Online Statistical Inference for Contextual Bandits via Stochastic Gradient Descent

Xi Chen* Zehua Lai† He Li‡ Yichen Zhang§

Abstract

With the fast development of big data, it has been easier than before to learn the optimal decision rule by updating the decision rule recursively and making online decisions. We study the online statistical inference of model parameters in a contextual bandit framework of sequential decision-making. We propose a general framework for online and adaptive data collection environment that can update decision rules via weighted stochastic gradient descent. We allow different weighting schemes of the stochastic gradient and establish the asymptotic normality of the parameter estimator. Our proposed estimator significantly improves the asymptotic efficiency over the previous averaged SGD approach via inverse probability weights. We also conduct an optimality analysis on the weights in a linear regression setting. We provide a Bahadur representation of the proposed estimator and show that the remainder term in the Bahadur representation entails a slower convergence rate compared to traditional SGD due to the adaptive data collection.

Keywords: online inference, stochastic gradient descent, contextual bandits, Bahadur representation, quantile regression

1 Introduction

Following the seminal work of Robbins (1952), the stochastic multi-armed bandit problem has been studied extensively in the literature, where an agent aims to make optimal decisions sequentially

*Stern School of Business, New York University, Email: xc13@stern.nyu.edu

†Committee on Computational and Applied Mathematics, University of Chicago, Email: laizehua@uchicago.edu

‡Stern School of Business, New York University, Email: hli@stern.nyu.edu

§Krannert School of Management, Purdue University, Email: zhang@purdue.edu

among multiple arms and only the selected arm reveals rewards consequently. As the agent's choice is often influenced by additional covariates, also referred to as contexts, contextual bandit problems have gained renewed attention in the past decades (Woodroofe, 1979; Langford and Zhang, 2007, etc.). With the development of internet and data technology, contextual bandit algorithms play an important role in sequential decision-making applications, such as online advertisement (Li et al., 2010), precision medicine (Kim et al., 2011), e-commence (Qiang and Bayati, 2016; Chen et al., 2022), and public policy (Kasy and Sautmann, 2021). Such decisions are often referred to as recommendations, treatments, interventions, and public orders, while the rewards can be healthcare outcomes, welfare utility, revenue as well as any measure of satisfaction of decisions.

Most contextual bandit algorithms are built with the goal of learning the best action under different contexts. In sequential settings, it is often formulated as minimizing the expected cumulative regret that the practitioner would have received if she knows the optimal action. While the importance of this regret minimization is undisputed, *reliable uncertainty quantification* of the learned decision rule is evidently important in many featured applications. For example, in a personalized medicine application where the intervention decision is to choose t''he best medical treatment to optimize some health outcome, the risk for the selected treatment plays a critical and even sometimes life-threatening role in decision-making. Such examples call for the crucial need for a valid and reliable statistical inference procedure accompanying the decision-making process to provide guidance on policy interventions. Inferential studies help not only prompt risk alerts in recommendations, but also gain scientific knowledge of questions such as the effectiveness of medicines.

Particularly, consider a linear contextual bandit environment where the observed data at each decision point t is a triplet $\zeta_t = (X_t, A_t, Y_t)$ for all $t \geq 1$, consisting of covariate X_t, action A_t, and reward $Y_t = X_t^\top \theta^*_{A_t} + \epsilon_t$ where $\theta^*_{A_t} \in \mathbb{R}^d$ is unknown parameters of interest governed by a finite set of actions $\mathcal{A}$, and $\epsilon_t \in \mathbb{R}$ is the noise under certain modeling assumptions. For illustrative simplicity, we consider a binary action space $\mathcal{A} = \{0, 1\}$ corresponding to a duplet of underlying model parameters $(\theta^*_0, \theta^*_1) \in \mathbb{R}^d \times \mathbb{R}^d$, and actions $A_t \in \mathcal{A}$ are selected according to a policy $A_t \sim \pi(X_t, \mathcal{H}_{t-1})$ where $\mathcal{H}_{t-1}$ denotes the trajectory of observations until time $t-1$. At the time t, a typical policy π prefers the action with a higher mean reward $X_t^\top \theta^*_a$ for $a \in \mathcal{A}$, while reserving a

small probability to explore a random action to avoid potential myopic short-sighted exploitation. For example, in the widely-used ε-greedy policy,

$$\mathbb{P}\big(A_t = a \mid X_t, \theta_{0,t-1}, \theta_{1,t-1}\big) = (1-\varepsilon)\mathbb{1}\big\{a = \arg\max_{a\in\mathcal{A}} X_t^\top \theta_{a,t-1}\big\} + \frac{\varepsilon}{2}, \tag{1}$$

This procedure heavily relies on a series of estimators $(\theta_{0,t-1}, \theta_{1,t-1})$ on-the-fly, of the underlying model parameters. Despite that a return-oriented policy would undoubtedly favor the action with a higher reward, it is often as crucial to obtain the confidence of decisions, i.e., conducting statistical inference for (θ_0^*, θ_1^*) in the prescribed applications. This model of statistical inference of model parameters in decision-making problems appears recently in literature (See e.g., Chen, Lu and Song, 2021a; Zhang, Janson and Murphy, 2021, and a brief survey in Section 1.1 below). A typical inferential task provides a confidence interval of the underlying parameters (θ_0^*, θ_1^*) or significance levels when testing hypotheses of parameters, or its margin $\theta_1^* - \theta_0^*$.

Since the sequential decision-making problem relies on updating the estimator for every t throughout the horizon, it is important to provide a computationally efficient fully-online algorithm for both estimation and inference purposes. The existing literature of sequential decision-making mostly focuses on the convergence rate and efficiency, while computational efficiency and storage applicability of the estimation algorithm is often optimistically neglected. As such, they often provide online decision-making procedures governed by an offline scheme of parameter estimation. At each iteration t, an "offline" M-estimator $(\theta_{0,t}, \theta_{1,t})$ is often obtained using the sample path $\{(X_1, y_1),\ (X_1, y_2), \ldots, (X_t, y_t)\}$ up to time t. For example, when using the linear estimator, the computation cost accumulates in a non-scalable manner to at least $\mathcal{O}(T^3)$ over the entire horizon T.

To facilitate computationally efficient online inference, we adopt the stochastic gradient descent (SGD) algorithms in conducting statistical inference in fully-online decision-making. SGD, dated back to Robbins and Monro (1951), has been widely used in large-scale stochastic optimization thanks to its computational and storage efficiency. Its averaged version (ASGD) enables the statistical inference (Polyak and Juditsky, 1992), and inference procedures have been recently analyzed by (Chen et al., 2020; Fang, Xu and Yang, 2018; Lee et al., 2022a, and others). A detailed survey of SGD inference is provided in Section 1.1 below. SGD fits well into the online decision-making

scheme, as the underlying parameter (θ_0^*, θ_1^*) is the solution to the following stochastic optimization under certain modeling assumptions,

$$\theta_a^* \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \mathbb{E}\left[\ell\big(\theta; (X_t, Y_t)\big) \mid A_t = a\right], \quad a \in \mathcal{A}, \tag{2}$$

where the function $\ell \in \mathbb{R}^d \to \mathbb{R}$ will be constructed according to different modeling assumptions. For example, in a linear contextual bandit $Y_t = X_t^\top \theta_{A_t}^* + \epsilon_t$ with *i.i.d.* covariates X_t and mean-zero noise $\{\epsilon_t\}$, a natural choice of $\ell\big(\theta; (X_t, Y_t)\big) = \big(y_t - X_t^\top \theta\big)^2$ is the squared loss. As the outcome Y_t at every time t is adaptively collected upon the decision of action A_t, only one of the $\big(\theta_{0,t}, \theta_{1,t}\big)$ is updated. To compensate missing updates, a generalized SGD updates

$$\theta_t = \theta_{t-1} - \eta_t w_t \nabla \ell\big(\theta_{t-1}; (X_t, Y_t)\big). \tag{3}$$

with a weighting parameter w_t determined by the decision policy $\pi(X_t, \mathcal{H}_{t-1})$. This procedure first appeared in Chen, Lu and Song (2021b) where they used the inverse probability weighting (IPW) for an ε-greedy policy (See (11) below for the explicit form of the IPW). As a consequence, the weighted stochastic gradient $w_t \nabla \ell\big(\theta_{t-1}; (X_t, Y_t)\big)$ is proved to be an unbiased estimator of a weighted population loss function where the weight is independent to the entire the historical information. While the unbiasedness property of stochastic gradient and its independence from the prior trajectory clear the technical difficulty of theoretical analysis of the asymptotic normality of the IPW-weighted ASGD estimator, IPW increases its asymptotic variance by a factor of $1/\varepsilon$. With such a factor, the proposed algorithm leads to a highly-volatile estimator in practice and entails an overly wide confidence interval while making inferential calls. Designing ameliorate decision-making algorithms to enhance the asymptotic efficiency of the estimator remains challenging yet important.

In this paper, we allow a general choice of the weighting parameter w_t in (3), which admits IPW weights as a special case, and derive the explicit formula for the asymptotic distribution of the generalized-weighting ASGD algorithm, thus provides us a way to compare different choices of w_t and even optimize over w_t for some simple models. Our proposed estimator greatly improves the asymptotic efficiency over IPW-ASGD and achieves comparable efficiency as if the practitioner picks one arm steadily. This estimator helps construct narrow yet reliable confidence intervals

for the underlying parameter of interest. The analysis also reveals a recommendation of optimal choices of weights w_t in certain policies. To overcome the technical challenge raised in dependent weighting parameters, we propose a new definition of the loss function, which is different from the loss function used in classical SGD literature (e.g., Chen et al., 2020) and adaptive SGD literature (Chen, Lu and Song, 2021b). We use two parameters θ and θ' to separate the effect of weighting parameters in SGD and that of decision-making procedures in the local geometric landscape of the loss function.

As a separate interest, our framework allows non-smooth loss functions such as quantile loss. In contrast to linear regression, quantile regression provides estimates of a range of conditional quantiles of the reward Y_t. Since contextual bandit problems often appear in an interactive environment, the underlying reward model is more likely to differ across the distribution of the rewards and contexts or involves outliers. Linear regression methods estimate only the mean effects which is usually an incomplete summary of the effect of exposures for certain outcomes. For example, when recommending health care interventions, associations between health care and health outcomes can be highly different among individuals at high-, median-, and low-level utilization of health care. Quantile regression finds ubiquitous applications in many fields such as operations management of business inventory and risk management of financial assets Rockafellar and Uryasev (2002); Ban and Rudin (2019). Therefore, it is worth exploring the use of quantile-based objective functions in sequential decision-making problems. In this paper, we establish a general framework that allows certain nonsmooth objective functions including quantile regression.

We emphasize the technical challenges and summarize the methodology contribution and theoretical advances in the following facets.

- We study the online statistical inference of model parameters in a contextual bandit framework of sequential decision-making. We adopt the existing fully-online re-weighting algorithm for SGD but extend it in two directions: for a general choice of weights and handling non-smooth loss functions via stochastic subgradient. An important example is the quantile loss functions with applications in newsvendor problems and risk management. Moreover, this example provides robustness due to the fact that the objective function is globally Lipschitz. We

establish the asymptotic normality result and characterize how the asymptotic covariance depends on the weight choice.

- We show that SGD under ε-greedy policies with inverse probability weighting (IPW) in Chen, Lu and Song (2021b) suffers from an unbounded asymptotic variance when the exploration rate, ε is close to 0, i.e., the relative efficiency of adaptive models versus non-adaptive models diverges to infinity. Our proposed algorithm features a general policy with a flexible specification of the weights to avoid such deficiency and obtain a bounded relative efficiency. We further provide some practical insights into the optimal weight specification.

- Beyond the asymptotic normality of the proposed estimator, we further establish an analysis of the higher-order remainder term in its Bahadur representation. In classical *i.i.d.* SGD settings, the essential part of the remainder term achieves the rate of $\mathcal{O}\big(t^{-\frac{\alpha}{2}}\big)$, which can be arbitrarily close to the order of the regular offline M-estimator under smooth objectives as α gets close to 1. On the contrary, under the adaptive decision-making environment, the reminder term has a slower rate of $\mathcal{O}\big(t^{-\alpha+\frac{1}{2}}+t^{-\frac{\alpha}{4}}\big)$. As α approaches 1, the remainder term gets closer to $\mathcal{O}\big(n^{-\frac{1}{4}}\big)$, which matches the rate of some nonsmooth M-estimators. This slower rate can be considered as the effect of the discontinuous indicator function for the ε-greedy policy.

1.1 Related works

Online statistical inference for model parameters in SGD The asymptotic distribution of averaged stochastic gradient descent (ASGD) is first given in Ruppert (1988) in Polyak and Juditsky (1992). Since then, there has been a rapid growth of interest recently in conducting statistical inference for model parameters in stochastic gradient algorithms. Chen et al. (2020) proposed two online estimators (plug-in and batch-means) in constructing estimators of limiting covariance matrix of ASGD, of which Zhu, Chen and Wu (2021) extended the batch-means to overlapped batches. Fang, Xu and Yang (2018) proposed a perturbation-based resampling procedure to conduct inference for ASGD. Su and Zhu (2018) proposed a tree-structured inference scheme to construct confidence intervals. Lee et al. (2022a,b) generalized the results in Polyak and Juditsky (1992) to a

functional central limit theorem and proposed an online inference procedure called random-scaling for smooth objectives and quantile regression, respectively.

Statistical inference in online decision-making problems Chen, Lu and Song (2021a) studied the asymptotic distribution of the parameters under a linear contextual bandit framework. Deshpande et al. (2018); Khamaru et al. (2021) considered adaptive linear regression where the vector contexts are correlated over time. Zhang, Janson and Murphy (2021, 2022) conducted statistical inference for M-estimators in contextual bandit and non-Markovian environments. Hao et al. (2019) used multiplier bootstrap to offer uncertainty quantification for exploration in the bandit settings. Chen, Lu and Song (2021b) conducted statistical inference of the model parameters via SGD. There also exists related statistical inference literature in reinforcement learning as a well-known online decision-making setting. Ramprasad et al. (2022) conducted statistical inference for TD (and GTD) learning. Shi et al. (2021) constructed the confidence interval for policy values in Markov decision processes. Shi et al. (2022) conducted statistical inference for confounded Markov decision processes. Chen, Song and Jordan (2022) developed the confidence interval for heterogeneous Markov decision processes.

1.2 Notations and organization of the paper

We first introduce some notations in our paper. For any pair of positive integers $m < n$, we use $[m : n]$ as a shorthand for the discrete set of $\{m, m+1, \ldots, n\}$. For any vector $\theta \in \mathbb{R}^d$, we use $\theta_{[m:n]}$ to denote the vector consisting of the m-th to n-th coordinates of θ. Similarly, $\theta_{[m:n],t}$ is the corresponding subvector of θ_t.

For convenience, let $\|\cdot\|$ denote the standard Euclidean norm for vectors and the spectral norm for matrices. We use the standard Loewner order notation $\Sigma \succeq 0$ if a matrix Σ is positive semi-definite. Denote I_d as the identity matrix in $\mathbb{R}^{d\times d}$. For any square matrix Σ, $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ represent the smallest and the largest eigenvalues, respectively. We also introduce $\mathbb{1}(\cdot)$ for the indicator function, and $\lesssim$ is used for inequalities with omitted constants.

The remainder of the paper is organized as follows. In Section 2, we consider the environment where we collect data adaptively. We describe the weighted version SGD under this setting and

give two illustrative examples of the classical regression problems. In Section 4, we first introduce the technical assumptions before we present the asymptotic distribution for general weighted SGD under this adaptive data collection scheme, along with a comparison on the statistical efficiency to the previous result proposed in Chen, Lu and Song (2021b). We further justify our assumptions under the two illustrative regression examples and show the asymptotic normality for these two cases. Section 4.1 gives the finite-sample rate for our SGD update under adaptive environment. We compare our result with the traditional SGD rate, where the slower rate is due to the adaptively collected data. Simulation studies and real data analyses in Section 5 lend numerical support to the theoretical claims in this paper, which also provides hands-on guidelines to practitioners.

2 Problem Setup

We consider a contextual bandit environment where the observed data at each decision point t is a triplet $\zeta_t = (X_t, A_t, Y_t)$ for all $t \geq 1$, consisting of covariate X_t, action A_t, and reward Y_t. In this paper, we consider a finite action space, i.e., $A_t \in \mathcal{A}$ and $|\mathcal{A}| < \infty$. We assume a stochastic contextual bandit environment in which $\{X_t, Y_t(A) : A \in \mathcal{A}\} \overset{i.i.d}{\sim} \mathcal{P} \in \mathbf{P}$ for all $t \geq 1$. The contextual bandit environment distribution $\mathcal{P}$ is in a space of possible environment distributions $\mathbf{P}$.

We define the trajectory until time t as $\mathcal{H}_t := \{X_s, A_s, Y_s\}_{s=1}^t$ for $t \geq 1$ and $\mathcal{H}_0 := \emptyset$. Actions $A_t \in \mathcal{A}$ are selected according to some policy $A_t \sim \pi(X_t, \mathcal{H}_{t-1})$, which defines action distribution. Even though the covariate reward tuples are *i.i.d.*, the observed data $\{X_t, A_t, Y_t\}_{t\geq 1}$ are not because the actions are selected using policies $\pi(X_t, \mathcal{H}_{t-1})$ which is a function of past data, $\mathcal{H}_{t-1}$. Non-independence of observations is a key property of adaptively collected data.

We are interested in constructing confidence regions for some unknown $\theta^* \in \mathbb{R}^d$. Specifically, we assume that θ^* is a conditionally maximizing value of some loss function $\ell(\theta; \zeta)$, i.e., for $\mathcal{P} \in \mathbf{P}$,

$$\theta^*(\mathcal{P}) \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \, \mathbb{E}_{\mathcal{P}_{Y|X}} \left[\ell(\theta; \zeta) \mid X, A \right]. \tag{4}$$

Note that θ^* does not depend on (X, A) and it is an implicit modeling assumption that such a θ_0 exists for a given $\ell(\theta; \zeta)$. note that under the finite action space setting where $|\mathcal{A}| < \infty$, θ^* is the concatenated vector of all optimal θ for each $A \in \mathcal{A}$, and this implicit assumption is very natural.

For example, in a classical regression setting, a natural choice for $\ell(\theta;\zeta)$ is as follows,

$$\ell(\theta;\zeta_t) = \rho\left(Y_t - X_t^T\theta_{A_t}\right),$$

where $\theta \in \mathbb{R}^d$ is the concatenated vector of $\theta_{A_t} \in \mathbb{R}^p$ for all possible choices of $A_t \in \mathcal{A}$ and $d = p|\mathcal{A}|$. Here $\rho(\cdot)$ is some convex loss function. Under the linear regression case, the $\rho(\cdot)$ is the quadratic loss function and under the quantile regression setting, $\rho_\tau(u) = u(\tau - \mathbb{1}(u < 0))$ with a given $0 < \tau < 1$.

In this paper, we would like to conduct estimation and statistical inference for the unknown θ^* given our adaptively collected data. Let θ_0 denote any given initial estimation, the stochastic gradient descent scheme (SGD) (Robbins and Monro, 1951) iteratively updates the parameter as follows,

$$\theta_t = \theta_{t-1} - \eta_t \nabla\ell(\theta_{t-1};\zeta_t), \tag{5}$$

where η_t is a positive non-increasing sequence referred to as the step-size sequence and $\nabla\ell$ is the gradient for smooth individual loss function ℓ. Note that ℓ can be non-smooth as long as $\nabla\ell$ exists almost surely. For the SGD update above, under the traditional *i.i.d.* setting where $\zeta_t = (X_t, Y_t)$, the classical result by Polyak and Juditsky (1992) uses the average $\bar{\theta}_t^{(\texttt{SGD})} = t^{-1}\sum_{s=0}^{t-1}\theta_s$ as the final estimator to accelerate the estimation. They characterize the limiting distribution and statistical efficiency of the averaged SGD, i.e.,

$$\sqrt{t}\big(\bar{\theta}_t^{(\texttt{SGD})} - \theta^*\big) \Longrightarrow \mathcal{N}\big(0, H^{(\texttt{SGD})-1}S^{(\texttt{SGD})}H^{(\texttt{SGD})-1}\big),$$

given a series of predetermined learning rates $\eta_t = \eta_0 t^{-\alpha}$ for $\eta_0 >$ and $0.5 < \alpha < 1$. Here $H^{(\texttt{SGD})}$ and $S^{(\texttt{SGD})}$ are the Hessian and Gram matrix at $\theta = \theta^*$ for some population loss function under traditional *i.i.d.* setting. The asymptotic covariance $H^{(\texttt{SGD})-1}S^{(\texttt{SGD})}H^{(\texttt{SGD})-1}$ is often known as the "sandwich" covariance structure. For model well-specified settings, this asymptotic covariance matrix matches the inverse Fisher information matrix and thus the resulting averaged estimator $\bar{\theta}_t^{(\texttt{SGD})}$ is asymptotically efficient.

We now provide some popular statistical models as illustrative examples, and we will refer to these examples throughout the paper.

Example 2.1 (Linear Regression)**.** *Consider a two-arm linear contextual bandit problem where*

$$\mathbb{E}[Y_t \mid A_t, X_t] = (1 - A_t)\big(X_t^\top \theta^*_{[1:p]}\big) + A_t\big(X_t^\top \theta^*_{[p+1:2p]}\big),$$

where $\theta^* \in \mathbb{R}^d$ *is the concatenated vector of* $\theta^*_{[1:p]}$ *and* $\theta^*_{[p+1:2p]}$ *and* $\theta^*_{[1:p]} \neq \theta^*_{[p+1:2p]}$, $\{X_t, Y_t(A) : A \in \mathcal{A}\} \overset{i.i.d}{\sim} \mathcal{P} \in \mathbf{P}$ *for all* $t \geq 1$. *The true reward* Y_t *is generated by* $\mathbb{E}[Y_t \mid A_t, X_t] + \mathcal{E}_t$ *where* $\{\mathcal{E}_t\}$ *are* i.i.d. *random error with mean zero and variance* σ^2.

Under linear regression model, our loss function ℓ *is defined as*

$$\ell(\theta; \zeta_t) = \frac{1}{2}(1 - A_t)(Y_t - X_t^\top \theta_{[1:p]})^2 + \frac{1}{2}A_t(Y_t - X_t^\top \theta_{[p+1:2p]})^2.$$

To address the exploration-and-exploitation dilemma, we consider the traditional ε*-greedy policy where the probability of action* A_t *is defined as,*

$$\mathbb{P}(A_t = 0 \mid X_t, \theta_{t-1}) = (1 - \varepsilon)\mathbb{1}\{X_t^\top \theta_{[1:p],t-1} > X_t^\top \theta_{[p+1:2p],t-1}\} + \frac{\varepsilon}{2}, \tag{6}$$

for some constant $\varepsilon \in (0, 1)$. *In practice, the* ε *is often set as some small constant close to zero. Note that this setting can be relaxed to a deterministic sequence* $\{\varepsilon_t\}$ *which converges to some constant* $\varepsilon_\infty \in (0, 1)$.

Example 2.2 (Quantile Regression)**.** *Under the same data generating process in Example 2.1,*

$$\mathbb{E}[Y_t \mid A_t, X_t] = (1 - A_t)X_t^\top \theta^*_{[1:p]} + A_t X_t^\top \theta^*_{[p+1:2p]}.$$

The true reward Y_t *is generated by* $\mathbb{E}[Y_t \mid A_t, X_t] + \mathcal{E}_t$ *where* $\{\mathcal{E}_t\}$ *are* i.i.d. *random error such that,* $\mathbb{P}(\mathcal{E} \leq 0) = \tau$ *for some given quantile level* $\tau \in (0, 1)$. *Now consider a quantile loss such that*

$$\ell(\theta; \zeta_t) = (1 - A_t)\rho_\tau(Y_t - X_t^\top \theta_{[1:p]}) + A_t\rho_\tau(Y_t - X_t^\top \theta_{[p+1:2p]}),$$

where $\rho_\tau(u) = u(\tau - \mathbb{1}(u < 0))$.

3 SGD with weighted stochastic gradients

Under our adaptive data collection scheme, we now consider a generalized version of the vanilla SGD (5). The following is an SGD update with weighted stochastic gradient under adaptively

collected data setting,

$$\theta_t = \theta_{t-1} - \eta_t w_t \nabla \ell(\theta_{t-1}; \zeta_t). \tag{7}$$

Here the gradient weights w_t only depends on the triplet (A_t, X_t, θ_{t-1}). For example, in the previous work of (Chen, Lu and Song, 2021b), $w_t = 1/2\pi\left(X_t, \theta_{t-1}\right)$.

Given our path of $\{\theta_t\}_{t\geq 1}$, we assume the policy $\pi\left(X_t, \mathcal{H}_{t-1}\right)$ depend on the history $\mathcal{H}_{t-1}$ only through θ_{t-1}, our estimator from the latest step, i.e., $A_t \sim \pi\left(X_t, \theta_{t-1}\right)$. One of the common algorithms following this rule is ϵ-greedy. Note that this can be relaxed to $A_t \sim \pi\left(X_t, \Phi_{t-1}\right)$ for some statistic Φ_{t-1} relies on the history $\theta_0, \cdots, \theta_{t-1}$ and it converges to θ^* when θ_t converges to θ^*. Another common algorithm in the classical contextual bandit literature fall under this relaxed setting is Thompson sampling.

To facilitate our analysis of the asymptotic behavior of SGD update (7), we define the function $\mathcal{L}_{\theta'}(\theta)$ as follows,

$$\mathcal{L}_{\theta'}(\theta) = \mathbb{E}_{\mathcal{P}}\left[\mathbb{E}_{\pi(X,\theta')}\left(w(\theta'; X, A)\ell(\theta; X, A, Y) \mid X\right)\right], \tag{8}$$

where $A \sim \pi(X, \theta')$, $\theta', \theta \in \mathbb{R}^d$, and gradient weight w depending on θ', action A and covariate X. Below we will always use the expression $\nabla\mathcal{L}_{\theta'}(\theta)$ to represent the partial gradient of $\mathcal{L}_{\theta'}(\theta)$ with respect to the variable θ, i.e.,

$$\nabla\mathcal{L}_{\theta'}(\theta) = \frac{\partial}{\partial\theta}\mathcal{L}_{\theta'}(\theta) \in \mathbb{R}^d, \quad \nabla^2\mathcal{L}_{\theta'}(\theta) = \frac{\partial^2}{\partial\theta^2}\mathcal{L}_{\theta'}(\theta) \in \mathbb{R}^{d\times d}.$$

Finally, we denote $\xi_{\theta'}(\theta; \zeta)$ as the gap between our weighted SGD update and the population gradient of our loss defined in (8), i.e.,

$$\xi_{\theta'}(\theta; \zeta) = w(\theta'; X, A)\nabla\ell(\theta; \zeta) - \nabla\mathcal{L}_{\theta'}(\theta), \tag{9}$$

By definition, we can easily verify that $w(\theta'; X, A)\nabla\ell(\theta; \zeta)$ is an unbiased estimator of $\nabla\mathcal{L}_{\theta'}(\theta)$, $\mathbb{E}[\xi_{\theta'}(\theta; \zeta)] = 0$.

In the previous work of Chen, Lu and Song (2021b), the loss function is defined with respect to some pre-determined stable policy π_{stable}, i.e.,

$$\tilde{\mathcal{L}}(\theta) = \mathbb{E}_{\mathcal{P}}\left[\mathbb{E}_{\pi_{\text{stable}}}\left(\ell(\theta; X, A, Y) \mid X\right)\right], \tag{10}$$

where $A \sim \pi_{\text{stable}}$ and π_{stable} is a Bernoulli$(1/|\mathcal{A}|)$, uniformly distributed on the action space $\mathcal{A}$. To match the SGD update with the loss function $\tilde{\mathcal{L}}(\cdot)$, they choose the IPW weighted SGD such that $w_t = \frac{\pi_{\text{stable}}}{\pi(X,\theta)}$. This weighting scheme corrects the sampling distribution of the action A_t towards the Bernoulli distribution under the stable policy. However, this definition cannot be extended to a general weighting scheme and the resulting asymptotic covariance matrix could be extremely large as we will see in the discussion after Theorem 4.2 in Section 4. Our framework allows a much broader class of weighting schemes, and our theoretical analysis relies heavily on our definition of the loss function $\mathcal{L}_{\theta'}(\theta)$ in (8). By expressing the loss using two different variables θ and θ', we separate the loss $\ell(\theta;\zeta)$ from the policy $\pi(X,\theta')$ and the weight $w(\theta';X,A)$, as we have a focus on the local geometry of $\ell(\theta;\zeta)$ instead of the local geometry of $\pi(X,\theta')$ and $w(\theta';X,A)$.

We now revisit the two aforementioned motivating examples and illustrate the weighted SGD algorithm for the two models.

Example 2.1 (Continued). *Under the linear regression model, the weighted SGD (7) writes as*

$$\theta_{[1:p],t} = \theta_{[1:p],t-1} - \eta_t w_t X_t (X_t^\top \theta_{[1:p],t-1} - Y_t), \quad A_t = 0;$$
$$\theta_{[p+1:2p],t} = \theta_{[p+1:2p],t-1} - \eta_t w_t X_t (X_t^\top \theta_{[p+1:2p],t-1} - Y_t), \quad A_t = 1.$$

Example 2.2 (Continued). *Under the quantile regression model, the weighted SGD (7) writes as*

$$\theta_{[1:p],t} = \theta_{[1:p],t-1} + \eta_t w_t \left[\tau - \mathbb{1}(Y_t - X_t^\top \theta_{[1:p],t-1} < 0)\right] X_t, \quad A_t = 0;$$
$$\theta_{[p+1:2p],t} = \theta_{[p+1:2p],t-1} + \eta_t w_t \left[\tau - \mathbb{1}(Y_t - X_t^\top \theta_{[p+1:2p],t-1} < 0)\right] X_t, \quad A_t = 1.$$

Some typical choices of gradient weight w_t are the inverse probability weighting (IPW) introduced in Chen, Lu and Song (2021b) which corrects the action distribution to some deterministic stable policy,

$$w_t(A_t, X_t, \theta_{t-1}) = \frac{1}{2\,\mathbb{P}(A_t \mid X_t, \theta_{t-1})}, \tag{11}$$

as well as the square-root importance weights used in Hammersley (2013) and Zhang, Janson and Murphy (2021),

$$w_t(A_t, X_t, \theta_{t-1}) = \sqrt{\frac{1}{2\,\mathbb{P}(A_t \mid X_t, \theta_{t-1})}}. \tag{12}$$

3.1 Optimal weights in linear regression settings

Under the linear model where X is normally distributed (see Remark 4.1), we consider a class of power functions $f_\gamma(\varepsilon) = \varepsilon^\gamma$. This class of weights covers the IPW-type weighted SGD (11) where $\gamma = -1$, the square-root importance weighted SGD (12) where $\gamma = -1/2$, and the vanilla SGD (5) where $\gamma = 0$ (all up to some constants). It is possible to write down the expression for S and H under this setting, i.e.,

$$S = \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix}, \; H = \begin{bmatrix} H_1 & 0 \\ 0 & H_2 \end{bmatrix},$$

where

$$S_1 = \sigma^2 \left[(1-\frac{\varepsilon}{2}) f^2 (1-\frac{\varepsilon}{2}) G_1^* + \frac{\varepsilon}{2} f^2(\frac{\varepsilon}{2}) G_2^* \right], \qquad S_2 = \sigma^2 \left[\frac{\varepsilon}{2} f^2(\frac{\varepsilon}{2}) G_1^* + (1-\frac{\varepsilon}{2}) f^2 (1-\frac{\varepsilon}{2}) G_2^* \right],$$
$$H_1 = (1-\frac{\varepsilon}{2}) f(1-\frac{\varepsilon}{2}) G_1^* + \frac{\varepsilon}{2} f(\frac{\varepsilon}{2}) G_2^*, \qquad H_2 = \frac{\varepsilon}{2} f(\frac{\varepsilon}{2}) G_1^* + (1-\frac{\varepsilon}{2}) f(1-\frac{\varepsilon}{2}) G_2^*,$$

and G_1^*, G_2^* are defined at $\theta' = \theta^*$, i.e.,

$$G_1^* = \Phi(a^*) I_p + \frac{1}{\sqrt{2\pi}} a^* e^{\frac{a^{*2}}{2}} \frac{(\theta^*_{[1:p]} - \theta^*_{[p+1:2p]})(\theta^*_{[1:p]} - \theta^*_{[p+1:2p]})^\top}{\left\| \theta^*_{[1:p]} - \theta^*_{[p+1:2p]} \right\|^2},$$
$$G_2^* = (1 - \Phi(a^*)) I_p - \frac{1}{\sqrt{2\pi}} a^* e^{\frac{a^{*2}}{2}} \frac{(\theta^*_{[1:p]} - \theta^*_{[p+1:2p]})(\theta^*_{[1:p]} - \theta^*_{[p+1:2p]})^\top}{\left\| \theta^*_{[1:p]} - \theta^*_{[p+1:2p]} \right\|^2},$$
$$a^* = \frac{\mu^\top (\theta^*_{[p+1:2p]} - \theta^*_{[1:p]})}{\sqrt{\left\| \theta^*_{[p+1:2p]} - \theta^*_{[1:p]} \right\|^2 + (\mu^\top (\theta^*_{[1:p]} - \theta^*_{[p+1:2p]}))^2}}.$$

Here the existence of G_1^* and G_2^* are assured by the implicit non-degenerate model assumption such that $\theta^*_{[1:p]} \neq \theta^*_{[p+1:2p]}$.

Denote v to be the vector $(\theta^*_{[1:p]} - \theta^*_{[p+1:2p]})/\|\theta^*_{[1:p]} - \theta^*_{[p+1:2p]}\|$. Since H_1, H_2, G_1, G_2 all have the form $bI + cvv^\top$ for some constants b and c, they can be simultaneously diagonalized by a set of basis $v, v_1, \ldots, v_{p-1}$ with corresponding eigenvalues $\{b+c, b, b, \ldots, b\}$, where $\{v_i\}_{i=1}^{p-1}$ can be any $(p-1)$ orthonormal vectors in the $(p-1)$-dimensional space that is orthogonal to v. We can further express $H^{-1}SH^{-1}$ as follows,

$$H^{-1}SH^{-1} = \sigma^2 \begin{bmatrix} c_1 I + c_2 vv^\top & 0 \\ 0 & c_3 I + c_4 vv^\top \end{bmatrix}, \tag{13}$$

where

$$
\begin{aligned}
c_1 &= \frac{(1-\frac{\varepsilon}{2})^{1+2\gamma}\Phi(a) + (\frac{\varepsilon}{2})^{1+2\gamma}(1-\Phi(a))}{((1-\frac{\varepsilon}{2})^{1+\gamma}\Phi(a) + (\frac{\varepsilon}{2})^{1+\gamma}(1-\Phi(a)))^2}, \\
c_2 &= \frac{(1-\frac{\varepsilon}{2})^{1+2\gamma}(\Phi(a) + \frac{1}{\sqrt{2\pi}}ae^{\frac{a^2}{2}}) + (\frac{\varepsilon}{2})^{1+2\gamma}(1-\Phi(a) - \frac{1}{\sqrt{2\pi}}ae^{\frac{a^2}{2}})}{((1-\frac{\varepsilon}{2})^{1+\gamma}(\Phi(a) + \frac{1}{\sqrt{2\pi}}ae^{\frac{a^2}{2}}) + (\frac{\varepsilon}{2})^{1+\gamma}(1-\Phi(a) - \frac{1}{\sqrt{2\pi}}ae^{\frac{a^2}{2}}))^2} - c_1, \\
c_3 &= \frac{(1-\frac{\varepsilon}{2})^{1+2\gamma}(1-\Phi(a)) + (\frac{\varepsilon}{2})^{1+2\gamma}\Phi(a)}{((1-\frac{\varepsilon}{2})^{1+\gamma}(1-\Phi(a)) + (\frac{\varepsilon}{2})^{1+\gamma}\Phi(a))^2}, \\
c_4 &= \frac{(1-\frac{\varepsilon}{2})^{1+2\gamma}(1-\Phi(a) - \frac{1}{\sqrt{2\pi}}ae^{\frac{a^2}{2}}) + (\frac{\varepsilon}{2})^{1+2\gamma}(\Phi(a) + \frac{1}{\sqrt{2\pi}}ae^{\frac{a^2}{2}})}{((1-\frac{\varepsilon}{2})^{1+\gamma}(1-\Phi(a) - \frac{1}{\sqrt{2\pi}}ae^{\frac{a^2}{2}}) + (\frac{\varepsilon}{2})^{1+\gamma}(\Phi(a) + \frac{1}{\sqrt{2\pi}}ae^{\frac{a^2}{2}}))^2} - c_3.
\end{aligned}
$$

The eigenvalues of the asymptotic covariance matrix are $c_1, c_1+c_2, c_3, c_3+c_4$ in the above equations. Notice that this result matches the result in Chen, Lu and Song (2021b), as their IPW-type weighted SGD can be seen as a special case in our framework. In practice, for the ε-greedy policy, the parameter ε is usually taken as some sufficiently small constant close to zero and the covariance of IPW-SGD estimator would be much larger. The proposed algorithm overcomes this problem by allowing a much broader set of weight schemes. It can be inferred from (13) that $\gamma \geq -1/2$ gives a finite covariance matrix even when ε is close to zero. This includes the vanilla SGD (5) and the square-root importance weighted SGD (12) but excludes IPW-SGD. In fact, all the eigenvalues of the asymptotic covariance matrix have the following form with respect to γ, i.e.,

$$
g(\gamma) = \frac{(1-\varepsilon)^{1+2\gamma}b + \varepsilon^{1+2\gamma}(1-b)}{((1-\varepsilon)^{1+\gamma}b + \varepsilon^{1+\gamma}(1-b))^2}, \quad b \in (0,1).
$$

Therefore, the minimum is obtained at $\gamma = 0$ for all $b \in (0,1)$. Therefore, we can conclude that under the settings in Remark 4.1, the vanilla SGD has an asymptotic covariance matrix that is strictly better than any other asymptotic covariance matrix obtained from a power-law weighted scheme, $f(\varepsilon) = \varepsilon^\gamma$. The following Remark 3.1 concludes the above discussion, with detailed derivation relegated to the supplementary material. This analysis can be further extended to general policies and any weighting scheme w under the linear regression setting with respect to different weights.

Remark 3.1 (Optimal weights in linear regression)**.** *Under Assumption 1 to Assumption 5, the vanilla SGD with weight* 1 *has the optimal asymptotic covariance matrix in the linear regression*

setting, i.e., $\Sigma_{\text{val}} \preceq \tilde{\Sigma}$*, where* Σ_{val} *is the asymptotic covariance matrix of vanilla SGD and* $\tilde{\Sigma}$ *is any other asymptotic covariance matrix.*

4 Asymptotic normality

We provide the main theoretical results in this section with detailed proof relegated to the supplementary material. We first introduce some regularity assumptions on the population loss function $\mathcal{L}_{\theta'}(\theta)$, the individual loss function $\ell(\theta;\zeta)$, and the gradient weight $w(\theta';X,A)$.

Assumption 1. *There exists some constants* $\underline{w},\overline{w}$*, such that* $0<\underline{w}<w_t<\overline{w}$ *for all* $t\geq 1$*.*

Assumption 2. *The loss function* $\mathcal{L}_{\theta'}(\theta)$ *is convex with respect to* $\theta\in\mathbb{R}^d$*, continuously differentiable with respect to* $\theta\in\mathbb{R}^d$*, and twice continuously differentiable with respect to* θ *at* θ^**. Moreover, there exists some constants* $\delta,\lambda>0$*, such that* $\langle\nabla\mathcal{L}_\theta(\theta),\theta-\theta^*\rangle>0,\ \forall\theta\neq\theta^*$ *and*

$$\langle\nabla\mathcal{L}_\theta(\theta),\theta-\theta^*\rangle\geq\lambda\|\theta-\theta^*\|^2,\ \ \forall\,\theta\in\{\theta:\|\theta-\theta^*\|\leq\delta\}.$$

Assumption 3. *The Hessian matrix* $\nabla^2\mathcal{L}_{\theta'}(\theta)\in\mathbb{R}^{d\times d}$ *exists for all* $(\theta;\theta')\in\mathbb{R}^d\times\mathbb{R}^d$ *and the Hessian matrix at* $(\theta^*;\theta^*)$ *is positive definite, i.e.,* $H\triangleq\nabla^2\mathcal{L}_{\theta^*}(\theta^*)\succ 0$*. Moreover, the Hessian matrix* $\nabla^2\mathcal{L}_{\theta'}(\theta)$ *is* K*-Lipschitz continuous at* (θ^*,θ^*)*, i.e.,*

$$\left\|\nabla^2\mathcal{L}_{\theta'}(\theta)-\nabla^2\mathcal{L}_{\theta^*}(\theta^*)\right\|\leq K\|\theta-\theta^*\|+K\|\theta'-\theta^*\|,$$

for all (θ,θ') *such that* $\|\theta-\theta^*\|+\|\theta'-\theta^*\|\leq 2\delta$*.*

Assumption 4. *For any action* $A\in\mathcal{A}$ *and covariate* X*, we further assume,*

$$\mathbb{E}_{\mathcal{P}_{Y|X}}\left(\|\nabla\ell(\theta;\zeta)\|^2\mid X,A\right)\leq\phi(X)(1+\|\theta-\theta^*\|^2),$$

for some function $\phi(\cdot)$ *such that* $\mathbb{E}[\phi(X)]=\kappa$ *for some constant* $\kappa>0$*. We also assume the Gram matrix of* $\xi_{\theta'}(\theta;\zeta)$ *at* $(\theta^*;\theta^*)$*,* $S\triangleq\mathbb{E}[\xi_{\theta^*}(\theta^*;\zeta)\xi_{\theta^*}(\theta^*;\zeta)^\top]$*, exists.*

Assumption 5. *Let* $\Delta(X,\theta)=\mathrm{d}_{TV}(\pi(X,\theta),\pi(X,\theta^*))$ *be the total variation distance of* $\pi(X,\theta)$ *and* $\pi(X,\theta^*)$*. For function* $\phi(X)$ *defined in Assumption 4, we have* $\lim_{\theta\to\theta^*}\mathbb{E}_{\mathcal{P}_X}[\Delta(X,\theta)\phi(X)]=0$*,*

$$\lim_{\theta\to\theta^*}\mathbb{E}_{\mathcal{P}_{Y|X}}\left[\|\nabla\ell(\theta;\zeta)-\nabla\ell(\theta^*;\zeta)\|^2\mid X,A\right]=0,\quad\lim_{\theta\to\theta^*}\mathbb{E}_{\mathcal{P}_X}\left[|w(\theta;X,A)-w(\theta^*;X,A)|^2\phi(X)\mid A\right]=0.$$

Assumption 1 is a common assumption on the weights applied to the stochastic gradient, which is used in many adaptive setting literature, e.g., Chen, Lu and Song (2021a), Chen, Lu and Song (2021b), and Zhang, Janson and Murphy (2021). The convexity and continuity on the population loss $\mathcal{L}$ in Assumption 2 is a standard requirement in traditional SGD literature (Polyak and Juditsky, 1992; Chen et al., 2020; Chen, Lu and Song, 2021b; Duchi and Ruan, 2021). We can also find similar arguments in the SGD literature mentioned above for Assumption 2 to Assumption 4, whereas we generalize the previous assumptions on our loss function $\mathcal{L}_\theta(\theta)$ with an extra variable θ'. Assumption 5 further gives some regularity on the function $\phi(\cdot)$ defined in Assumption 4. Later we will further verify our assumptions on the two examples we mentioned above, i.e., the linear regression and the quantile regression. It is noteworthy to mention that, in Assumption 4 and Assumption 5, we only implicitly assume $\nabla\ell$ exists almost surely under $\mathcal{P}_{Y|X}$. Therefore, our assumption is not restricted to smooth loss function ℓ, it also covers many non-smooth statistical problems like quantile regression and robust regression.

In the Remark 4.1 below, we illustrate our definition of $\mathcal{L}_{\theta'}(\theta)$ and our assumptions above using a special case where the covariate X follows a normal distribution.

Remark 4.1. *Under the settings in Example 2.1 with the ε-greedy policy (6). Assume X follows a standard normal distribution, i.e., $X_t \sim \mathcal{N}(\mu, I_p)$. Assume $w_t(A_t, X_t, \theta_{t-1})$ is a function of $\mathcal{P}(A_t \mid X_t, \theta_{t-1})$ for some smooth function $f(\cdot)$, i.e., $w_t(A_t, X_t, \theta_{t-1}) = f(\mathcal{P}(A_t \mid X_t, \theta_{t-1}))$. We have for any ε,*

$$\mathcal{L}_{\theta'}(\theta) = (\theta^* - \theta)^\top G(\theta^* - \theta) + \frac{\sigma^2}{2}\left[(1-\frac{\varepsilon}{2})f(1-\frac{\varepsilon}{2}) + \frac{\varepsilon}{2}f(\frac{\varepsilon}{2})\right],$$

where we denote $\Phi(\cdot)$ as the c.d.f. for standard normal distribution and

$$G = \begin{bmatrix} (1-\frac{\varepsilon}{2})f(1-\frac{\varepsilon}{2})G_1 + \frac{\varepsilon}{2}f(\frac{\varepsilon}{2})G_2 & 0 \\ 0 & (1-\frac{\varepsilon}{2})f(1-\frac{\varepsilon}{2})G_2 + \frac{\varepsilon}{2}f(\frac{\varepsilon}{2})G_1 \end{bmatrix}.$$

$$G_1 = \Phi(a)\, I_p + \frac{1}{\sqrt{2\pi}} a e^{\frac{a^2}{2}} \nu'\nu'^\top, \quad G_2 = (1-\Phi(a))\, I_p - \frac{1}{\sqrt{2\pi}} a e^{\frac{a^2}{2}} \nu'\nu'^\top,$$

where ν' is the normalized margin between the two arms of θ',

$$\nu' = (\theta'_{[1:p]} - \theta'_{[p+1:2p]}) / \left\|\theta'_{[1:p]} - \theta'_{[p+1:2p]}\right\|, \quad \text{and} \quad a = \frac{\mu^\top \nu'}{\sqrt{1 + (\mu^\top \nu')^2}}.$$

We now state our first main result that characterizes the limiting distribution of the averaged weighted SGD iterates defined in (7).

Theorem 4.2. *Under Assumption 1 to Assumption 5, the averaged SGD estimator $\bar{\theta}_t$ converges θ^* almost surely when $t \to \infty$ and*

$$\sqrt{t}(\bar{\theta}_t - \theta^*) \to \mathcal{N}(0, H^{-1}SH^{-1}),$$

where $H = \nabla^2 \mathcal{L}_{\theta^*}(\theta^*)$ and $S = \mathbb{E}[\xi_{\theta^*}(\theta^*;\zeta)\xi_{\theta^*}(\theta^*;\zeta)^\top]$.

We relegate the proof of Theorem 4.2 to Section A of the supplementary materials. To emphasize the technical challenge in the theoretical analysis, our loss function $\mathcal{L}$ in (8) is not defined by the stable policy as in the prior works (Chen, Lu and Song, 2021b). The action $A_t \sim \pi(X_t, \theta_{t-1})$ and $A^* \sim \pi(X_t, \theta^*)$ are no longer in the same probability space, and therefore we specify a coupling between A_t and A^* to compare them. A natural choice is the coupling such that

$$\Delta(X, \theta) = d_{TV}(\pi(X, \theta), \pi(X, \theta^*)) = \frac{1}{2}\sum_{i=1}^{|\mathcal{A}|} |p_i - q_i| = \mathcal{P}(A \neq A^*), \tag{14}$$

where $p_i = \mathbb{P}(A = A_i), q_i = \mathbb{P}(A^* = A_i)$.

To further illustrate our assumptions and central limit theorem result in Theorem 4.2, we validate them under two examples we mentioned above, i.e., linear regression (Example 2.1) and quantile regression (Example 2.2). We will show that under ε-greedy policy defined in Equation (6), Theorem 4.2 holds for these two cases.

In Corollary 4.3 below, we demonstrate that Assumptions 1–5 are quite natural and can be satisfied by the linear regression example we discussed in Example 2.1.

Corollary 4.3. *Consider the linear setting defined in Example 2.1, assume that*

(a) The covariate X has finite $\mathbb{E}_{\mathcal{P}_X}\|X\|^4$ and $\mathbb{E}_{\mathcal{P}_X}[XX^\top] \succ 0$;

(b) The probability density function of X, $p(x)$, is smooth and $\int_{x^\top \theta^*_{[1:p]} = x^\top \theta^*_{[p+1:2p]}} x \otimes x \otimes x p(x)\,\mathrm{d}x$ exists;

(c) *For the gradient weights* $\{w_t\}$ *defined in* (7)*, assume* $w_t(A_t, X_t, \theta_{t-1})$ *is a function of* $\mathbb{P}(A_t \mid X_t, \theta_{t-1})$*, i.e.,*

$$w_t(A_t, X_t, \theta_{t-1}) = f(\mathbb{P}(A_t \mid X_t, \theta_{t-1})),$$

for some Lipschitz continuous function $f(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ *and* $f(\cdot)$ *is positive and bounded within interval* $(0, 1)$*.*

Under the above conditions, Assumptions 1–5 are satisfied and the asymptotic normality in Theorem 4.2 exists.

As discussed earlier, our assumption allows a much broader setting than the class of smooth individual loss functions. Under our assumptions, the individual loss function $\ell(\theta; \zeta)$ can be non-smooth. We will justify this argument in the quantile regression example below.

Corollary 4.4. *Consider the quantile regression setting defined in Example 2.2, assume that*

(a) *The covariate* X *has finite* $\mathbb{E}_{\mathcal{P}_X}[XX^\top]$ *and* $\mathbb{E}_{\mathcal{P}_X}[XX^\top] \succ 0$*;*

(b) *The p.d.f. of* X*, denoted as* $p(x)$*, is smooth and* $\int_{x^\top \theta^*_{[1:p]} = x^\top \theta^*_{[p+1:2p]}} x \otimes x \otimes xp(x)\, \mathrm{d}x$ *exists;*

(c) *The p.d.f. of* $\mathcal{E}$*, denoted as* $q(x)$*, is smooth and bounded. Also,* $q(0) > 0$ *and* $q'(x)$ *is bounded;*

(d) *For the gradient weights* $\{w_t\}$ *defined in* (7)*, assume* $w_t(A_t, X_t, \theta_{t-1})$ *is a function of* $\mathbb{P}(A_t \mid X_t, \theta_{t-1})$*, i.e.,*

$$w_t(A_t, X_t, \theta_{t-1}) = f(\mathbb{P}(A_t \mid X_t, \theta_{t-1})),$$

for some Lipschitz continuous function $f(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ *and* $f(\cdot)$ *is positive and bounded within interval* $(0, 1)$*.*

Under above conditions, the Assumption 1 to Assumption 5 are satisfied and the asymptotic normality in Theorem 4.2 exists.

Corollary 4.4 states that we can also obtain the limiting distribution for some non-smooth loss functions like a quantile loss.

Remark 4.5. *In Corollary 4.3 and Corollary 4.4 above, we use the ε-greedy policy with fixed constant $\varepsilon \in (0,1)$ throughout the whole SGD process. This policy can be relaxed to a general ε_t-greedy policy, for some deterministic sequence $\{\varepsilon_t\}$ varying with respect to time t, such that $\varepsilon_t \in (0,1)$ and $\varepsilon_t \to \varepsilon_\infty$. The asymptotic normality result also holds under this setting, we defer the discussion and technical details to Section D in the supplementary material.*

In order to provide statistical inference for the model parameter, we need to estimate the variance of $\widehat{\theta}_t$, $H^{-1}SH^{-1}$, as we established in Theorem 4.2, in a fully online fashion. A few options have been provided from SGD inference literature, e.g., the plug-in estimator (Chen et al., 2020; Chen, Lu and Song, 2021b), the batch-means estimator (Chen et al., 2020; Zhu, Chen and Wu, 2021), the bootstrap estimator (Fang, Xu and Yang, 2018), the random scaling estimator (Lee et al., 2022a). Among the above, the plug-in estimator is expected to achieve the best numerical behavior as evident from classical SGD approaches. In this paper, we use the plug-in estimator Chen et al. (2020) for smooth loss functions ℓ, and leave the other methods as an interesting future work. In adaptive settings, the online plugin estimators for S and H are given by,

$$\widehat{S}_n = \frac{1}{n}\sum_{t=1}^{n} w_t^2 \nabla \ell(\theta_{t-1};\zeta_t)\nabla \ell(\theta_{t-1};\zeta_t)^\top, \quad \widehat{H}_n = \frac{1}{n}\sum_{t=1}^{n} w_t \nabla^2 \ell(\theta_{t-1};\zeta_t).$$

With the plug-in estimators $(\widehat{S}_t, \widehat{H}_t)$, an online plug-in inference procedure can be provided by replacing S and H in the asymptotic covariance matrix in Theorem 4.2 to $(\widehat{S}_t, \widehat{H}_t)$. We defer the detailed procedure to Section 5.2 below and the consistency proof of these estimators to Section E of the supplementary material.

4.1 Bahadur representations

In this section, we present the Bahadur representation of our weighted SGD update (7) under the adaptive data collection environment. Specifically, under classical non-adaptive SGD settings (5), the Bahadur representation is established in Polyak and Juditsky (1992) as,

$$\sqrt{t}\Sigma^{-1/2}(\bar{\theta}_t - \theta^*) = W + R_t,$$

where W is the leading term in the central limit theorem, i.e., W is a weighted sum of *i.i.d.* random variables, it converges to a standard normal distribution as $t \to \infty$. The R_t term is the remainder

term which converges faster than the leading term W under common regularity conditions.

Theorem 4.6. *Under the conditions in Theorem 4.2 and ε-greedy algorithm defined in (6), we further assume:*

(a) There exists constant $C_1 > 0$, such that $\int_{x^\top \theta_{[1:p]} = x^\top \theta_{[p+1:2p]}} x \otimes x \otimes xp(x)\,\mathrm{d}x \le C_1$ for all θ;

(b) Given θ, θ^*, the following inequality holds for some constant $C_2 > 0$,

$$\mathbb{E}\left[\left|\mathbb{1}\left(X^\top \theta^*_{[1:p]} > X^\top \theta^*_{[p+1:2p]}\right) - \mathbb{1}\left(X^\top \theta_{[1:p]} > X^\top \theta_{[p+1:2p]}\right)\right| (1 + \|X\|^4)\right] \le C_2 \|\theta - \theta^*\|.$$

We have the following decomposition

$$\begin{aligned}
\sqrt{t}\Sigma^{-1/2}(\bar{\theta}_t - \theta^*) = & \underbrace{\frac{1}{\sqrt{t}} \sum_{i=1}^{t-1} \Sigma_t^{-1/2} Q_i^t \xi_{\theta^*}(\theta^*; \zeta_i)}_{W} + \underbrace{\frac{1}{\sqrt{t}} \sum_{i=1}^{t-1} \Sigma^{-1/2} Q_i^t (\xi_{\theta_{i-1}}(\theta_{i-1}; \zeta_i) - \xi_{\theta^*}(\theta^*; \zeta_i))}_{R_1} \\
& + \underbrace{\frac{1}{\sqrt{t}\eta_0} \Sigma^{-1/2} Q_0^t (\theta_0 - \theta^*)}_{R_2} + \underbrace{\frac{1}{\sqrt{t}} \sum_{i=1}^{t-1} \Sigma^{-1/2} Q_i^t (\mathcal{L}_{\theta_i}(\theta_i) - H(\theta_i - \theta^*))}_{R_3} \\
& + \underbrace{\frac{1}{\sqrt{t}} \sum_{i=1}^{t-1} (\Sigma^{-1/2} - \Sigma_t^{-1/2}) Q_i^t \xi_{\theta^*}(\theta^*; \zeta_i)}_{R_4} \\
= & W + R_1 + R_2 + R_3 + R_4, \qquad (15)
\end{aligned}$$

where $\mathbb{E}[W] = 0, \mathbb{E}[WW^\top] = I_d$, $\Sigma_t = \frac{1}{t}\sum_{i=1}^{t-1} Q_i^t S Q_i^t$, and $Q_i^t = \eta_i \sum_{j=i}^{t-1} \prod_{k=i+1}^{j} (I_d - \eta_k H)$ for $t > 0$. Furthermore, we have,

$$\mathbb{E}\|R_1\|^2 \lesssim t^{-\frac{\alpha}{2}}, \quad \mathbb{E}\|R_2\|^2 \lesssim t^{-1}, \quad \mathbb{E}\|R_3\| \lesssim t^{-\alpha+\frac{1}{2}}, \quad \mathbb{E}\|R_4\|^2 \lesssim t^{2\alpha-2}.$$

In Theorem 4.6, the remainder term is decomposed into four terms. Here R_1 is an accumulated error produced by the initialization via stochastic gradient approximation in each iteration (i.e., the difference between the stochastic gradient evaluated at θ_t and θ^*). The term R_2 is the deviation directly produced by the arbitrary initialization θ_0. The term R_3 characterizes the quadratic approximation for a general loss function. Indeed, when the loss function is quadratic (e.g., linear regression in Example 2.1), we have $R_3 = 0$ since the population Hessian matrix is identical for any

θ and the loss function $\mathcal{L}$ is completely characterized by the multiplication of H and $\theta - \theta^*$. The term R_4 is a non-asymptotic compensation to the main asymptotic normal approximation term W, i.e., $W + R_4$ is the non-asymptotic error when the algorithm is initialized at the truth θ^* for a quadratic model.

We defer the proof details to Section C of the supplementary material. Note that to derive the above decomposition, we require a slightly stronger condition (condition (a) in the theorem statement) compared with condition (b) in Corollary 4.3. The second condition in Theorem 4.6 requires a certain level of continuity of the distribution of covariate X. These extra conditions can be easily satisfied, e.g., when X obeys a non-degenerate normal distribution. To characterize the decomposition, we need a generalization of the coupling we defined in Equation (14). Now let us consider the $(|\mathcal{A}| - 1)$-simplex $S = \{(x_1, \ldots, x_{|\mathcal{A}|}) \mid x_i \geq 0, \sum x_i = 1\}$. It has $|\mathcal{A}|$ vertices given by $V_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ where 1 is in the i-th coordinate. Take a point P uniformly from S. For any categorical distribution with probability $(p_1, \ldots, p_{|\mathcal{A}|})$, define $K = (p_1, \ldots, p_{|\mathcal{A}|})$. The probability that P lies in the sub-simplex with vertices $\{V_1, \ldots, \widehat{V_i}, \ldots, V_{|\mathcal{A}|}, K\}$ (V_i is deleted) is exactly p_i. Thus, K gives a partition of S that has the required categorical distribution and we can use this to define the action A. Furthermore, given two different distributions K, K', it is easy to see that the quantity $\mathcal{P}(A \neq A')$ is bounded by $Cd_{TV}(K, K')$, where C is some constant which only depends on $|\mathcal{A}|$.

Given the characterization of the remainder terms of the Bahadur representation of $\bar{\theta}_t$, we now emphasize the difference in the convergence rate of the adaptive SGD and the classical SGD results (Polyak and Juditsky, 1992; Shao and Zhang, 2022). The difference appears only in the term R_1. For the classical SGD, the corresponding term R_1 satisfies $\mathbb{E}\|R_1\|^2 \lesssim t^{-\alpha}$, whereas for the adaptive SGD, we have $\mathbb{E}\|R_1\|^2 \lesssim t^{-\frac{\alpha}{2}}$. This slower convergence rate is caused by our adaptive data collection scheme. As a consequence, under our setting, the remainder term has a rate of $\mathcal{O}\left(t^{-\alpha+\frac{1}{2}} + t^{-\frac{\alpha}{4}} + t^{\alpha-1}\right)$. Minimizing the order of the rate over $\alpha \in (\frac{1}{2}, 1)$, we have that the optimal convergence rate of the remainder term is $\mathcal{O}(t^{-0.2})$ where $\alpha = 0.8$.

Remark 4.7. *Under the conditions in Corollary 4.3 and ε-greedy algorithm defined in (6), assume that the distribution of X non-degenerated normal. Then assumptions (a)(b) of Theorem 4.6 hold.*

Note that $R_3 = 0$ in the linear regression setting and R_2 depends on the initialization of the SGD algorithm. We can indeed establish a lower bound for R_1 as $\mathbb{E}\|R_1\|^2 \gtrsim t^{-\frac{1}{2}}$ (the detailed proof is provided in Section C of the supplementary material). The lower bound on R_1 does not match the upper bound established in Theorem 4.6 but it guarantees a strictly slower convergence than the classical SGD setting where $\mathbb{E}\|\tilde{R}_1\|^2 \lesssim t^{-\alpha}$, as we assume $\alpha \in (1/2, 1)$.

5 Numerical Experiments

In this section, we investigate the empirical performance of the proposed estimators and their performance on normal approximation. We further construct the confidence intervals using a plug-in estimator of the asymptotic covariance matrices and report their coverage rates. The performance of the proposed estimation and inference is also validated on a real dataset.

5.1 Normal approximation

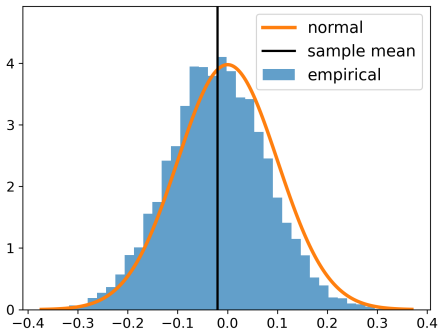
We verify Theorem 4.2 under linear regression and quantile regression (Example 2.1 and Example 2.2). For both examples, we have $\theta^* \in \mathbb{R}^{20}$ and

$$Y_t = (1 - A_t)X_t^\top \theta^*_{[1:10]} + A_t X_t^\top \theta^*_{[11:20]} + \mathcal{E}_t.$$

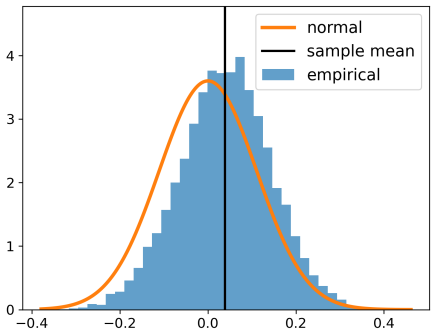
In our numerical experiments below, the sample size is fixed as $80,000$. The covariate $X_t \sim \mathcal{N}(0, I_{10})$ and the error term $\{\mathcal{E}_s\}_{s=1}^t$ is an *i.i.d.* sample with standard deviation $\sigma = 0.1$. We use ε-greedy policy (6) to select actions, and set $\varepsilon = 0.02$.

For the SGD update (7), we specify the step sizes as $\eta_t = \eta * \max(t, 300)^{-\alpha}$. As indicated in Theorem 4.6, the optimal value for the parameter α in the step size should be $\alpha = 0.8$. We specify $\alpha = 0.8$ for both linear regression and quantile regression. We compare three weighting schemes below, vanilla SGD (5), square-root IPW SGD (12), and IPW SGD (11).

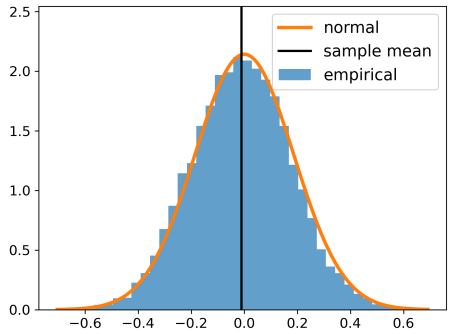
We first present the results for linear regression, where the error term $\mathcal{E} \sim \mathcal{N}(0, \sigma^2)$. In Figure 1 below, we plot the empirical distribution of $\sqrt{t}(\bar{\theta}_t - \theta^*)$ using $10,000$ Monte-Carlo simulations. As can be inferred from the plots, the vanilla SGD and the square-root importance weight SGD have much smaller standard deviation compared with IPW SGD, this finding matches our discussion in



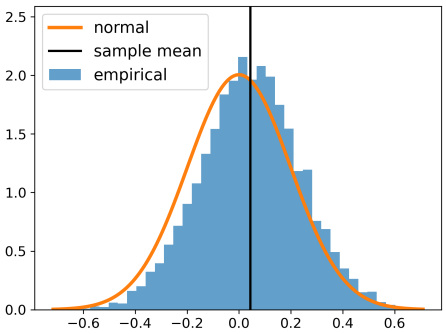
(a) vanilla arm 0



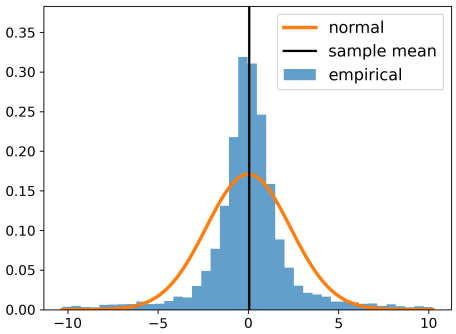
(b) vanilla arm 1



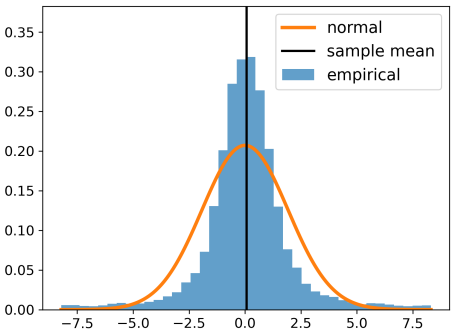
(c) square-root importance arm 0



(d) square-root importance arm 1

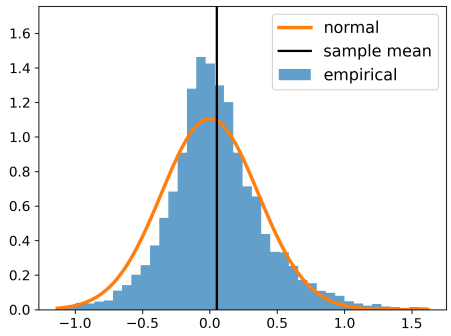


(e) IPW arm 0

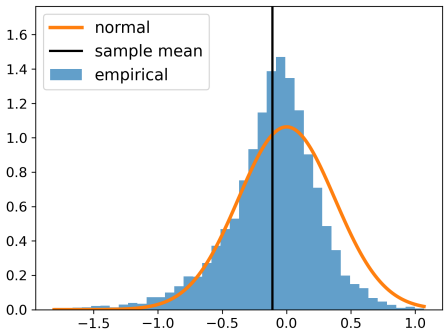


(f) IPW arm 1

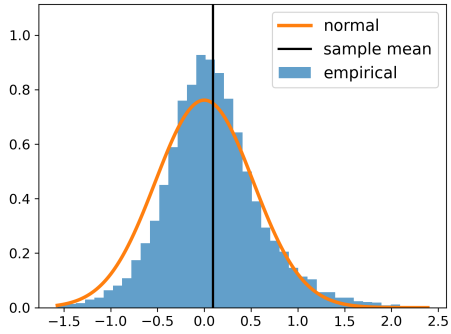
Figure 1: SGD on linear regression with different weights. We report the empirical distribution (one dimension for each arm) of $\sqrt{t}(\bar{\theta}_t - \theta^*)$ for $10,000$ Monte-Carlo simulations. We also plot the density function of a zero-mean normal distribution that match the second-order moments.



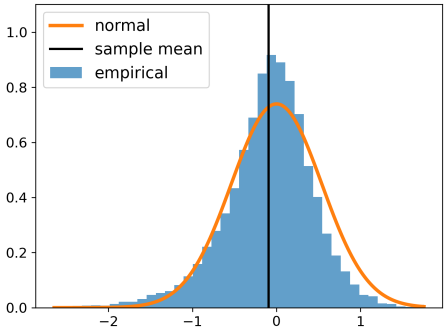
(a) vanilla arm 0



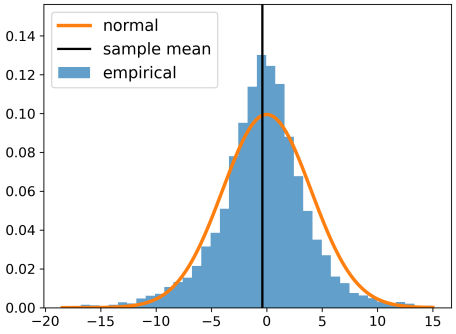
(b) vanilla arm 1



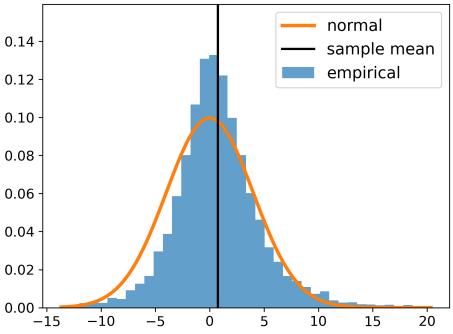
(c) square-root IPW arm 0



(d) square-root IPW arm 1



(e) IPW arm 0



(f) IPW arm 1

Figure 2: SGD on quantile regression with different weights. We report the empirical distribution (one dimension for each arm) of $\sqrt{t}(\bar{\theta}_t - \theta^*)$ for $10,000$ Monte-Carlo simulations. We also plot the density function of a zero-mean normal distribution that match the second-order moments.

Section 3.

We also conduct simulations on quantile regression in Figure 2 below with quantile level $\tau = 0.75$, and the error term has standard deviation $\sigma = 0.1$ and $\mathbb{P}(\mathcal{E} \leq 0) = \tau$.

5.2 Online statistical inference

In this section, we demonstrate the online plug-in inference procedure based on the limiting distribution of our proposed estimator $\bar{\theta}_t$ in Theorem 4.2. As we mentioned in the previous section, the plug-in estimator constructs a pair $(\widehat{S}_n, \widehat{H}_n)$ to estimate (S, H) in the asymptotic covariance matrix $H^{-1}SH^{-1}$.

$$\widehat{S}_n = \frac{1}{n}\sum_{t=1}^{n} w_t^2 \nabla \ell(\theta_{t-1}; \zeta_t) \nabla \ell(\theta_{t-1}; \zeta_t)^\top, \quad \widehat{H}_n = \frac{1}{n}\sum_{t=1}^{n} w_t \nabla^2 \ell(\theta_{t-1}; \zeta_t).$$

The consistency of the plugin estimator is established under the following additional assumption, which can be easily verified for the linear regression example.

Assumption 6. *For any action $A \in \mathcal{A}$ and covariate X, we assume that $\nabla^2 \ell(\theta; \zeta)$ exists and $\mathbb{E}_{\mathcal{P}_{Y|X}}\left(\|\nabla^2 \ell(\theta; \zeta)\|^2 \mid X, A\right)$ is bounded by $\psi(X)(1 + \|\theta - \theta^*\|^2)$ for some function $\psi(\cdot)$ such that $\mathbb{E}[\psi(X)] < \infty$. We have,*

$$\begin{aligned}
&\lim_{\theta \to \theta^*} \mathbb{E}_{\mathcal{P}_X}[\Delta(X, \theta)\psi(X)] = 0, \\
&\lim_{\theta \to \theta^*} \mathbb{E}_{\mathcal{P}_{Y|X}}\left[\|\nabla^2 \ell(\theta; \zeta) - \nabla^2 \ell(\theta^*; \zeta)\|^2 \mid X, A\right] = 0, \\
&\lim_{\theta \to \theta^*} \mathbb{E}_{\mathcal{P}_X}\left[|w(\theta; X, A) - w(\theta^*; X, A)|^2 \psi(X) \mid A\right] = 0.
\end{aligned}$$

Proposition 5.1. *Under Assumption 1 to Assumption 6, the plug-in estimators are consistent, i.e., $\widehat{S}_n \to S$ and $\widehat{H}_n \to H$ in probability.*

The proof is presented in Section E of the supplementary materials. Under the same setting as in Section 5.1, we show the inference results for linear regression in Table 1. The comparison of the three candidate weighted-SGD schemes is clearly stated. Both the vanilla method and sqrt-IPW provide a valid conference interval, while IPW provides a much wider confidence interval than its oracle.

Weight & Arm	Sample size	Plug-in Cov.	Oracle Cov.	Plug-in Len.	Oracle Len.
vanilla Arm 0	2×10^4	0.78 (0.14)	0.73 (0.15)	0.63 (0.03)	0.55
	8×10^4	0.88 (0.09)	0.86 (0.09)	0.57 (0.01)	0.55
vanilla Arm 1	2×10^4	0.89 (0.09)	0.83 (0.12)	0.63 (0.03)	0.55
	8×10^4	0.94 (0.07)	0.93 (0.08)	0.58 (0.01)	0.55
sqrt-IPW Arm 0	2×10^4	0.78 (0.14)	0.72 (0.15)	0.82 (0.12)	0.72
	8×10^4	0.88 (0.10)	0.87 (0.11)	0.74 (0.04)	0.72
sqrt-IPW Arm 1	2×10^4	0.84 (0.12)	0.78 (0.14)	0.83 (0.13)	0.72
	8×10^4	0.91 (0.09)	0.90 (0.10)	0.75 (0.05)	0.72
IPW Arm 0	2×10^4	0.81 (0.15)	0.47 (0.32)	19.18 (34.94)	2.79
	8×10^4	0.85 (0.14)	0.62 (0.33)	13.04 (28.04)	2.79
IPW Arm 1	2×10^4	0.82 (0.15)	0.51 (0.32)	16.76 (32.12)	2.79
	8×10^4	0.86 (0.13)	0.65 (0.32)	11.47 (25.80)	2.79

Table 1: Inference results of linear regression with different weighting schemes. Averaged coverage rate and average length of the confidence intervals are reported for plug-in estimator and oracle estimator. We also include standard error in the parentheses.

5.3 Real data analysis

In this section, we apply our online estimation and inference framework to Yahoo! Today module user click-log dataset and conduct statistical inference for model parameters. We use the news recommendation and user response records on May 1$^{\text{st}}$, 2009. On this day, we consider the two most recommended (recommended $405,888$ times) articles, No.109510 and No.109520 for analysis.

We follow the experiment settings in Chen, Lu and Song (2021b). The action A_t is specified to be 1 when Article No.109510 is recommended and $A_t = 0$ when Article No.109520 is recommended. The original user features have six covariates, where the first five sum up to one, and the sixth is a constant 1. In our experiments below, we keep the second to fifth covariates in the original features as $X_{[2:5]}$ and specify $X_{[1]} = 1$ as the intercept.

As the reward Y_t is binary, we consider a logistic regression model and set $Y_t = 1$ if the user clicks on the article link and $Y_t = -1$ if not. The logistic loss is defined as follows,

$$\ell(\theta; \zeta_t) = (1 - A_t) \log \left\{1 + \exp\left(-y_t(X_t^\top \theta_{[1:5]})\right)\right\} + A_t \log \left\{1 + \exp\left(-y_t(X_t^\top \theta_{[6:10]})\right)\right\}. \tag{16}$$

Under the weighted SGD setting (7), we have

$$\theta_{[1:5],t} = \theta_{[1:5],t-1} + \eta_t w_t \left[1 + \exp\left(y_t X_t^\top \theta_{[1:5],t-1}\right)\right]^{-1} y_t X_t, \quad A_t = 0;$$
$$\theta_{[6:10],t} = \theta_{[6:10],t-1} + \eta_t w_t \left[1 + \exp\left(y_t X_t^\top \theta_{[6:10],t-1}\right)\right]^{-1} y_t X_t, \quad A_t = 1.$$

We use the ε-greedy algorithm (6). In order to match our online decision-making process with our offline dataset, we keep the entry if the recorded offline action matches the action given by our online ε-greedy algorithm with two specifications of $\varepsilon \in \{0.2, 0.02\}$.

We now present the online statistical inference results. For our SGD update, we use the same settings as above experiments, i.e., 300-step meltdown and $\alpha = 0.8$. We compare three weighting schemes below, vanilla SGD (5), square-root importance weight SGD (12), and IPW SGD (11). Table 2 below gives the result for $\varepsilon = 0.2$ and Table 3 gives the result for $\varepsilon = 0.02$. In both table, the vanilla SGD and the square-root importance SGD have smaller standard errors and smaller p-values. There are also more insignificant parameters for IPW SGD. The results of IPW SGD are worse when we decrease the value of ε, matches our findings in Theorem 4.2 and discussions in Section 3.

Weight & Arm	Parameter	Estimate	S.E.	95% LB	95% UB	t-value	p-value
vanilla Arm 0	θ_1	-2.56	0.04	-2.64	-2.48	-65.52	0.00
	θ_2	-0.26	0.08	-0.43	-0.1	-3.11	0.00
	θ_3	-0.48	0.07	-0.62	-0.34	-6.8	0.00
	θ_4	-0.23	0.06	-0.34	-0.12	-4.09	0.00
	θ_5	-0.9	0.07	-1.03	-0.77	-13.65	0.00
vanilla Arm 1	θ_6	-2.55	0.05	-2.65	-2.44	-47.77	0.00
	θ_7	-0.24	0.08	-0.4	-0.09	-3.06	0.00
	θ_8	-0.45	0.07	-0.58	-0.32	-6.76	0.00
	θ_9	-0.41	0.11	-0.62	-0.19	-3.71	0.00
	θ_{10}	-0.91	0.07	-1.05	-0.77	-12.31	0.00
sqrt-IPW Arm 0	θ_1	-2.52	0.05	-2.62	-2.43	-52.85	0.00
	θ_2	-0.3	0.11	-0.51	-0.09	-2.79	0.01
	θ_3	-0.49	0.09	-0.66	-0.31	-5.56	0.00
	θ_4	-0.28	0.07	-0.4	-0.15	-4.25	0.00
	θ_5	-0.8	0.09	-0.97	-0.63	-9.33	0.00
sqrt-IPW Arm 1	θ_6	-2.51	0.05	-2.61	-2.41	-49.35	0.00
	θ_7	-0.28	0.08	-0.43	-0.13	-3.6	0.00
	θ_8	-0.45	0.06	-0.58	-0.33	-7.1	0.00
	θ_9	-0.42	0.11	-0.63	-0.2	-3.83	0.00
	θ_{10}	-0.81	0.07	-0.94	-0.68	-12.02	0.00
IPW Arm 1	θ_1	-2.64	0.1	-2.85	-2.44	-25.54	0.00
	θ_2	-0.28	0.19	-0.64	0.08	-1.51	0.13
	θ_3	-0.51	0.15	-0.8	-0.23	-3.49	0.00
	θ_4	-0.24	0.16	-0.55	0.07	-1.54	0.12
	θ_5	-0.91	0.16	-1.23	-0.59	-5.64	0.00
IPW Arm 1	θ_6	-2.47	0.03	-2.53	-2.4	-76.6	0.00
	θ_7	-0.22	0.06	-0.33	-0.11	-3.83	0.00
	θ_8	-0.51	0.05	-0.6	-0.42	-11.08	0.00
	θ_9	-0.37	0.05	-0.47	-0.27	-7.4	0.00
	θ_{10}	-0.88	0.05	-0.98	-0.78	-17.67	0.00

Table 2: Real data analysis with online statistic inference. We use ε-greedy algorithm with $\varepsilon = 0.2$.

Weight & Arm	Parameter	Estimate	S.E.	95% LB	95% UB	t-value	p-value
vanilla Arm 0	θ_1	-2.55	0.04	-2.63	-2.48	-68.62	0.00
	θ_2	-0.31	0.09	-0.47	-0.14	-3.61	0.00
	θ_3	-0.45	0.07	-0.6	-0.31	-6.18	0.00
	θ_4	-0.23	0.05	-0.33	-0.12	-4.29	0.00
	θ_5	-0.88	0.07	-1.01	-0.75	-13.45	0.00
vanilla Arm 1	θ_6	-2.54	0.06	-2.66	-2.42	-41.76	0.00
	θ_7	-0.29	0.09	-0.45	-0.12	-3.36	0.00
	θ_8	-0.42	0.07	-0.57	-0.28	-5.88	0.00
	θ_9	-0.42	0.19	-0.79	-0.04	-2.18	0.03
	θ_{10}	-0.89	0.08	-1.04	-0.73	-11.25	0.00
sqrt-IPW Arm 0	θ_1	-2.49	0.05	-2.58	-2.4	-54.74	0.00
	θ_2	-0.31	0.13	-0.57	-0.05	-2.37	0.02
	θ_3	-0.45	0.12	-0.68	-0.21	-3.74	0.00
	θ_4	-0.29	0.06	-0.41	-0.17	-4.78	0.00
	θ_5	-0.82	0.08	-0.98	-0.66	-9.8	0.00
sqrt-IPW Arm 1	θ_6	-2.48	0.08	-2.64	-2.33	-31.13	0.00
	θ_7	-0.29	0.1	-0.5	-0.09	-2.84	0.00
	θ_8	-0.42	0.09	-0.6	-0.25	-4.69	0.00
	θ_9	-0.4	0.25	-0.9	0.09	-1.6	0.11
	θ_{10}	-0.82	0.1	-1.01	-0.63	-8.49	0.00
IPW Arm 0	θ_1	-2.75	0.33	-3.4	-2.11	-8.37	0.00
	θ_2	-0.22	0.57	-1.35	0.9	-0.39	0.70
	θ_3	-0.8	0.5	-1.78	0.18	-1.59	0.11
	θ_4	0.11	0.39	-0.65	0.87	0.28	0.78
	θ_5	-0.9	0.51	-1.89	0.09	-1.78	0.08
IPW Arm 1	θ_6	-2.4	0.09	-2.57	-2.23	-27.81	0.00
	θ_7	-0.33	0.14	-0.6	-0.07	-2.46	0.01
	θ_8	-0.33	0.08	-0.48	-0.17	-4.17	0.00
	θ_9	-0.55	0.3	-1.14	0.05	-1.81	0.07
	θ_{10}	-1.14	0.2	-1.53	-0.76	-5.81	0.00

Table 3: Real data analysis with online statistic inference. We use ε-greedy algorithm with $\varepsilon = 0.02$.

References

Ban, Gah-Yi and Cynthia Rudin (2019). The big data newsvendor: Practical insights from machine learning. *Operations Research 67*(1), 90–108.

Chen, Elynn Y, Rui Song, and Michael I Jordan (2022). Reinforcement learning with heterogeneous data: estimation and inference. *arXiv preprint arXiv:2202.00088*.

Chen, Haoyu, Wenbin Lu, and Rui Song (2021a). Statistical inference for online decision making: In a contextual bandit setting. *Journal of the American Statistical Association 116*(533), 240–255.

Chen, Haoyu, Wenbin Lu, and Rui Song (2021b). Statistical inference for online decision making via stochastic gradient descent. *Journal of the American Statistical Association 116*(534), 708–719.

Chen, Xi, Jason D Lee, Xin T Tong, and Yichen Zhang (2020). Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics 48*(1), 251–273.

Chen, Xi, Zachary Owen, Clark Pixton, and David Simchi-Levi (2022). A statistical learning approach to personalization in revenue management. *Management Science 68*(3), 1923–1937.

Deshpande, Yash, Lester Mackey, Vasilis Syrgkanis, and Matt Taddy (2018). Accurate inference for adaptive linear models. In *International Conference on Machine Learning*, pp. 1194–1203. PMLR.

Duchi, John C and Feng Ruan (2021). Asymptotic optimality in stochastic optimization. *The Annals of Statistics 49*(1), 21–48.

Fang, Yixin, Jinfeng Xu, and Lei Yang (2018). Online bootstrap confidence intervals for the stochastic gradient descent estimator. *The Journal of Machine Learning Research 19*(1), 3053–3073.

Hammersley, John (2013). *Monte carlo methods*. Springer Science & Business Media.

Hao, Botao, Yasin Abbasi Yadkori, Zheng Wen, and Guang Cheng (2019). Bootstrapping upper confidence bound. *Advances in Neural Information Processing Systems 32*.

Kasy, Maximilian and Anja Sautmann (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica 89*(1), 113–132.

Khamaru, Koulik, Yash Deshpande, Lester Mackey, and Martin J Wainwright (2021). Near-optimal inference in adaptive linear regression. *arXiv preprint arXiv:2107.02266*.

Kim, Edward S, Roy S Herbst, Ignacio I Wistuba, J Jack Lee, George R Blumenschein, Anne Tsao, David J Stewart, Marshall E Hicks, Jeremy Erasmus, Sanjay Gupta, et al. (2011). The BATTLE trial: Personalizing therapy for lung cancerthe BATTLE trial: Personalizing therapy for lung cancer. *Cancer Discovery 1*(1), 44–53.

Langford, John and Tong Zhang (2007). The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in Neural Information Processing Systems 20*.

Lee, Sokbae, Yuan Liao, Myung Hwan Seo, and Youngki Shin (2022a). Fast and robust online inference with stochastic gradient descent via random scaling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 36, pp. 7381–7389.

Lee, Sokbae, Yuan Liao, Myung Hwan Seo, and Youngki Shin (2022b). Fast inference for quantile regression with millions of observations. *arXiv preprint arXiv:2209.14502*.

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

Polyak, Boris T and Anatoli B Juditsky (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization 30*(4), 838–855.

Qiang, Sheng and Mohsen Bayati (2016). Dynamic pricing with demand covariates. *arXiv preprint arXiv:1604.07463*.

Ramprasad, Pratik, Yuantong Li, Zhuoran Yang, Zhaoran Wang, Will Wei Sun, and Guang Cheng (2022). Online bootstrap inference for policy evaluation in reinforcement learning. *Journal of the American Statistical Association*, 1–14.

Robbins, Herbert (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society 58*(5), 527–535.

Robbins, Herbert and Sutton Monro (1951). A stochastic approximation method. *The Annals of Mathematical Statistics 22*(3), 400–407.

Rockafellar, R Tyrrell and Stanislav Uryasev (2002). Conditional value-at-risk for general loss distributions. *Journal of banking & finance 26*(7), 1443–1471.

Ruppert, David (1988). Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering.

Shao, Qi-Man and Zhuo-Song Zhang (2022). Berry–esseen bounds for multivariate nonlinear statistics with applications to M-estimators and stochastic gradient descent algorithms. *Bernoulli 28*(3), 1548–1576.

Shi, Chengchun, Shengxing Zhang, Wenbin Lu, and Rui Song (2021). Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*.

Shi, Chengchun, Jin Zhu, Shen Ye, Shikai Luo, Hongtu Zhu, and Rui Song (2022). Off-policy confidence interval estimation with confounded markov decision process. *Journal of the American Statistical Association*, 1–12.

Su, Weijie J and Yuancheng Zhu (2018). Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent. *arXiv preprint arXiv:1802.04876*.

Woodroofe, Michael (1979). A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association 74*(368), 799–806.

Zhang, Kelly, Lucas Janson, and Susan Murphy (2021). Statistical inference with M-estimators on adaptively collected data. *Advances in Neural Information Processing Systems 34*, 7460–7471.

Zhang, Kelly W, Lucas Janson, and Susan A Murphy (2022). Statistical inference after adaptive sampling in non-markovian environments. *arXiv preprint arXiv:2202.07098*.

Zhu, Wanrong, Xi Chen, and Wei Biao Wu (2021). Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, 1–12.