Accurate Inference for Adaptive Linear Models

Yash Deshpande [1] **Lester Mackey** [2] **Vasilis Syrgkanis** [2] **Matt Taddy** [2 3]

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.

1. Introduction

Consider a dataset of n sample points $(y_i, \boldsymbol{x}_i)_{i \le n}$ where y_i represents an observed outcome and $\boldsymbol{x}_i \in \mathbb{R}^p$ an associated vector of covariates. In the standard linear model, the outcomes and covariates are related through a parameter β:

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

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 problems. A common additional assumption is that the covariate vector $\boldsymbol{x}_i$ for a given datapoint i is independent of the other sample point outcomes $(y_j)_{j \neq i}$ and the inherent variation $(\varepsilon_j)_{j \in [n]}$. This paper is motivated by experiments where the sample $(y_i, \boldsymbol{x}_i)_{i \le n}$ is not completely randomized but rather *adaptively* chosen. By adaptive, we mean that the choice of the data point $(y_i, \boldsymbol{x}_i)$ is guided from inferences on past data $(y_j, \boldsymbol{x}_j)_{j<i}$. 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: Acquiring labels y_i is potentially costly, and the experimenter aims to learn with as few outcomes as possible. At time i, based on prior data $(y_j, \boldsymbol{x}_j)_{j \le i-1}$ the experimenter chooses a new data point $\boldsymbol{x}_i$ to label based on its value in learning.
3. Time series analysis: Here, the data points $(y_i, \boldsymbol{x}_i)$ are naturally ordered in time, with $(y_i)_{i \le n}$ denoting a time series and the covariates $\boldsymbol{x}_i$ include observations from the prior 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 $(\boldsymbol{x}_i)_{i \le n}$ 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 & Nowak, 2008; Audibert & Bubeck, 2009; Bubeck et al., 2012; Rusmevichientong & Tsitsiklis, 2010). Our motivation is the natural follow-up question of accurate *ex post* inference in the standard statistical sense:

[1]Department of Mathematics, MIT [2]Microsoft Research New England [3]Booth School of Business, University of Chicago. Correspondence to: Yash Deshpande <yash@mit.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

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 (CLT) 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. Due to page constraints, all proofs are given in Appendix A in the supplementary information.

2. Main results: W-decorrelation

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. Formally, 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 post* 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 $\boldsymbol{X}_n = [\boldsymbol{x}_1^{\mathsf{T}}, \dots \boldsymbol{x}_n^{\mathsf{T}}] \in \mathbb{R}^{n \times p}$ is the design matrix and $\boldsymbol{y}_n = [y_1, \dots y_n] \in \mathbb{R}^n$. When data collection is non-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 & Wei (1982) extend these results to the current scenario:

Theorem 1 (Theorems 1, 3 (Lai & 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:*

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

Further assume the following stability condition: there exists a deterministic sequence of matrices $\boldsymbol{A}_n$ such that (iii) $\boldsymbol{A}_n^{-1} (\boldsymbol{X}_n{}^{\mathsf{T}} \boldsymbol{X}_n)^{1/2} \to \mathrm{I}_p$ and (iv) $\max_i \left\| \boldsymbol{A}_n^{-1} \boldsymbol{x}_i \right\|_2 \to 0$ 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 $[\widehat{\beta}_{\mathsf{OLS},1} - 1.96 \widehat{\sigma} (\boldsymbol{X}_n^{\mathsf{T}} \boldsymbol{X}_n)^{-1}_{11}, \widehat{\beta}_{\mathsf{OLS},1} + 1.96 \widehat{\sigma} (\boldsymbol{X}_n^{\mathsf{T}} \boldsymbol{X}_n)^{-1}_{11}]$ 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: the CLT need not hold for the OLS estimator (Lai & Wei, 1982; Lai & Siegmund, 1983).
2. The rate of convergence to the asymptotic CLT 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 & Robbins, 1985). At each time point $i \leq n$, the experimenter (or data collecting policy) chooses an arm $k \in \{1, 2, \ldots, p\}$ 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 vector $\boldsymbol{x}_i$ 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 & Tsitsiklis, 2010; Li et al., 2010; Deshpande & 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 computing sample means for each arm. The stability condition of Theorem 1 requires that $N_k(n)$, 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 & 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 $(\log n)^{-1/2}$ 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:

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

Here the least squares estimate is given by $\widehat{\beta}_{\mathsf{OLS}} = \sum_{i \leq n-1} y_{i+1} y_i / \sum_{i \leq n-1} y_{i-1}^2$. 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 $\beta - 1$ (cf. (Chan & Wei, 1987)). A 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

We propose to decorrelate the OLS estimator by constructing:

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

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

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

We interpret b as a 'bias' and v as a 'variance'. This is based on 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.

[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.

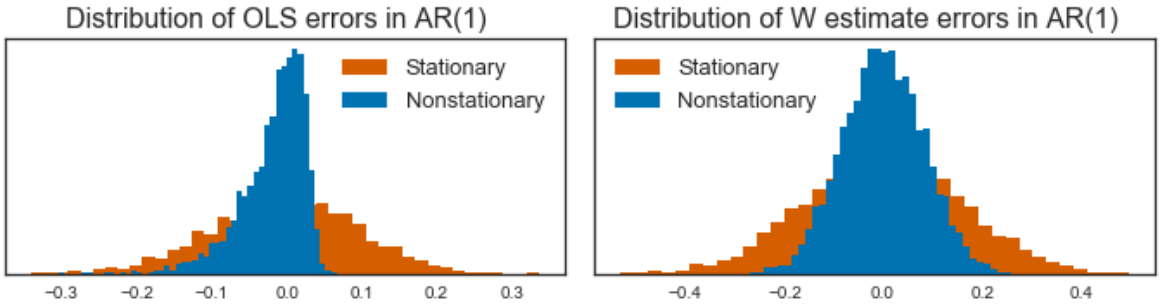

Figure 1. The distribution of errors for (left) the OLS estimator for stationary and (nearly) nonstationary AR(1) time series and (right) error distribution for both models after decorrelation. $n = 50$, $\varepsilon_i \sim \mathsf{N}(0, 1)$.

Lemma 2. *Assume $\boldsymbol{W}_n$ is well-adapted. Then:*

$$\|\beta - \mathbb{E}\{\widehat{\beta}^d\}\|_2 \leq \mathbb{E}\{\|\mathsf{b}\|_2\},$$
$$\mathrm{Var}(\mathsf{v}) = \sigma^2 \mathbb{E}\{\boldsymbol{W}_n \boldsymbol{W}_n^{\mathsf{T}}\}.$$

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:

$$\boldsymbol{W}_n = \arg\min_{\boldsymbol{W}} \|\mathrm{I}_p - \boldsymbol{W}\boldsymbol{X}_n\|_F^2 + \lambda \mathsf{Tr}(\boldsymbol{W}\boldsymbol{W}^{\mathsf{T}}). \quad (3)$$

Solving the above in closed form yields ridge estimators for β, and by continuity, also the standard least squares estimator. Departing from (Zhang & Zhang, 2014; Javanmard & Montanari, 2014a), we solve the above in an *online* fashion in order to obtain a well-adapted $\boldsymbol{W}_n$. We define, $\boldsymbol{W}_0 = 0$, $\boldsymbol{X}_0 = 0$, and recursively $\boldsymbol{W}_n = [\boldsymbol{W}_{n-1} \boldsymbol{w}_n]$ for

$$\boldsymbol{w}_n = \arg\min_{\boldsymbol{w} \in \mathbb{R}^p} \|\mathrm{I} - \boldsymbol{W}_{n-1}\boldsymbol{X}_{n-1} - \boldsymbol{w}\boldsymbol{x}_n^T\|_F^2 + \lambda \|\boldsymbol{w}\|_2^2 .$$

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 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. Interpretation as reverse implicit SGD

While we motivated $\boldsymbol{W}$-decorrelation decorrelation as an online procedure for optimizing the bias-variance trade-off objective (3), it holds a dual interpretation as implicit stochastic gradient descent (SGD) (see, e.g., Kulis & Bartlett, 2010), also known as incremental proximal minimization (Bertsekas, 2011) or the normalized least mean squares filter (Nagumo & 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 $(y_n, \boldsymbol{x}_n)$ and ending with the initial observation $(y_1, \boldsymbol{x}_1)$; this recipe yields the parameter updates $\widehat{\beta}_0 = \widehat{\beta}_{\mathsf{OLS}}$ and

$$\begin{aligned}
\widehat{\beta}_{i+1} &= \widehat{\beta}_i + \boldsymbol{x}_{n-i}(y_{n-i} - \langle \boldsymbol{x}_{n-i}, \widehat{\beta}_{i+1} \rangle)/\lambda \\
&= (\mathrm{I}_p + \boldsymbol{x}_{n-i}\boldsymbol{x}_{n-i}^{\mathsf{T}}/\lambda)^{-1}(\widehat{\beta}_i + y_{n-i}\boldsymbol{x}_{n-i}/\lambda) \\
&= (\mathrm{I}_p - \boldsymbol{x}_{n-i}\boldsymbol{x}_{n-i}^{\mathsf{T}}/(\lambda + \|\boldsymbol{x}_{n-i}\|_2^2))\widehat{\beta}_i \\
&\quad + y_{n-i}\boldsymbol{x}_{n-i}/(\lambda + \|\boldsymbol{x}_{n-i}\|_2^2).
\end{aligned}$$

Unrolling the recursion, we obtain $\widehat{\beta}_n = \widehat{\beta}_{\mathsf{OLS}} + \sum_{i=1}^n y_i \boldsymbol{w}_i$ with each $\boldsymbol{w}_i$ precisely as in Algorithm 1:

$$\boldsymbol{w}_i = \textstyle\prod_{j=1}^n (\mathrm{I}_p - \boldsymbol{x}_j \boldsymbol{x}_j^{\mathsf{T}}/(\lambda + \|\boldsymbol{x}_j\|_2^2)).$$

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*

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

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:*

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

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 λ, 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:*

$$\mathbb{E}\{\|\mathrm{I}_p - \boldsymbol{W}_n\boldsymbol{X}_n\|_{op}^2\} = o(1/\log n), \tag{4}$$

$$\mathbb{P}\{\lambda_{\min}(\boldsymbol{X}_n\boldsymbol{X}_n^\mathsf{T}) \le \lambda(n)\log n\} \le 1/n, \tag{5}$$

Then we have

$$\frac{\mathbb{E}\{\|\mathsf{b}\|_2^2\}}{\mathsf{Tr}\{\mathrm{Var}(\mathsf{v})\}} = o(1).$$

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 $\lambda \ge 1$:*

$$\mathbb{E}\{\tfrac{\boldsymbol{x}_i\boldsymbol{x}_i^\mathsf{T}}{\lambda+\|\boldsymbol{x}_n\|_2^2}|\mathcal{G}_{i-1}\} \succcurlyeq \tfrac{\mu_n(i)}{\lambda}\mathrm{I}_p, \tag{6}$$

$$\textstyle\sum_i \mu_n(i) \equiv n\bar{\mu}_n \ge K\sqrt{n}, \tag{7}$$

   for a large enough constant K. Here we keep $\lambda(n) \asymp n\bar{\mu}_n/\log n$.

2. *The matrices $(\boldsymbol{x}_i\boldsymbol{x}_i^\mathsf{T})_{i\le n}$ commute and $\lambda(n)\log n$ is (at most) the $1/n^{th}$ percentile of $\lambda_{\min}(\boldsymbol{X}_n^\mathsf{T}\boldsymbol{X}_n)$.*

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 $\|\mathrm{I} - \boldsymbol{W}_n\boldsymbol{X}_n\|_{op}$, 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 $\beta \in \mathbb{R}^p$ is now impossible regardless of the choice of $\boldsymbol{W}_n$. For 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. Proposition 5 the intuition that the data collecting policy should explore the full parameter space. For multi-armed bandits, policies such as epsilon-greedy and Thompson sampling satisfy this assumption with appropriate $\mu_n(i)$.

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.

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

Input: sample $(y_i, \mathbf{x}_i)_{i\le n}$, regularization λ, 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 $\boldsymbol{w}_i = (\mathrm{I}_p - \boldsymbol{W}_{i-1}\boldsymbol{X}_i^\mathsf{T})\boldsymbol{x}_i/(\lambda + \|\boldsymbol{x}_i\|_2^2)$, for $i = 1, 2, \ldots, n$.
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: decorrelated estimate $\widehat{\beta}^d$ and CI interval $I(\boldsymbol{v},\alpha) = [\langle \boldsymbol{v}, \widehat{\beta}^d\rangle - \hat{\sigma}(\boldsymbol{v})\Phi^{-1}(1-\alpha), \langle \boldsymbol{v}, \widehat{\beta}^d\rangle + \hat{\sigma}(\boldsymbol{v})\Phi^{-1}(1-\alpha)]$.

Before we move to the 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 or conditional likelihoods 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. Moreover, propensity scores or conditional likelihoods are ineffective when data collection policies make adaptive decisions that are deterministic given the history. A important example is that of UCB algorithms for bandits, which make deterministic choices of arms.

2.4. A central limit theorem and confidence intervals

Our final result is a simple CLT that provides an alternative to the stability condition of Theorem 1 and standard martingale CLTs.[2] 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. Define, for any vector $\boldsymbol{t} \in \mathbb{R}^p$, the conditional variance $\sigma_i(\boldsymbol{t}) \equiv \sum_{j\le i}\langle \boldsymbol{w}_j, \boldsymbol{t}\rangle^2$. We make the following crucial moment stability assumption on the conditional covariance:

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

$$\sup_{\|\boldsymbol{t}\|_2\le 1}\textstyle\sum_{i=1}^{m(n)}\mathbb{E}\{|\mathbb{E}\{\varepsilon_i^a\sigma_m(\boldsymbol{t})^k|\mathcal{F}_{i-1}\} - \mathbb{E}\{\varepsilon_i^a|\mathcal{F}_{i-1}\}\mathbb{E}\{\sigma_m(\boldsymbol{t})^k\}|\mathcal{F}_{i-1}\}| = o_n(1).$$

Theorem 6 (Martingale CLT)**.** *Let $(w_i(n), \varepsilon_i(n), \mathcal{F}_i(n))_{i\le m(n)}$ be a triangular martingale difference array. Here for each $n \ge 1$, $\mathcal{F}_i(n)$ is a non-decreasing sequence of sub-sigma-algebras, $\varepsilon_i(n)$*

[2] Many standard martingale CLTs (see, e.g., Lai & Wei, 1982; Dvoretzky, 1972) demand the convergence of $\sum_i \boldsymbol{w}_i\boldsymbol{w}_i^\mathsf{T}/n$ to a constant, but this convergence condition is violated in many examples of interest, including the AR examples in Section 4.

are i.i.d. (uniformly) bounded random variables with $\mathbb{E}\{\varepsilon_i|\mathcal{F}_{i-1}\} = 0, \mathbb{E}\{\varepsilon_i^2|\mathcal{F}_{i-1}\} = 1$ *and* $w_i \in \mathrm{m}\mathcal{F}_{i-1}$ *is predictable and bounded by 1 almost surely. Suppose a Lyapunov condition holds, i.e.* $\sum_i \mathbb{E}\{w_i^3\} = o_n(1)$. *Then* $(\sum_i \boldsymbol{w}_i\boldsymbol{w}_i^\mathsf{T})^{-1/2}\sum_i \boldsymbol{w}_i\varepsilon_i \stackrel{\mathrm{d}}{\Rightarrow} \mathsf{N}(0, \sigma^2\mathrm{I}_p)$. *In particular, for any bounded, continuous* $\varphi : \mathbb{R}^p \to \mathbb{R}$

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

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 CLT. Assumption 1 is an alternate form of stability. 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 & Wei, 1982; Dvoretzky, 1972). We have the following

Lemma 7. *Consider a martingale sequence* $\sum_i \boldsymbol{w}_i\varepsilon_i$ *as in Theorem 6. 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 CLT in hand, one can now assign confidence intervals in the standard fashion, based on the assumption that the bias is negligible. For instance, we have the following result on two-sided confidence intervals.

Proposition 8. *Fix any* $\alpha > 0$. *Suppose that the data collection process satisfies the assumptions of Theorems 4 and 6. Set* $\lambda = \lambda(n)$ *as in Theorem 4, and let* $\widehat{\sigma}$ *be a consistent estimate of* σ *as in Theorem 1. Define* $\boldsymbol{Q} = \widehat{\sigma}^2\boldsymbol{W}_n\boldsymbol{W}_n^\mathsf{T}$ *and the interval* $I(a,\alpha) = [\widehat{\beta}_a^d - \sqrt{Q_{aa}}\Phi^{-1}(1-\alpha/2), \widehat{\beta}_a^d + \sqrt{Q_{aa}}\Phi^{-1}(1-\alpha/2)$. *Then*

$$\limsup_{n\to\infty} \mathbb{P}\{\beta_a \notin I(a,\alpha)\} \leq \alpha.$$

3. Related work

There is extensive work in statistics and econometrics on stochastic regression models (Wei, 1985; Lai, 1994; Chen et al., 1999; Heyde, 2008) and non-stationary time series (Shumway & Stoffer, 2006; Enders, 2008; Phillips & Perron, 1988). The former extend Theorem 1 using similar assumptions, while the latter consider specific time series models and optimal testing restricted to those cases. We instead focus on literature from sequential decision-making, online 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 & Robbins, 1985) has spurred a vast literature 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 & Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011; Jamieson et al., 2014) based on martingale concentration and the law of iterated logarithm. While these results can certainly be used to compute valid confidence intervals, they 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 it. 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 conservative intervals.

A closely-related line of work is that of learning from logged data (Li et al., 2011; Dudík et al., 2011; Swaminathan & Joachims, 2015) and policy learning (Athey & Wager, 2017; Kallus, 2017). The focus here is efficiently estimating the reward (or value) of a certain test policy using data collected from a different policy. For linear models, this reduces to accurate prediction which is directly related to the estimation error on the parameters β. While our work shares some features, 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 & 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 & Zubizarreta (2017) which propose balancing covariates or residuals for causal inference in the potential outcomes framework.

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. It is an interesting, and open, direction to consider

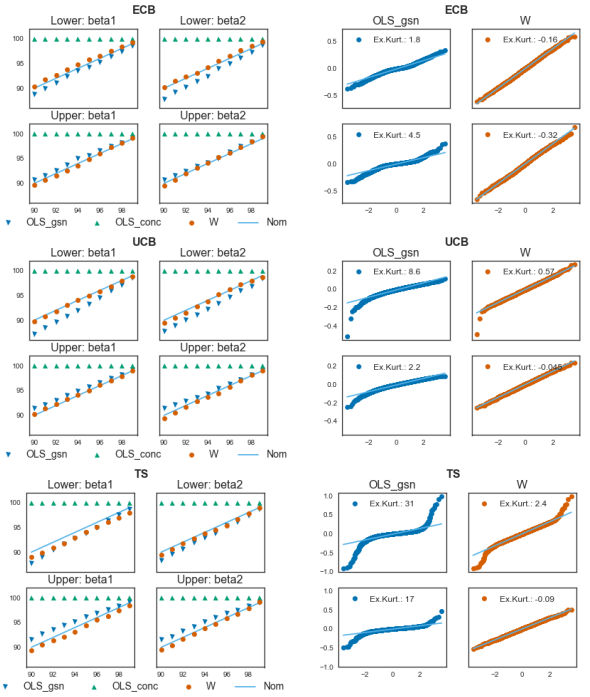

Figure 2. Left: One-sided confidence region coverage for OLS and decorrelated $\boldsymbol{W}$-decorrelated estimator trials. Right: Quantile-quantile (QQ) plots and empirical excess kurtosis (inset) for the OLS and $\boldsymbol{W}$-decorrelated estimator errors for each parameter β_k.

approaches that can combine the strengths of our approach and that of (Nie et al., 2017).

4. Experiments

In this section we empirically validate the decorrelated estimators in two scenarios that involve sequential dependence in covariates. Continuing from Section 2, our first scenario is a simple experiment of multi-armed bandits while the second scenario is autoregressive time series data. In these cases, we compare the empirical coverage and typical widths of confidence intervals for parameters obtained via three methods: (i) classical OLS theory, (ii) concentration inequalities and (iii) decorrelated estimates.

4.1. Multi-armed bandits

In this section, we demonstrate the utility of the $\boldsymbol{W}$-estimator for a stochastic multi-armed bandit setting. Villar et al. (2015) studied this problem in the context of patient allocation in clinical trials. Here the trial proceeds in a sequential fashion with the i^{th} patient given one of p treatments, encoded as $\boldsymbol{x}_i = \boldsymbol{e}_a$ with $a \in [p]$, and y_i denoting the outcome observed. We model the outcome as $y_i = \langle \boldsymbol{x}_i, \beta \rangle + \varepsilon_i$ where $\varepsilon_i \sim \mathsf{Unif}([-1, 1])$ with β being the mean outcome of the p treatments.

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 $\mu_0 = 0.3$ and variance $\sigma_0^2 = 0.33$) on each unknown mean outcome parameter $\beta = (0.3, 0.31)$ 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 (6) 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, see Jamieson et al. (2014). Our goal is to produce confidence intervals for the mean effect β_a of each treatment based on the data adaptively collected from these standard bandit algorithms.

We repeat this simulation 4000 times. From each trial simulation, we estimate the parameters β using both OLS and the $\boldsymbol{W}$-estimator with $\lambda = \hat{\lambda}_{10\%,\pi}$ which is the 10^{th} percentile of $\lambda_{\min}(n)$ achieved by the policy $\pi \in \{\text{ECB}, \text{UCB}, \text{TS}\}$. 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-based concentration inequality regions (OLS_{conc}) (Abbasi-Yadkori et al., 2011, Sec. 4). Figure 2 (left column) shows that the OLS Gaussian have lower tail regions that typically overestimate coverage and upper tail regions that typically underestimate coverage. This is consistent with the observation that the sample means are biased negatively (Nie et al., 2017). The concentration OLS tail bounds are all conservative, producing nearly 100% coverage, irrespective of the nominal level. Meanwhile, the decorrelated intervals provide faithful empirical coverage for every scenario apart from a few cases in Thompson sampling.

Figure 2 (right column) shows the QQ 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 reduced for every policy by at least an order of

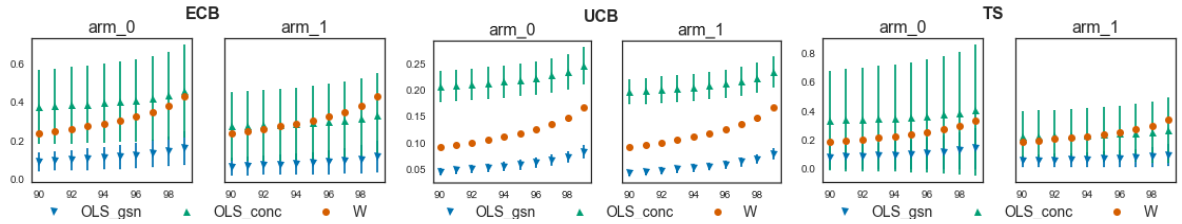

Figure 3. Mean 2-sided confidence interval widths (error bars show 1 standard deviation) for the 2 arms in the MAB experiment.

magnitude. Indeed, it is nearly 0 for ECB and UCB.

Figure 3 summarizes the distribution of 2-sided interval widths produced by each method for each arm. As expected, the $\boldsymbol{W}$-decorrelation intervals are wider than those of $\mathsf{OLS}_{\text{gsn}}$ but compare favorably with those provided by $\mathsf{OLS}_{\text{conc}}$. For UCB and for 'arm_0' for all policies, the mean $\mathsf{OLS}_{\text{conc}}$ widths are always largest. For 'arm_1' in the ECB and TS policies, $\boldsymbol{W}$-decorrelation yields smaller intervals than $\mathsf{OLS}_{\text{conc}}$ for moderate confidence levels and larger for high confidence levels.

4.2. Autoregressive time series

In this section, we consider the classical AR(p) model where $y_i = \sum_{\ell \le p} \beta_\ell y_{i-\ell} + \varepsilon_i$.. We generate data for the model with parameters $p = 2, n = 50, \beta = (0.95, 0.2)$, $y_0 = 0$ and $\varepsilon_i \sim \mathsf{Unif}([-1, 1])$; all estimates are computed over 4000 monte carlo iterations.

We plot the coverage confidences for various values of the nominal 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 $\boldsymbol{W}$-estimate errors are nearly Gaussian. We obtain the following improvements over the comparison methods of OLS standard errors $\mathsf{OLS}_{\text{gsn}}$ and concentration inequality widths $\mathsf{OLS}_{\text{conc}}$ (Abbasi-Yadkori et al., 2011)

The Gaussian OLS confidence regions systematically give incorrect empirical coverage. Meanwhile, the concentration inequalities provide very conservative intervals, with nearly 100% coverage, irrespective of the nominal level. In contrast, our decorrelated intervals achieve empirical coverage that closely approximates the nominal confidence levels.

These coverage improvements are enabled by an increase in width over that of $\mathsf{OLS}_{\text{gsn}}$, but the $\boldsymbol{W}$-estimate widths are systematically smaller than those of the concentration inequalities. Note also that the widths of $\mathsf{OLS}_{\text{conc}}$ vary significantly across runs, while the $\boldsymbol{W}$-estimate widths have minimal variability. Empirically, this demonstrates the stability condition of Theorem 1.

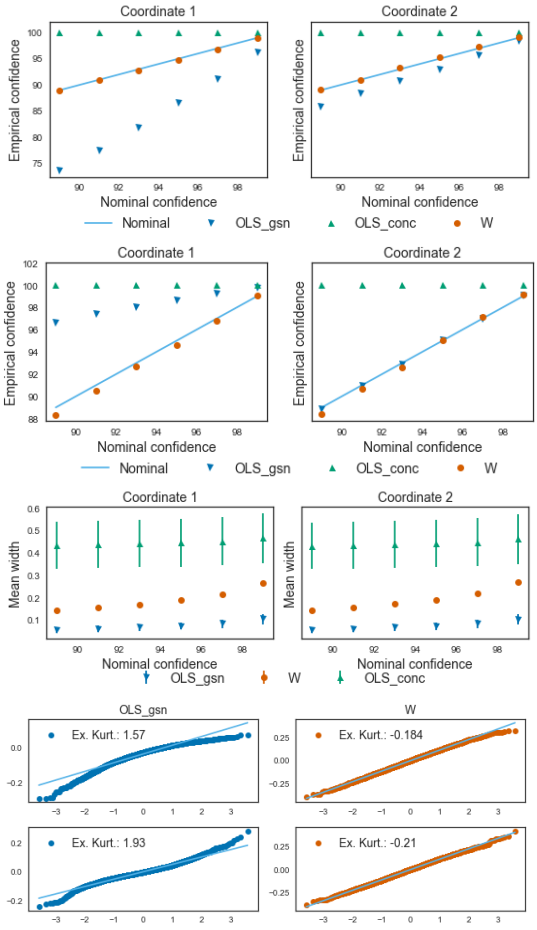

Figure 4. Lower (Top left) and upper (Top right) coverage probabilities for OLS with Gaussian intervals, OLS with concentration inequality intervals and decorrelated $\boldsymbol{W}$-decorrelated estimate intervals. QQ plot with kurtosis inset (bottom right) errors in OLS estimate and W-decorrelated estimate. Mean confidence widths (bottom left) for OLS, concentration and $\boldsymbol{W}$-decorrelated estimates. Error bars show one standard deviation.

References

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

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

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

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

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

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

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

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

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

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

Chen, Kani, Hu, Inchi, Ying, Zhiliang, 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Lai, Tze Leung and Wei, Ching Zong. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, pp. 154–166, 1982.

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

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

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

Li, Lihong, Chu, Wei, Langford, John, and Wang, Xuanhui. 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*, pp. 297–306. ACM, 2011.

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

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

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

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

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

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

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

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

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

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

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

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

Villar, Sofia, Bowden, Jack, and Wason, James. 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.

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

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

Zhang, Cun-Hui and Zhang, Stephanie S. 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.