain a well-adapted $\boldsymbol{W}_n$. We define, recursively $\boldsymbol{W}_n = [\boldsymbol{W}_{n-1}\boldsymbol{w}_n]$:

$$\boldsymbol{w}_n = \arg\min_{\boldsymbol{w}\in\mathbb{R}^p} \|\mathrm{I} - \boldsymbol{W}_{n-1}\boldsymbol{X}_{n-1} - \boldsymbol{w}\boldsymbol{x}_n^T\|_F^2 + \lambda\|\boldsymbol{w}\|^2.$$

As in the case of the offline optimization, we may obtain closed form formulae for the columns $\boldsymbol{w}_i$ (see Algorithm 1). The method as specified requires $O(np^2)$ additional computational overhead, which is typically minimal compared to computing $\widehat{\beta}_{\mathsf{OLS}}$, or even a regularized version like the ridge or lasso estimate. We refer to $\widehat{\beta}^d$ as a $\boldsymbol{W}$*-estimate* or a $\boldsymbol{W}$*-decorrelated estimate.*

2.2 Implicit stochastic gradient descent in reverse

While we motivated $\boldsymbol{W}$-decorrelation decorrelation as an online procedure for optimizing the bias-variance tradeoff objective (6), it holds a dual interpretation as implicit stochastic gradient descent (SGD) [see, e.g., Kulis and Bartlett, 2010] (also known as incremental proximal minimization [Bertsekas, 2011] or the normalized least mean squares filter [Nagumo and Noda, 1967] in this context) with step-size λ applied to the least-squares objective, $\frac{1}{n}\sum_{i=1}^n(y_i - \langle\beta, \boldsymbol{x}_i\rangle)^2$. Importantly, to obtain the well-adapted form of our updates, one must apply implicit SGD in reverse, starting with the final observation $(\boldsymbol{x}_n, y_n)$ and ending with the initial observation $(\boldsymbol{x}_1, y_1)$; this recipe yields the parameter updates $\widehat{\beta}_0 = \widehat{\beta}_{\mathsf{OLS}}$ and

$$\begin{aligned}\widehat{\beta}_{i+1} &= \widehat{\beta}_i + \boldsymbol{x}_{n-i}(y_{n-i} - \langle \boldsymbol{x}_{n-i}, \widehat{\beta}_{i+1}\rangle)/\lambda \\ &= (\mathrm{I}_p + \boldsymbol{x}_{n-i}\boldsymbol{x}_{n-i}^T/\lambda)^{-1}(\widehat{\beta}_i + y_{n-i}\boldsymbol{x}_{n-i}/\lambda) \\ &= (\mathrm{I}_p - \boldsymbol{x}_{n-i}\boldsymbol{x}_{n-i}^T/(\lambda + \|\boldsymbol{x}_{n-i}\|^2))\widehat{\beta}_i + y_{n-i}\boldsymbol{x}_{n-i}/(\lambda + \|\boldsymbol{x}_{n-i}\|^2).\end{aligned}$$

Unrolling the recursion, we obtain the equivalent form $\widehat{\beta}_n = \widehat{\beta}_{\mathsf{OLS}} + \sum_{i=1}^n y_i \boldsymbol{w}_i$ for

$$\boldsymbol{w}_i = (\mathrm{I}_p - \boldsymbol{x}_1\boldsymbol{x}_1^T/(\lambda + \|\boldsymbol{x}_1\|^2))(\mathrm{I}_p - \boldsymbol{x}_2\boldsymbol{x}_2^T/(\lambda + \|\boldsymbol{x}_2\|^2)) \cdots (\mathrm{I}_p - \boldsymbol{x}_{i-1}\boldsymbol{x}_{i-1}^T/(\lambda + \|\boldsymbol{x}_{i-1}\|^2))\boldsymbol{x}_i/(\lambda + \|x_i\|^2),$$

which precisely matches the updates given in Algorithm 1.

2.3 Bias and variance

We now examine the bias and variance control for $\widehat{\beta}^d$. We first begin with a general bound for the variance:

Theorem 3 (Variance control)**.** *For any $\lambda \geq 1$ set non-adaptively, we have that*

$$\mathsf{Tr}\{\mathrm{Var}(\mathsf{v})\} \leq \frac{\sigma^2}{\lambda}(p - \mathbb{E}\{\|\mathrm{I}_p - \boldsymbol{W}_n\boldsymbol{X}_n\|_F^2\}).$$

In particular, $\mathsf{Tr}\{\mathrm{Var}(\mathsf{v})\} \leq \sigma^2 p/\lambda$. Further, if $\|\boldsymbol{x}_i\|_2^2 \leq C$ for all i:

$$\mathsf{Tr}\{\mathrm{Var}(\mathsf{v})\} \asymp \frac{\sigma^2}{\lambda}(p - \mathbb{E}\{\|\mathrm{I}_p - \boldsymbol{W}_n\boldsymbol{X}_n\|_F^2\}).$$

This theorem suggests that one must set λ as large as possible to minimize the variance. While this is accurate, one must take into account the bias of $\widehat{\beta}^d$ and its dependence on the regularization λ. Indeed, for large regularization, one would expect that $\mathrm{I}_p - \boldsymbol{W}_n\boldsymbol{X}_n \approx \mathrm{I}_p$, which would not help control the bias. In general, One would hope to set λ, thereby determining $\widehat{\beta}^d$, at a level where its bias is negligible in comparison to the variance. The following theorem formalizes this:

Theorem 4 (Variance dominates MSE)**.** *Recall that the matrix $\boldsymbol{W}_n$ is a function of λ. Suppose that there exists a deterministic sequence $\lambda(n)$ such that under the collection policy:*

$$\mathbb{E}\{\|\mathrm{I}_p - \boldsymbol{W}_n\boldsymbol{X}_n\|_2^2\} = o(1/\log n), \tag{7}$$

$$\mathbb{P}\{\lambda_{\min}(\boldsymbol{X}_n\boldsymbol{X}_n^{\mathsf{T}}) \leq \lambda(n)\} \leq 1/n, \tag{8}$$

Then we have

$$\frac{\mathbb{E}\{\|\mathsf{b}\|^2\}}{\mathsf{Tr}\{\mathrm{Var}(\mathsf{v})\}} = o(1).$$

The conditions of Theorem 4, in particular the bias condition on $\mathrm{I}_p - \boldsymbol{W}_n\boldsymbol{X}_n$ are quite general. In the following proposition, we verify some sufficient conditions under which the premise of Theorem 4 hold.

Proposition 5. *Either of the following conditions suffices for the requirements of Theorem 4.*

1. *The data collection policy satisfies for some sequence $\mu_n(i)$ and for all $\lambda \geq 1$:*

$$\mathbb{E}\left\{\frac{\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}}}{\lambda + \|\boldsymbol{x}_n\|^2}\middle|\mathcal{G}_{i-1}\right\} \succcurlyeq \frac{\mu_n(i)}{\lambda}\mathrm{I}_p, \tag{9}$$

$$\sum_i \mu_n(i) \equiv n\bar{\mu}_n \geq K\sqrt{n}, \tag{10}$$

for a large enough constant K. The choice here is $\lambda(n) \asymp n\bar{\mu}_n/\log n$.

Algorithm 1: $\boldsymbol{W}$-Decorrelation Method

```
    Input: $n$ samples $(y_i, \mathrm{x}_i)_{i\le n}$, unit vector $\boldsymbol{v} \in \mathbb{R}^p$, confidence level $\alpha \in (0,1)$, noise estimate $\hat{\sigma}^2$.
    Compute: $\widehat{\beta}_{\mathsf{OLS}} = (\boldsymbol{X}_n^{\mathsf{T}}\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n\boldsymbol{y}_n$.
    Setting $\boldsymbol{W}_0 = 0$, compute $\boldsymbol{W}_i = [\boldsymbol{W}_{i-1}\boldsymbol{w}_i]$ with $\boldsymbol{w}_i = (\mathrm{I}_p - \boldsymbol{W}_{i-1}\boldsymbol{X}_i^{\mathsf{T}})\boldsymbol{x}_i/(\lambda + \|\boldsymbol{x}_i\|^2)$.
    Compute $\widehat{\beta}^d = \widehat{\beta}_{\mathsf{OLS}} + \boldsymbol{W}_n(y - \boldsymbol{X}_n\widehat{\beta}_{\mathsf{OLS}})$ and $\hat{\sigma}(\boldsymbol{v}) = \hat{\sigma}\langle \boldsymbol{v}, \boldsymbol{W}_n\boldsymbol{W}_n^{\mathsf{T}}\boldsymbol{v}\rangle^{1/2}$
    Output: interval $I(\boldsymbol{v}, \alpha) = [\langle \boldsymbol{v}, \widehat{\beta}^d\rangle - \hat{\sigma}(\boldsymbol{v})\Phi^{-1}(-\alpha), \langle \boldsymbol{v}, \widehat{\beta}^d\rangle + \hat{\sigma}(\boldsymbol{v})\Phi^{-1}(-\alpha)]$.
```

2. The matrices $(\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}})_{i\le n}$ commute and $\lambda(n) \asymp \lambda_{\min}/\log n$ with probability at least $1/n$.

It is useful to consider the intuition for the sufficient conditions given in Proposition 5. By Lemma 2, note that the bias is controlled by $\|\mathrm{I} - \boldsymbol{W}_n\boldsymbol{X}_n\|$, which increases with λ. Consider a case in which the samples $\boldsymbol{x}_i$ lie in a strict subspace of $\mathbb{R}^p$. In this case, controlling the bias uniformly over $\beta_0 \in \mathbb{R}^p$ is now impossible regardless of the choice of $\boldsymbol{W}_n$. As an example, in a multi-armed bandit problem, if the policy does not sample a specific arm, there is no information available about the reward distribution of that arm. Intuitively, the data collection policy should ‘explore’ most of the space, which is formalized in Proposition 5. For multi-armed bandits, one can interpret this assumption as a guarantee of ‘minimum exploration’. Indeed, policies such as epsilon-greedy and Thompson sampling [Thompson, 1933] satisfy it with appropriate choices of $\mu_n(i)$.

Given sufficient exploration, Proposition 5 recommends a reasonable value to set for the regularization parameter. In particular setting λ to a value such that $\lambda \ll \lambda_{\min}$ occurs with high probability suffices to ensure that the $\boldsymbol{W}$-decorrelated estimate is approximately unbiased. Correspondingly, the MSE (or equivalently variance) of the $\boldsymbol{W}$-decorrelated estimate need not be smaller than that of the original OLS estimate. Indeed the variance scales as $1/\lambda$, which exceeds with high probability the $1/\lambda_{\min}$ scaling for the MSE. This is the cost paid for removing most of the bias in the OLS estimate. Before we move to the specific inference results, note that the procedure requires only access to high probability lower bounds on $\lambda_{\min}$, which intuitively quantifies the exploration of the data collection policy. In comparison with methods such as propensity score weighting or conditional likelihood optimization, this represents rather coarse information about the data collection process. In particular, given access to propensity scores, one can simulate the process to extract appropriate values for the regularization $\lambda(n)$. This is the approach we take in the experiments of Section 4.

Propensity score methods are also ineffective when data collection policies make adaptive decisions that are deterministic given the history. A pertinent example is that of UCB algorithms for bandits, which make history-dependent deterministic decisions to pull arms. Finally, we note that the cost of increased variance can be mitigated by the use of information on the data collection policy, as demonstrated in [Nie et al., 2017, Dimakopoulou et al., 2017]. It is an interesting open problem to develop a middle ground between these approaches.

2.4 A central limit theorem and confidence intervals

Our final result is a simple central limit theorem that provides an alternative to the stability condition of Theorem 1 and standard martingale central limit theorems. We state it for martingales of the form of $\sum_i \boldsymbol{w}_i\varepsilon_i$, as required, but a form for general martingales also holds true. We make the following crucial moment stability assumption:

Assumption 1. *For $a = 1, 2$, and positive integer m*

$$\sup_{\|\boldsymbol{t}\|\le 1} \frac{1}{n^{a/2}} \sum_{i=1}^{n} \mathbb{E}\big\{\big|\mathbb{E}\{\varepsilon_i^a \langle \boldsymbol{w}_i, \boldsymbol{t}\rangle^m | \mathcal{F}_{i-1}\} - \mathbb{E}\{\varepsilon_i^a | \mathcal{F}_{i-1}\}\mathbb{E}\{\langle \boldsymbol{w}_i, \boldsymbol{t}\rangle^m | \mathcal{F}_{i-1}\}\big|\big\} = o_{m,a}(1).$$

Theorem 6 (Martingale CLT)**.** *Let ε_i be a $\mathcal{F}_i$-adapted sequence, and $\boldsymbol{w}_i$ be $\mathcal{F}_i$-predictable. In addition to the stability assumption 1, suppose that (i) $\mathbb{E}\{\varepsilon_i | \mathcal{F}_{i-1}\} = 0$, (ii) $\mathbb{E}\{\varepsilon_i^2 | \mathcal{F}_{i-1}\} = \sigma^2$, that (iii) ε_i are* *subgaussian, i.e.* $\mathbb{E}\{e^{\eta\varepsilon_i}|\mathcal{F}_{i-1}\} \le e^{\eta^2 B^2/2}$ *almost surely, for a constant B and (iv) the predictable sequence* $\boldsymbol{w}_i$ *is bounded almost surely. Then* $(\sum_i \boldsymbol{w}_i\boldsymbol{w}_i^\mathsf{T})^{-1/2}\sum_i \boldsymbol{w}_i\varepsilon_i \overset{\mathrm{d}}{\Rightarrow} \mathsf{N}(0,\sigma^2\mathrm{I}_p)$*. In particular, for any bounded, continuous function* $\varphi : \mathbb{R}^p \to \mathbb{R}$

$$\lim_{n\to\infty} \mathbb{E}\Big\{\varphi\Big(\frac{1}{\sqrt{n}}\sum_i \boldsymbol{w}_i\varepsilon_i\Big)\Big\} - \mathbb{E}\Big\{\varphi\Big(\Big[\frac{\sigma^2}{n}\sum_i \boldsymbol{w}_i\boldsymbol{w}_i^\mathsf{T}\Big]^{1/2}\boldsymbol{\xi}\Big)\Big\} = 0,$$

where $\boldsymbol{\xi} \sim \mathsf{N}(0,\mathrm{I}_p)$ *is independent of* $\sum_i \boldsymbol{w}_i\boldsymbol{w}_i^\mathsf{T}$*.*

The assumptions (*iii*) and (*iv*) are made for simplicity of the proof, which uses the usual Fourier-analytic approach to prove the central limit theorem [Billingsley, 2008]. These can likely be relaxed significantly to standard third moment assumptions as in a Lyapunov central limit theorem. Assumption 1 is an alternate form of stability. The essence of this assumption is that it controls the dependence of the conditional covariance of $\sum_i \boldsymbol{w}_i\varepsilon_i$ on the first two conditional moments of the martingale increments ε_i. In words, it states that conditioning on the conditional covariance $\sum_i \boldsymbol{w}_i\boldsymbol{w}_i^\mathsf{T}$ does not change the first two moments of the random variables ε_i by much. In particular, this holds given a quantitative version of the stability condition of Lai and Wei [1982], Dvoretzky [1972]. For instance, if a non-random sequence $\boldsymbol{A}_n$ satisfies $\boldsymbol{A}_n^{-1}\sum_i \boldsymbol{w}_i\boldsymbol{w}_i^\mathsf{T} - \mathrm{I}_p = o(n^{-1/2})$, then Assumption 1 holds.

With a central limit theorem in hand, one can now assign confidence intervals in the standard fashion, based on the assumption that the bias is negligible. For instance, it is not hard to show the following result on two-sided confidence intervals.

Proposition 7. *Fix any $\alpha > 0$. Suppose that the data collection process satisfies the assumptions of Theorem 4. Setting $\lambda = \lambda(n)$ as in Theorem 4 we have, for any fixed coordinate $a \in [p]$*

$$\limsup_{n\to\infty} \mathbb{P}\big\{\beta_a \notin [\widehat{\beta}^d_a - \widehat{\sigma}\sqrt{Q_{aa}}\Phi^{-1}(1-\alpha/2), \widehat{\beta}^d_a + \widehat{\sigma}\sqrt{Q_{aa}}\Phi^{-1}(1-\alpha/2)\big\} \le \alpha.$$

Here $\widehat{\sigma}^2$ *is a consistent estimator for the variance, as in Theorem 1, and* $\boldsymbol{Q} = \boldsymbol{W_n}\boldsymbol{W_n}^\mathsf{T}$*.*

3 Related work

There is a long line of work in statistics literature extending the results of Lai and Wei [1982] to a variety of different models and estimators [Wei, 1985, Lai, 1994, Chen et al., 1999, Heyde, 2008]. These results are in a similar flavor as Theorem 1 in that they demonstrate (*i*) consistency for natural estimators such as least squares or (quasi-)likelihood optimizers under weak assumptions and (*ii*) appropriate central limit theorems under an additional stability assumption similar to that of Theorem 1. In the same vein, there is much work in statistics and econometrics on non-stationary time series, including testing for unit roots in autoregressive processes and inference on (near) unit parameters (see [Shumway and Stoffer, 2006, Enders, 2008, Dickey and Fuller, 1979, Phillips and Perron, 1988] and references therein). Methods from this line of work are focused on specific time series models and do not extend to the general setup we consider. We instead focus on literature from sequential decision-making, offline learning, policy learning and causal inference that more closely resembles our work in terms of goals, techniques and scope of applicability.

The seminal work of Lai and Robbins [Robbins, 1985, Lai and Robbins, 1985] has spurred a vast literature in statistics and computer science on multi-armed bandit problems and sequential experiments that propose allocation algorithms based on confidence bounds (see Bubeck et al. [2012] and references therein). A variety of confidence bounds and corresponding rules have been proposed [Auer, 2002, Dani et al., 2008, Rusmevichientong and Tsitsiklis, 2010, Abbasi-Yadkori et al., 2011, Jamieson et al., 2014] based largely on tools in martingale theory of concentration and law of iterated logarithm. Such results can certainly be used to compute valid confidence intervals. However such bounds are conservative for a few reasons. First, they do not explicitly account for bias in OLS estimates and, correspondingly, must be wider to account for this. Second, obtaining optimal constants in the concentration inequalities can require sophisticated tools even for non-adaptive data [Ledoux, 1996, 2005]. This is evidenced in all of our experiments which show that concentration inequalities yield valid, but overly conservative intervals.

A closely-related line of work is that of learning from logged data [Li et al., 2011, Dudík et al., 2011, Swaminathan and Joachims, 2015] and policy learning [Athey and Wager, 2017, Kallus, 2017]. These focus on efficiently estimating the reward (or value) of a certain test policy using data collected from a different policy. For linear models, this reduces to accurately estimating the prediction error which is directly related to the estimation error on the parameters β. While our work shares some features, particularly resembling the techniques of [Kallus, 2017], we focus on unbiased estimation of the parameters and obtaining accurate confidence intervals for linear functions of the parameters. Some of the work on learning from logged data also builds on propensity scores and their estimation [Imbens, 2000, Lunceford and Davidian, 2004], which are well-studied in econometrics and causal inference. In particular, our techniques also closely resemble those of Athey et al. [2016], Wang and Zubizarreta [2017] which propose balancing covariates or residuals for causal inference in the potential outcomes framework.

As covered in the previous section, Villar et al. [2015] empirically demonstrate the presence of bias for a number of multi-armed bandit algorithms. Recent work by Dimakopoulou et al. [2017] also shows a similar effect in contextual bandits. Along with a result on the sign of the bias, Nie et al. [2017] also propose conditional likelihood optimization methods to estimate parameters of the linear model. Through the lens of selective inference, they also propose methods to randomize the data collection process that simultaneously lower bias and reduce the MSE. Their techniques rely on considerable information about (and control over) the data generating process, in particular the probabilities of choosing a specific action at each point in the data selection. This can be viewed as lying on the opposite end of the spectrum from our work, which attempts to use only the data at hand, along with coarse aggregate information on the exploration inherent in the data generating process.

4 Experiments

In this section we empirically validate the decorrelated estimators introduced in the previous section in two scenarios that involve sequential dependence in covariates. Continuing from Section 2, our first scenario is a simple experiment of multi-armed bandits with gaussian rewards. The second scenario is that of autoregressive time series data. In these cases, we compare the coverage obtained by and typical widths for confidence intervals for individual parameters that we obtain from the decorrelated estimators.

4.1 Multi-armed bandits

The stochastic multi-armed bandit problem is a sequential decision making problem in which, on each time step i, a decision maker carries out one of a finite set of actions (formalized as $\boldsymbol{x_i} = \boldsymbol{e_a}$ for $a \in \{1, \dots, p\}$) and receives a stochastic reward $y_i = \langle \boldsymbol{x}_i, \beta \rangle + \varepsilon_i = \beta_k + \sigma \varepsilon_i$ related to the selected action. The rewards received then inform the subsequent actions taken. Villar et al. [2015] studied this problem in the context of patient allocation in clinical trials, where each datapoint represents a patient in a trial, each action corresponds to a treatment that can be administered, and each y_i represents an outcome following treatment. To demonstrate the utility of our decorrelated W estimator in this setting, we reproduce a modified form of their simulation.

Specifically, we sequentially assign one of $p = 2$ treatments to each of $n = 444$ patients using one of three policies (i) an ε-greedy policy (called ECB or Epsilon Current Belief), (*ii*) a practical UCB strategy based on the law of iterated logarithm (UCB) Jamieson et al. [2014] and (iii) Thompson sampling Thompson [1933]. The ECB and TS sampling strategies are Bayesian. They place an independent Gaussian prior (with mean $\mu_0 = 0.5$ and variance $\sigma_0^2 = 0.5$) on each unknown mean outcome parameter β_k and form an updated posterior belief concerning β following each treatment administration $\boldsymbol{x_i}$ and observation y_i. For ECB, the treatment administered to patient i is, with probability $1 - \varepsilon = .9$, the treatment with the largest posterior mean; with probability $1 - \varepsilon$, a uniformly random treatment is administered instead, to ensure sufficient exploration of all treatments. Note that this strategy satisfies condition 9 with $\mu_n(i) = \varepsilon/p$. For TS, at each patient i, a sample $\widehat{\beta}$ of the mean treatment effect is drawn from the posterior belief. The treatment assigned to patient is the one maximizing the sampled mean treatment, i.e. $a_*(i) = \arg\max_{a\in[p]} \widehat{\beta}_a$. In UCB, the algorithm maintains a score for each arm $a \in [p]$ that is a combination of the mean reward that the arm achieves and the empirical uncertainty of the reward. For each patient i, the UCB algorithm chooses the arm maximizing this score, and updates the score according to a fixed rule. For details on the specific implementation, we refer the reader to Jamieson et al. [2014].

In this regime, with just 2 treatments, we can expect a purely random policy to faithfully provide valid confidence intervals using the classical theory. The setup here is chosen to show how the adaptivity in data collection causes the estimators' behavior to dramatically differ from classical theory predicted by Theorem 1 Lai and Wei [1982].

We repeat this simulation 10000 times with $\sigma^2 = 0.5$. From each trial simulation, we estimate the parameters β using both OLS and our decorrelated $\boldsymbol{W}$-estimator with $\lambda = \hat{\lambda}_{10\%,\pi}$ which is the 10th percentile of $\lambda_{\min}(n)$ achieved by the policy $\pi \in \{\mathrm{ECB}, \mathrm{UCB}, \mathrm{TS}\}$. This choice is guided by Corollary 4. We compare the quality of $\boldsymbol{W}$-decorrelated estimator confidence regions, OLS Gaussian confidence regions ('OLS_gsn'), and the OLS multi-armed bandit concentration inequality regions ('OLS_conc') proposed in [Abbasi-Yadkori et al., 2011, Sec. 4]. Figure 2 (left column) shows that the Gaussian OLS lower tail regions typically overestimate coverage, while upper tail regions underestimate them. This is consistent with the observation that the sample means are biased negatively Nie et al. [2017]. The concentration OLS tail bounds are all very conservative, producing nearly 100% coverage, irrespective of the nominal level. Meanwhile, the decorrelated intervals provide faithful empirical coverage for all nominal levels for every scenario except for a few cases in Thompson sampling.

Figure 2 (right column) shows the quantile-quantile plots of OLS and $\boldsymbol{W}$-estimator errors for each parameter β_a. As in the AR experiment of the next section, the distribution of OLS errors is distinctly non-Gaussian with considerable excess kurtosis for every policy. Conversely, for the $\boldsymbol{W}$ estimator the excess kurtosis is vastly reduced for every policy. Indeed, it is nearly 0 for ECB and UCB. For Thompson sampling, the excess kurtosis is reduced by a factor of at least 4 compared to the initial values.

Figure 3 shows the mean width computed by the classical OLS, OLS concentration bounds and $\boldsymbol{W}$-decorrelation, for both arms. While $\boldsymbol{W}$-decorrelation has, on average, wider confidence intervals, they compare favorably with those provided by the concentration inequalities, particularly in the moderate confidence regimes. Moreover, the variation in mean widths is very small, as compared with that of the OLS-based methods. This shows that the stability condition assumed in Theorem 1 fails to hold for the standard methods. It would be interesting to show that the stability instead holds for $\boldsymbol{W}$-decorrelation under general conditions.

4.2 Autoregressive time series

In this section we $\boldsymbol{x_i} = [y_{i-p}, \dots, y_{i-1}]$ according to the classical AR(p) model where

$$y_i = \sum_{\ell \le p} \beta_\ell y_{i-\ell} + \varepsilon_i, \tag{11}$$

Here we consider the case $p = 1$, and demonstrate similar results for $p = 2$ in the Supplementary Material. We generate data for model with parameters $p = 1, n = 50, \beta = 0.99$, $y_0 = 0$ and $\varepsilon_i \sim \mathsf{N}(0, 1)$. As shown in Figure 1, the Gaussianity of the OLS estimate is intimately related to the stationarity of the series, in particular how close β is to 1 in terms of the number of data points n, or the length of the time series. We now evaluate the hypothesis testing procedures in terms of the following metrics: (i) Lower empirical coverage confidence $\mathbb{P}\{\beta \le \widehat{\beta} + \hat{\sigma}\Phi^{-1}(-\alpha)\}$, for various choices of the nominal confidence α, (ii) Upper empirical coverage probability $\mathbb{P}\{\beta \ge \widehat{\beta} - \hat{\sigma}\Phi^{-1}(-\alpha)\}$ and (iii) typical width of confidence region: $\mathbb{E}\{\hat{\sigma}\Phi^{-1}(-\alpha)\}$.

We plot the coverage confidences for various values of α in Figure 4 and the empirical widths (with standard errors on the widths) on the right panel of Figure 4. The QQ plot of the error distributions on the bottom right panel of Figure 4 shows that the OLS errors are skewed downwards, while the errors we obtain are appropriately Gaussian. We obtain the following improvements over the comparison methods of OLS

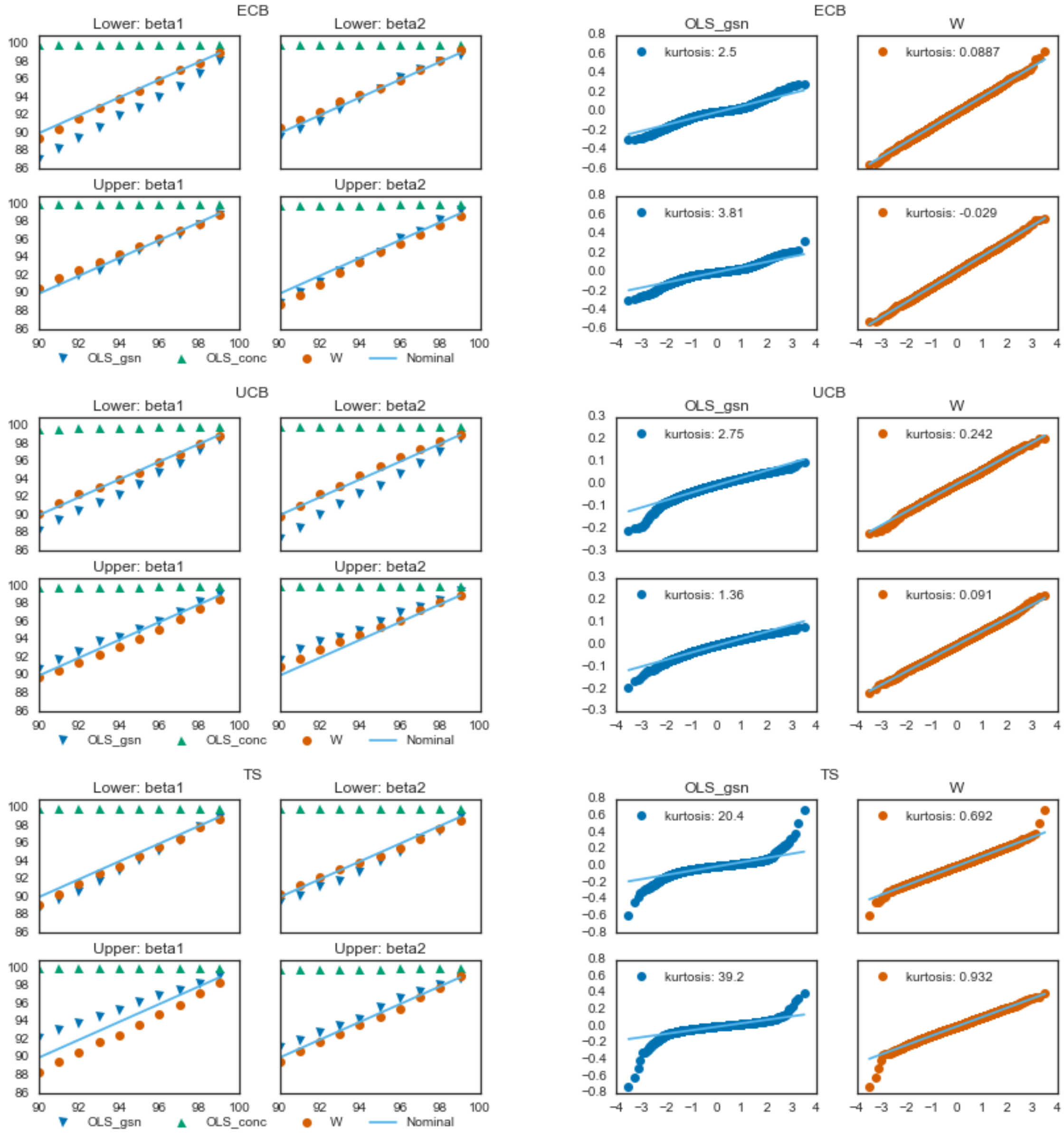

Figure 2: Left: One-sided confidence region coverage for OLS and decorrelated $\boldsymbol{W}$-decorrelated estimator across 10000 trials under ϵ-current best policy for multi-armed bandits. Right: Quantile-quantile (QQ) plots and empirical excess kurtosis (inset) for the OLS and $\boldsymbol{W}$-decorrelated estimator errors for each parameter β_k.

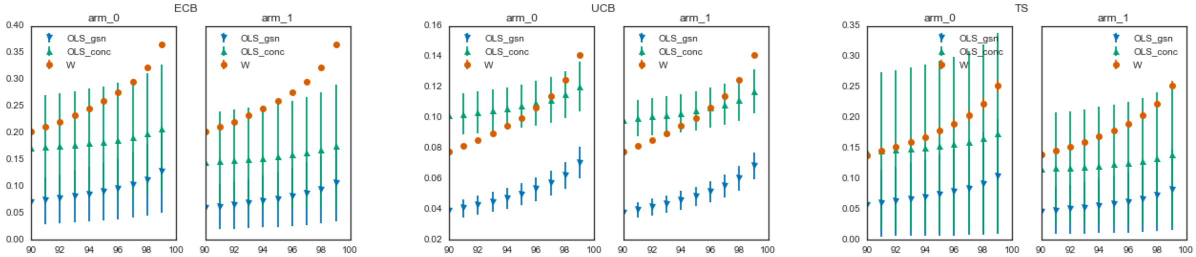

Figure 3: Mean 2-sided confidence interval widths (error bars show 1 standard deviation) for the 2 arms in the MAB experiment. The OLS-based widths vary significantly over runs of the experiment, while the $\boldsymbol{W}$- decorrelated estimate widths are very stable. This empirically demonstrates the failure of the stability condition in Theorem 1.

standard errors (labeled 'OLS') and concentration inequalities (labeled 'Conc') from Abbasi-Yadkori et al. [2011]

1. The Gaussian OLS confidence regions consistently underestimate the nominal coverage probability in the lower tail and overestimate the nominal coverage in the upper tail. Meanwhile, the concentration inequalities provide very conservative intervals that cover with nearly 100% probability, irrespective of the nominal level. In contrast, our decorrelated intervals achieve empirical coverage that closely approximates the nominal confidence levels.

2. In terms of width comparisons, our estimated widths compare favourably to the OLS as they are not much larger and are typically smaller than those obtained via concentration inequalities. Note also that the widths for OLS and the concentration bounds vary significantly over many runs, while our widths concentrate much better.

3. In Figure 4, we also plot the *empirical* coverage widths (labeled 'OLSEmp') for the OLS distribution as measured directly from the histogram in the left panel. Note that this can only be done as we know specifically the model, but it serves as an oracle bootstrap benchmark. Our confidence widths compare quite favorably to this oracle bootstrap as well.

   We also include results of a (nearly) non-stationary AR(2) experiment in Figure 5. All the above conclusions also hold for the AR(2) model.

5 Proofs

5.1 Proofs of Theorems 3 and 4

The proofs of the main results rely on the following simple lemma.

Lemma 8. *Consider the $\boldsymbol{W}$-estimate as defined in Algorithm 1. Assume $\|\boldsymbol{x}_i\|^2 \le C$. Then for any i,*

$$\|\mathrm{I}_p - \boldsymbol{W}_{i-1}\boldsymbol{X}_{i-1}\|_F^2 - \|\mathrm{I}_p - \boldsymbol{W}_i\boldsymbol{X}_i\|_F^2 \asymp 2\lambda\|\boldsymbol{w}_i\|^2 \tag{12}$$

Proof. This follows directly from the fact that $\boldsymbol{W}_i\boldsymbol{X}_i = \boldsymbol{W}_{i-1}\boldsymbol{X}_{i-1} + \boldsymbol{w}_i\boldsymbol{x}_i^\mathsf{T}$ and the following formula for $\boldsymbol{w}_i$:

$$\boldsymbol{w}_i = \frac{(\mathrm{I}_p - \boldsymbol{W}_{i-1}\boldsymbol{X}_{i-1})\boldsymbol{x}_i}{\lambda + \|\boldsymbol{x}_i\|^2} \tag{13}$$

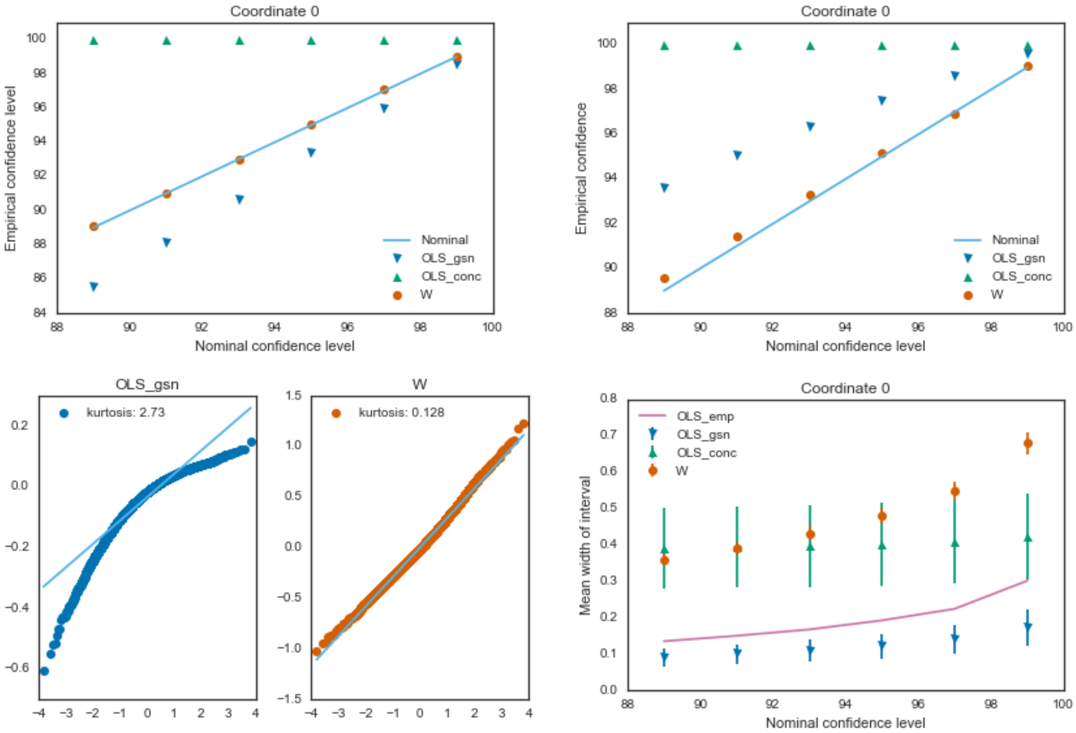

Figure 4: Lower (Top left) and upper (Top right) coverage probabilities for OLS with Gaussian intervals, OLS with concentration inequality intervals, and decorrelated $\boldsymbol{W}$-decorrelated estimate intervals. Note that 'Conc' has always 100% coverage. Bottom right: QQ plot for the distribution of errors of standard OLS estimate and the W-decorrelated estimate. Excess kurtosis is included inset. Right: Mean confidence widths for various estimators. The error bars show one (empirical) standard deviation of the confidence width.

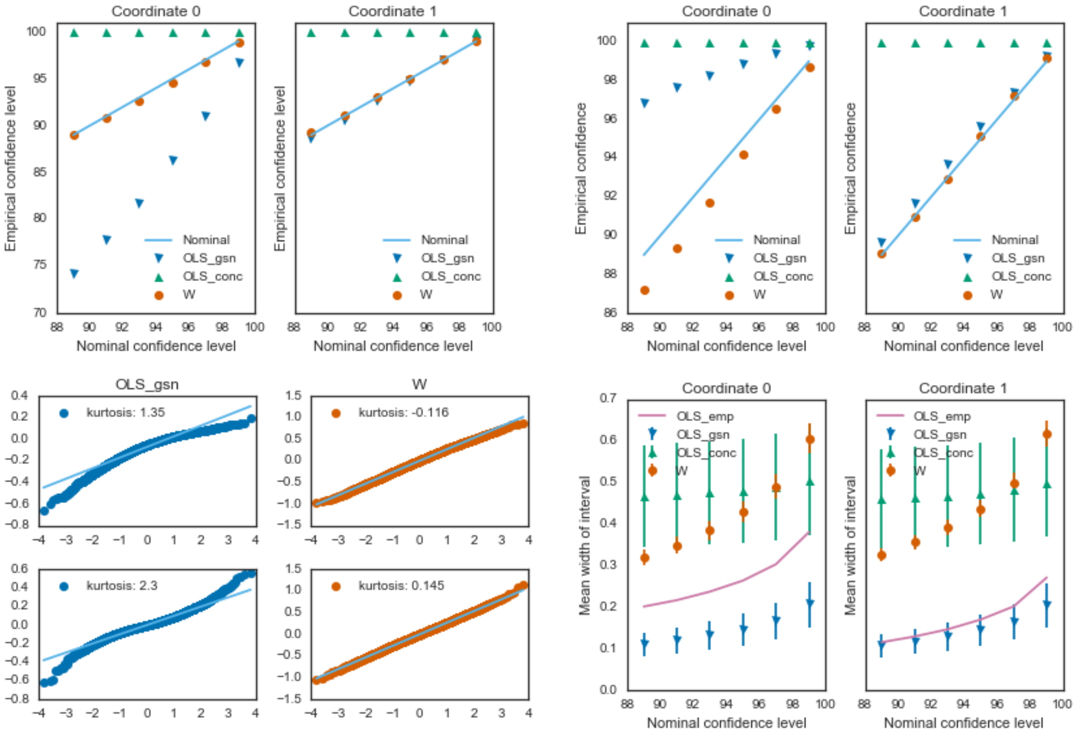

Figure 5: Lower (Top left) and upper (Top right) coverage probabilities for OLS with Gaussian intervals, OLS with concentration inequality intervals, and decorrelated $\boldsymbol{W}$-decorrelated estimate intervals. Note that 'Conc' has always 100% coverage. Bottom right: QQ plot for the distribution of errors of standard OLS estimate and the W-decorrelated estimate. Excess kurtosis is included inset. Right: Mean confidence widths for various estimators. The error bars show one (empirical) standard deviation of the confidence width.

which implies:

$$\|\mathrm{I}_p - \boldsymbol{W}_{i-1}\boldsymbol{X}_{i-1}\|_F^2 - \|\mathrm{I}_p - \boldsymbol{W}_i\boldsymbol{X}_i\|_F^2 = (2\lambda + \|\boldsymbol{x}_i\|^2)\|\boldsymbol{w}_i\|^2 \tag{14}$$

The result follows as $\|\boldsymbol{x}_i\|^2$ is bounded uniformly. □

We can now prove Theorems 3 and 4 in a straightforward fashion.

Proof of Theorem 3. We have:

$$\begin{aligned} \mathsf{Tr}\{\mathrm{Var}(\mathsf{v})\} &= \sigma^2\mathbb{E}\Big\{ \sum_i \|\boldsymbol{w}_i\|^2 \Big\} & (15)\\ &\asymp \frac{\sigma^2}{2\lambda}\Big(\|\mathrm{I}_p\|_F^2 - \mathbb{E}\big\{ \|\mathrm{I}_p - \boldsymbol{W}_n\boldsymbol{X}_n\|_F^2 \big\}\Big), & (16)\end{aligned}$$

where in the second line we use Lemma 8 and sum over the telescoping series in i. The result follows. □

Proof of Theorem 4. From Lemma 2 and Cauchy-Schwarz we have that

$$\|\beta - \mathbb{E}\{\widehat{\beta}\}\|^2 \le \mathbb{E}\{\|\mathrm{I}_p - \boldsymbol{W}_n\boldsymbol{X}_n\|_2^2\}\mathbb{E}\{\|\widehat{\beta}_{\mathsf{OLS}} - \beta\|^2\}. \tag{17}$$

Using Theorem 1, the second term is bounded by $p\sigma^2\mathbb{E}\{\log\lambda_{\max}/\lambda_{\min}\}$. We first show that this term is at most $p\sigma^2\log n/\lambda(n)$, under the conditions of Theorem 4. First, note that

$$\begin{aligned} \lambda_{\max} &\le \mathsf{Tr}\{\boldsymbol{X}_n\boldsymbol{X}_n^{\mathsf{T}}\} & (18)\\ &\le \sum_i \|\boldsymbol{x}_i^2\| & (19)\\ &\le Cn. & (20)\end{aligned}$$

With this and condition 8, we have that:

$$\begin{aligned} \mathbb{E}\bigg\{\frac{\log(\lambda_{\max})}{\lambda_{\min}}\bigg\} &\le \frac{\log n}{\lambda(n)} + O\Big(\frac{1}{n}\Big) & (21)\\ &= O\Big(\frac{\log n}{\lambda(n)}\Big). & (22)\end{aligned}$$

Combining with condition 7 and Theorem 3 gives the result. □

We split the proof of Proposition 5 for the different conditions independently in the following lemmas.

Lemma 9. *Suppose that the data collection process satisfies Eqs.9, 10. Then for any $\lambda \ge 1$ we have that:*

$$\mathbb{E}\big\{ \|\mathrm{I}_p - \boldsymbol{W}_n\boldsymbol{X}_n\|_F^2 \big\} \le p\exp\Big(-\frac{n\bar{\mu}_n}{\lambda}\Big) \tag{23}$$

Proof. Define $\boldsymbol{M}_i = \mathrm{I}_p - \boldsymbol{W}_i\boldsymbol{X}_i$. Then, from Lemma 8 and the closed form for $\boldsymbol{w}_i$ we have that:

$$\begin{aligned} \|\boldsymbol{M}_{i-1}\|_F^2 - \|\boldsymbol{M}_i\|_F^2 &= \frac{2\lambda + \|\boldsymbol{x}_i^2\|}{(\lambda + \|x_i\|^2)^2}\mathsf{Tr}\{\boldsymbol{M}_{i-1}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{M}_{i-1}^{\mathsf{T}}\} & (24)\\ &\ge \frac{1}{\lambda + \|\boldsymbol{x}_i\|^2}\mathsf{Tr}\{\boldsymbol{M}_{i-1}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{M}_{i-1}^{\mathsf{T}}\}. & (25)\end{aligned}$$

We now take expectations conditional on $\mathcal{G}_{i-1}$ on both sides. Observing that (i) $\boldsymbol{W}_n$, $\boldsymbol{X}_n$ and, therefore, $\boldsymbol{M}_n$ are well-adapted and (ii) using Condition 9, we have

$$\begin{aligned} \mathbb{E}\{\|\boldsymbol{M}_{i-1}\|_F^2\,|\mathcal{G}_{i-1}\} - \mathbb{E}\{\|\boldsymbol{M}_i\|_F^2\,|\mathcal{G}_{i-1}\} &\ge \frac{\mu_i(n)}{\lambda}\mathbb{E}\{\|\boldsymbol{M}_i\|_F^2\,|\mathcal{G}_{i-1}\}, & (26)\\ \text{or}\ \ \mathbb{E}\{\|\boldsymbol{M}_i\|_F^2\,|\mathcal{G}_{i-1}\} &\le \exp\Big(\frac{-\mu_i(n)}{\lambda}\Big)\mathbb{E}\{\|\boldsymbol{M}_{i-1}\|_F^2\,|\mathcal{G}_{i-1}\}. & (27)\end{aligned}$$

Removing the conditioning on $\mathcal{G}_{i-1}$ and iterating over $i = 1, 2, \ldots, n$ gives the claim. □

Lemma 10. *If the matrices $\{\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}}\}_{i\leq n}$ commute, we have that*

$$\|\mathrm{I}_p - \boldsymbol{W}_n\boldsymbol{X}_n\|_2 \leq \exp\Big(-\frac{\lambda_{\min}}{\lambda}\Big) \tag{28}$$

Proof. From the closed form in Lemma 8 and induction, we get that:

$$\mathrm{I}_p - \boldsymbol{W}_n\boldsymbol{X}_n = \prod_{i\leq n}\Big(\mathrm{I}_p - \frac{\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}}}{\lambda + \|x\|_i^2}\Big). \tag{29}$$

The scalar equality $\exp(a+b) = \exp(a)\exp(b)$ extends to commuting matrices $\boldsymbol{A}, \boldsymbol{B}$. Applying this to the terms in the product above, which commute by assumption:

$$\mathrm{I}_p - \boldsymbol{W}_n\boldsymbol{X}_n = \exp\Big[\sum_i \log\Big(\mathrm{I}_p - \frac{\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}}}{\lambda + \|x\|_i^2}\Big)\Big] \tag{30}$$

$$\preccurlyeq \exp\Big(-\sum_i \frac{\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}}}{\lambda}\Big), \tag{31}$$

using the fact that $\exp(\log(1-a)) \leq -a$. Finally, employing commutativity the fact that $\lambda_{\min}$ is the minimum eigenvalue of $\boldsymbol{X}_n^{\mathsf{T}}\boldsymbol{X}_n = \sum_i \boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}}$, the desired result follows. □

We can now prove Proposition 5.

Proof of Proposition 5. We need to satisfy conditions 7 and 8 for both the cases. Using either Lemma 9 or 10, with the appropriate choice of $\lambda(n)$ we have that

$$\mathbb{E}\{\|\mathrm{I}_p - \boldsymbol{W}_n\boldsymbol{X}_n\|_2^2\} = o(1/\log n), \tag{32}$$

thus obtaining condition 7. In fact, this can be made polynomially small with a constant factor smaller choice of $\lambda(n)$. Condition 8 only needs to be verified for the case of Lemma 9 or condition 9. It follows from a standard application of the matrix Azuma inequality Tropp [2012], the fact that $n\bar{\mu}_n \geq K\sqrt{n}$ and the fact that $\|\boldsymbol{x}\|_i^2$ are bounded. □

5.2 Proof of Theorem 10: Central limit theorem

It suffices to show the case for $p = 1$ of the theorem. In this case, it is not hard to show that the moment stability assumption of the main article subsumes the condition of the following:

Theorem 11. *Let X_i be a martingale difference sequence, adapted to the filtration $\mathcal{F}_i \subset \mathcal{F}$, with the predictable conditional covariance process $\sigma_i^2 = \mathbb{E}\{X_i^2|\mathcal{F}_{i-1}\}$. Define $\bar{\sigma}_i^2 = \sum_{\ell\leq i}\sigma_\ell^2$ and $S_n = \sum_{i\leq n} X_i$. Suppose, additionally, that $\{X_i\}_{i\leq n}$ satisfies:*

1. *For all i, $|X_i|$ are conditionally B-subgaussian for some constant B. almost surely.*
2. *For any fixed $t \in \mathbb{R}$,*

$$\frac{1}{\sqrt{n}}\sum_{i\leq n}\mathbb{E}\big|\mathbb{E}\{X_i\exp(-\bar{\sigma}_n^2t^2/n)|\mathcal{F}_{i-1}\}\big| = o_n(1), \tag{33}$$

$$\frac{1}{n}\sum_{i\leq n}\mathbb{E}\big|\mathbb{E}\{(X_i^2 - \sigma_i^2)\exp(-\bar{\sigma}_n^2t^2/n)|\mathcal{F}_{i-1}\}\big| = o_n(1), \tag{34}$$

Then, for any bounded and continous function $\varphi : \mathbb{R} \to \mathbb{R}$ we have that,

$$\big|\mathbb{E}\{\varphi(S_n/\sqrt{n})\} - \mathbb{E}\varphi(\bar{\sigma}_n\xi/\sqrt{n})\big| = o_n(1), \tag{35}$$

where $\xi \sim \mathsf{N}(0,1)$ is independent of $\bar{\sigma}_n$.

Proof. By Levy's continuity theorem it suffices to consider the subclass of functions $\varphi(x) = \exp(\mathrm{i}tx)$ for $t \in \mathbb{R}$. Note that, since ξ is independent of $\bar{\sigma}_n$, $\mathbb{E}\{\exp(\mathrm{i}t\bar{\sigma}_n\xi)|\mathcal{F}_i\} = \mathbb{E}\{\exp(-\bar{\sigma}_n^2 t^2/2)|\mathcal{F}_i\}$ a.s. for all t and i. The proof is mostly standard. Using the boundedness of X_i, note that $\bar{\sigma}_n^2 \le 2n$ a.s. Furthermore, for all i, $\bar{\sigma}_i^2 \in \mathrm{m}\mathcal{F}_{i-1}$ by definition. For simplicity, define the following errors that will ultimately be controlled by the conditions of the theorem:

$$\nu_i^1(t) \equiv \mathbb{E}\{X_i \exp(-\bar{\sigma}_n^2 t^2/n) - X_i|\mathcal{F}_{i-1}\} \tag{36}$$

$$\nu_i^2(t) \equiv \mathbb{E}\{(X_i^2 - \sigma_i^2)\exp(-\bar{\sigma}_n^2 t^2/n)\}. \tag{37}$$

We will show that

$$\begin{aligned} &|\mathbb{E}\{\exp(\mathrm{i}tS_i/\sqrt{n} + (\bar{\sigma}_i^2 - \bar{\sigma}_n^2)t^2/2n)|\mathcal{F}_{i-1}\} - \exp(\mathrm{i}tS_{i-1}/\sqrt{n} + \bar{\sigma}_{i-1}^2 t^2/2n)\mathbb{E}\{\exp(-\bar{\sigma}_n^2 t^2/2n)|\mathcal{F}_{i-1}\}| \\ &\qquad \le \quad \frac{t}{\sqrt{n}}|\nu_i^1(t)| + \frac{t^2}{2n}|\nu_i^2(t)| + \frac{C}{n^{3/2}} \end{aligned} \tag{38}$$

Taking expectations above and summing the above estimate from $i = 1$ to n yields

$$|\mathbb{E}\{\exp(\mathrm{i}tS_n/\sqrt{n})\} - \mathbb{E}\{\exp(-\bar{\sigma}_n^2 t^2/2n\}| \le \frac{t}{\sqrt{n}}\sum_i \mathbb{E}\{|\nu_i^1(t)|\} + \frac{t^2}{2n}\sum_{i=1}^n \mathbb{E}\{|\nu_i^2(t)|\} + \frac{C}{\sqrt{n}}. \tag{39}$$

By assumptions (33) and (34), this implies the claim for complex exponential test functions as required.

It remains to show the estimate Eq. (38). By Taylor expansion

$$\exp(\mathrm{i}tS_i/\sqrt{n}) = \exp(\mathrm{i}tS_{i-1}/\sqrt{n})\exp(\mathrm{i}tX_i/\sqrt{n}) \quad = \exp(\mathrm{i}tS_{i-1}/\sqrt{n})\Big(1 + \frac{\mathrm{i}tX_i}{\sqrt{n}} - \frac{t^2X_i^2}{2n}\Big) + O\Big(\frac{t^3|X_i|^3}{n^{3/2}}\Big). \tag{40}$$

Therefore

$$\begin{aligned} \mathbb{E}\{\exp(\mathrm{i}tS_i/\sqrt{n} + (\bar{\sigma}_i^2 - \bar{\sigma}_n^2)t^2/2n)|\mathcal{F}_{i-1}\} &= \exp(\mathrm{i}tS_{i-1}/\sqrt{n} + t^2\bar{\sigma}_i^2/2n)\times \\ &\mathbb{E}\Big\{\Big(1 + \frac{\mathrm{i}tX_i}{\sqrt{n}} - \frac{t^2X_i^2}{2n}\Big)\exp(-t^2\bar{\sigma}_n^2/2n)\Big|\mathcal{F}_{i-1}\Big\} + O\Big(\frac{t^3}{n^{3/2}}\Big) \qquad (41) \\ &= \exp(\mathrm{i}tS_{i-1}/\sqrt{n} + t^2\bar{\sigma}_i^2/2n)\mathbb{E}\{\exp(-t^2(\bar{\sigma}_n^2 + \sigma_i^2)/2n)|\mathcal{F}_{i-1}\} \\ &\quad + O\Big(\frac{t|\nu_i^1(t)|}{\sqrt{n}} + \frac{t^2|\nu_i^2(t)|}{n} + \frac{t^3}{n^{3/2}}\Big), \qquad (42) \end{aligned}$$

where we used the fact that $|X_i| \le 1$ and that $\exp(-z^2/2) = 1 - z^2/2 + O(z^4)$ for $|z| \le 1$. Rearranging, we obtain the required estimate in Eq. (38).

□

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670*, 2011.

Susan Athey and Stefan Wager. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017.

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.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Dimitri P Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical programming*, 129(2):163, 2011.

Patrick Billingsley. *Probability and measure.* John Wiley & Sons, 2008.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Rui M Castro and Robert D Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008.

Ngai H Chan and Ching-Zong Wei. Asymptotic inference for nearly nonstationary ar (1) processes. *The Annals of Statistics*, pages 1050–1063, 1987.

Kani Chen, Inchi Hu, Zhiliang Ying, et al. Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *The Annals of Statistics*, 27(4):1155–1163, 1999.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *COLT*, pages 355–366, 2008.

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

David A Dickey and Wayne A Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a):427–431, 1979.

Maria Dimakopoulou, Susan Athey, and Guido Imbens. Estimation considerations in contextual bandits. *arXiv preprint arXiv:1711.07077*, 2017.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

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

Walter Enders. *Applied econometric time series.* John Wiley & Sons, 2008.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 359–376, 2011.

Christopher C Heyde. *Quasi-likelihood and its application: a general approach to optimal parameter estimation.* Springer Science & Business Media, 2008.

Guido W Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3): 706–710, 2000.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lilucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439, 2014.

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

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

Nathan Kallus. Balanced policy evaluation and learning. *arXiv preprint arXiv:1705.07384*, 2017.

Brian Kulis and Peter L Bartlett. Implicit online learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 575–582, 2010.

TseLeung Lai and David Siegmund. Fixed accuracy estimation of an autoregressive parameter. *The Annals of Statistics*, pages 478–485, 1983.

Tze Leung Lai. Asymptotic properties of nonlinear least squares estimates in stochastic regression models. *The Annals of Statistics*, pages 1917–1930, 1994.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

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*, pages 154–166, 1982.

M. Ledoux. *Isoperimetry and Gaussian analysis*, volume 1648. Springer, Providence, 1996.

Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2005.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306. ACM, 2011.

Jared K Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960, 2004.

Jin-Ichi Nagumo and Atsuhiko Noda. A learning method for system identification. *IEEE Transactions on Automatic Control*, 12(3):282–287, 1967.

Xinkun Nie, Tian Xiaoying, Jonathan Taylor, and James Zou. Why adaptively collected data have negative bias and how to correct for it. 2017.

Peter CB Phillips and Pierre Perron. Testing for a unit root in time series regression. *Biometrika*, 75(2): 335–346, 1988.

Herbert Robbins. Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*, pages 169–177. Springer, 1985.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Daniel Russo. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418, 2016.

Robert H Shumway and David S Stoffer. *Time series analysis and its applications: with R examples*. Springer Science & Business Media, 2006.

Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16:1731–1755, 2015.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Sara Van de Geer, Peter Bühlmann, Yaacov Ritov, Ruben Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

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(2):199, 2015.

Yixin Wang and José R Zubizarreta. Approximate balancing weights: Characterizations from a shrinkage estimation perspective. *arXiv preprint arXiv:1705.00998*, 2017.

Ching-Zong Wei. Asymptotic properties of least-squares estimates in stochastic regression models. *The Annals of Statistics*, pages 1498–1508, 1985.

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