
Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

# Online Statistical Inference for Stochastic Optimization via Kiefer-Wolfowitz Methods

Xi Chen, Zehua Lai, He Li & Yichen Zhang

**To cite this article:** Xi Chen, Zehua Lai, He Li & Yichen Zhang (31 Jan 2024): Online Statistical Inference for Stochastic Optimization via Kiefer-Wolfowitz Methods, Journal of the American Statistical Association, DOI: 10.1080/01621459.2023.2296703

**To link to this article:** https://doi.org/10.1080/01621459.2023.2296703

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 00, NO. 0, 1–11: Theory and Methods
https://doi.org/10.1080/01621459.2023.2296703



# Online Statistical Inference for Stochastic Optimization via Kiefer-Wolfowitz Methods

Xi Chen[a], Zehua Lai[b], He Li[a], and Yichen Zhang[c]

[a]Stern School of Business, New York University, New York, NY; [b]CCAM, University of Chicago, Chicago, IL; [c]Mitchell E. Daniels, Jr. School of Business, Purdue University, West Lafayette, IN

**ABSTRACT**

This article investigates the problem of online statistical inference of model parameters in stochastic optimization problems via the Kiefer-Wolfowitz algorithm with random search directions. We first present the asymptotic distribution for the Polyak-Ruppert-averaging type Kiefer-Wolfowitz (AKW) estimators, whose asymptotic covariance matrices depend on the distribution of search directions and the function-value query complexity. The distributional result reflects the tradeoff between statistical efficiency and function query complexity. We further analyze the choice of random search directions to minimize certain summary statistics of the asymptotic covariance matrix. Based on the asymptotic distribution, we conduct online statistical inference by providing two construction procedures of valid confidence intervals. Supplementary materials for this article are available online.

**ARTICLE HISTORY**
Received November 2021
Accepted December 2023

**KEYWORDS**
Asymptotic normality; Kiefer-Wolfowitz stochastic approximation; Online inference; Stochastic optimization

## 1. Introduction

Stochastic optimization algorithms, introduced by Robbins and Monro (1951) and Kiefer and Wolfowitz (1952), have been widely used in statistical estimation, especially for large-scale datasets and online learning where the sample arrives sequentially (e.g., web search queries, transactional data). The Robbins–Monro algorithm (Robbins and Monro 1951), often known as the stochastic gradient descent, is perhaps the most popular algorithm in stochastic optimization and has found a wide range of applications in statistics and machine learning. Nevertheless, in many modern applications, the gradient information is not available. For example, the objective function may be embedded in a black box and the user can only access the noisy objective value for a given input. In such cases, the Kiefer-Wolfowitz algorithm (Kiefer and Wolfowitz 1952) becomes a natural choice as it is completely free of gradient computation. Despite being equipped with an evident computational advantage to avoid gradient measurements, the Kiefer-Wolfowitz algorithm has been historically out of practice as compared to the Robbins-Monro counterpart. Nonetheless, heralded by the big data era, there has been a restoration of the interest of gradient-free optimization in a wide range of applications in recent years (Conn, Scheinberg, and Vicente 2009; Nesterov and Spokoiny 2017). We briefly highlight a few of them to motivate our article.

- In some bandit problems, one may only have black-box access to individual objective values but not to their gradients (Flaxman, Kalai, and McMahan 2005; Shamir 2017). Other examples include graphical models and variational inference problems, where the objective is defined variationally (Wainwright and Jordan 2008), and the explicit differentiation can be difficult.
- In some scenarios, the computation of gradient information is possible but very expensive. For example, in the online sensor selection problem (Joshi and Boyd 2008), evaluating the stochastic gradient requires the inverse of matrices, which generates $\mathcal{O}(d^3)$ computation cost per iteration, where $d$ is the number of sensors in the network. In addition, the storage for gradient calculation also requires an $O(d^3)$ memory, which could be practically infeasible.

This article aims to study the asymptotic properties of the Kiefer-Wolfowitz stochastic optimization and conduct online statistical inference. In particular, we consider the problem,

$$\boldsymbol{\theta}^\star = \operatorname{argmin} F(\boldsymbol{\theta}), \quad \text{where} \tag{1}$$
$$F(\boldsymbol{\theta}) := \mathbb{E}_{\mathcal{P}_\zeta}\left[f(\boldsymbol{\theta};\boldsymbol{\zeta})\right] = \int f(\boldsymbol{\theta};\boldsymbol{\zeta})\mathrm{d}\mathcal{P}_\zeta,$$

where $f(\boldsymbol{\theta};\boldsymbol{\zeta})$ is a convex *individual loss function* for a data point $\boldsymbol{\zeta}$, $F(\boldsymbol{\theta})$ is the *population loss* function, and $\boldsymbol{\theta}^\star$ is the true underlying parameter of a fixed dimension $d$. Let $\boldsymbol{\theta}_0$ denote any given initial point. Given a sequentially arriving online sample $\{\boldsymbol{\zeta}_n\}$, the Robbins and Monro (1951) algorithm (RM), also known as the stochastic gradient descent (SGD), iteratively updates,

$$\text{(RM)} \qquad \boldsymbol{\theta}_n^{(\mathrm{RM})} = \boldsymbol{\theta}_{n-1}^{(\mathrm{RM})} - \eta_n g(\boldsymbol{\theta}_{n-1};\boldsymbol{\zeta}_n), \tag{2}$$

where $\{\eta_n\}$ is a positive nonincreasing step-size sequence, and $g(\boldsymbol{\theta};\boldsymbol{\zeta})$ denotes the stochastic gradient, that is, $g(\boldsymbol{\theta};\boldsymbol{\zeta}) = \nabla f(\boldsymbol{\theta};\boldsymbol{\zeta})$. In the scenarios that direct gradient measurements are

**CONTACT** Xi Chen xc13@stern.nyu.edu Stern School of Business, New York University, New York, NY.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

inaccessible to practitioners, the Kiefer and Wolfowitz (1952) algorithm (KW) becomes the natural choice, as

$$\text{(KW)} \qquad \boldsymbol{\theta}_n^{(\mathrm{KW})} = \boldsymbol{\theta}_{n-1}^{(\mathrm{KW})} - \eta_n \widehat{g}(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n), \tag{3}$$

where $\widehat{g}(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n)$ is an estimator of $g(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n)$. Under the univariate framework ($d = 1$), Kiefer and Wolfowitz (1952) considered the finite-difference approximation

$$\widehat{g}(\theta_{n-1}; \zeta_n) = \frac{f(\theta_{n-1} + h_n; \zeta_n) - f(\theta_{n-1}; \zeta_n)}{h_n}, \tag{4}$$

where $h_n$ is be a positive deterministic sequence that goes to zero. Blum (1954) later extended the algorithm to the multivariate case and proved its almost sure convergence. This pioneering work extended in various directions of statistics and control theory (see, e.g., Fabian 1967, 1980; Hall and Heyde 1980; Ruppert 1982; Chen 1988; Polyak and Tsybakov 1990; Spall 1992; Chen, Duncan, and Pasik-Duncan 1999; Spall 2000; Hall and Molchanov 2003; Dippon 2003; Mokkadem and Pelletier 2007; Broadie, Cicek, and Zeevi 2011). In the optimization literature, the Kiefer-Wolfowitz (KW) algorithm is often referred to as the gradient-free stochastic optimization, or zeroth-order SGD (Agarwal, Dekel, and Xiao 2010; Agarwal et al. 2011; Jamieson, Nowak, and Recht 2012; Ghadimi and Lan 2013; Duchi et al. 2015; Shamir 2017; Nesterov and Spokoiny 2017; Wang et al. 2018, among others).

For the (RM) algorithm in (2), Ruppert (1988) and Polyak and Juditsky (1992) characterize the limiting distribution and statistical efficiency of the *averaged iterate* $\overline{\boldsymbol{\theta}}_n^{(\mathrm{RM})} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\theta}_i^{(\mathrm{RM})}$ by

$$\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n^{(\mathrm{RM})} - \boldsymbol{\theta}^\star\right) \Longrightarrow \mathcal{N}\left(\mathbf{0}, H^{-1}SH^{-1}\right), \tag{5}$$

where $H = \nabla^2 F(\boldsymbol{\theta}^\star)$ is the Hessian matrix of $F(\boldsymbol{\theta})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^\star$, and $S = \mathbb{E}[\nabla f(\boldsymbol{\theta}^\star; \boldsymbol{\zeta}) \nabla f(\boldsymbol{\theta}^\star; \boldsymbol{\zeta})^\top]$ is the Gram matrix of the stochastic gradient. Under a well-specified model, this asymptotic covariance matrix matches the inverse Fisher information and the averaged (RM) estimator is asymptotically efficient. Based on the limiting distribution result (5), there are many recent research efforts devoted to statistical inference for (RM). A brief survey is conducted at the end of the introduction.

For the (KW) scheme, we can similarly construct the averaged Kiefer-Wolfowitz (AKW) estimator

$$\text{(AKW)} \qquad \overline{\boldsymbol{\theta}}_n^{(\mathrm{KW})} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\theta}_i^{(\mathrm{KW})}. \tag{6}$$

As compared to well-established asymptotic properties of (RM), study of the asymptotics of (AKW) is limited, particularly with a random sampling direction in multivariate (KW). In this article, we study the (KW) algorithm (3) with random search directions $\{\boldsymbol{v}_i\}_{i=1}^n \overset{\text{iid}}{\sim} \mathcal{P}_{\boldsymbol{v}}$, that is, at each iteration $i = 1, 2, \ldots, n$, a random direction $\boldsymbol{v}_i$ is sampled independently from $\mathcal{P}_{\boldsymbol{v}}$, and the (KW) gradient

$$\widehat{g}_{h_n, \boldsymbol{v}_n}(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n) = \frac{f(\boldsymbol{\theta}_{n-1} + h_n \boldsymbol{v}_n; \boldsymbol{\zeta}_n) - f(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n)}{h_n} \boldsymbol{v}_n. \tag{7}$$

Compared to the (RM) scheme, (KW) introduces additional randomness into the stochastic gradient estimator through $\{\boldsymbol{v}_n\}$. Indeed, as one can see from our main result in Theorem 3.3, (AKW) is no longer statistically efficient and its asymptotic covariance structure depends on the distribution $\mathcal{P}_{\boldsymbol{v}}$. It opens the room for the investigation on the impact of $\mathcal{P}_{\boldsymbol{v}}$ (see Section 3.1 for details). We further extend the estimator to use multiple function-value queries per step and establish an online statistical inference framework. We summarize our main results and contributions as follows,

- First, we quantify the asymptotic covariance structure of (AKW) in Theorem 3.3. Since the asymptotic distribution depends on the choice of the direction variable $\boldsymbol{v}$, we provide an introductory analysis on the asymptotic performance for different choices of random directions for constructing (AKW) estimators (see Section 3.1).
- The efficiency loss of (AKW) is due to the information constraint as one evaluates only *two* function values at each iteration. We analyze the (AKW) estimators in which multiple function queries can be assessed at each iteration, and show that the asymptotic covariance matrix decreases as the number of function queries $m+1$ increases (see Section 3.2). Moreover, (AKW) achieves asymptotic statistical efficiency as $m \to \infty$. We further show that when $\boldsymbol{v}$ is sampled without replacement from $\mathcal{P}_{\boldsymbol{v}}$ with a discrete uniform distribution of any orthonormal basis, (AKW) achieves asymptotic statistical efficiency with $d+1$ function queries per iteration.
- Based on the asymptotic distribution, we propose two online statistical inference procedures. The first one is using a plug-in estimator of the asymptotic covariance matrix, which separately estimates the Hessian matrix and Gram matrix of the (KW) gradients (with additional function-value queries, see Theorem 4.3). The second procedure is to characterize the distribution of intermediate (KW) iterates as a stochastic process and construct an asymptotically pivotal statistic by normalizing the (AKW) estimator, without directly estimating the covariance matrix. This inference procedure follows the "random scaling" method proposed in Lee et al. (2022a) that considers the online inference for the (RM) scheme. These two procedures have their advantages and disadvantages: the plug-in approach leads to better empirical performance but requires additional function-value queries to estimate the Hessian matrix, while the other one is more efficient in both computation and storage, though its finite-sample performance is inferior in practice when the dimension is large. A practitioner may choose the approach suitable to her computational resources and requirement of the inference accuracy.

Lastly, we provide a brief literature survey on the recent works for statistical inference for the (RM)-type SGD algorithms. Chen et al. (2020) developed a batch-means estimator of the limiting covariance matrix $H^{-1}SH^{-1}$ in (5), which only uses the stochastic gradient information (i.e., without estimating any Hessian matrices). Zhu, Chen, and Wu (2023) further extended the batch-means method in Chen et al. (2020) to a fully online covariance estimator. Lee et al. (2022a) extended the results in Polyak and Juditsky (1992) to a functional central limit theorem and used it to propose a novel online inference procedure that allows for efficient implementation, followed by Lee et al.

(2022b) and Chen et al. (2023) for application to quantile regression and generalized method of moments. Fang, Xu, and Yang (2018) presented a perturbation-based resampling procedure for inference. Su and Zhu (2018) proposed a tree-structured inference scheme, which splits the SGD into several threads to construct confidence intervals. Liang and Su (2019) introduced a moment-adjusted method and its corresponding inference procedure. Toulis and Airoldi (2017) considered the implicit SGD, and investigate the statistical inference problem under the variant. Duchi and Ruan (2021) studied the stochastic optimization problem with constraints and investigate its optimality properties. Chao and Cheng (2019) proposed a class of generalized regularized dual averaging (RDA) algorithms and make uncertainty quantification possible for online $\ell_1$-penalized problems. Shi et al. (2021) developed an online estimation procedure for high-dimensional statistical inference. Chen, Lu, and Song (2021) studied statistical inference of online decision-making problems via SGD in a contextual bandit setting.

### 1.1. Notations and Organization of the Article

We write vectors in boldface letters (e.g., $\boldsymbol{\theta}$ and $\boldsymbol{v}$) and scalers in lightface letters (e.g., $\eta$). For any positive integer $n$, we use $[n]$ as a shorthand for the discrete set $\{1, 2, \ldots, n\}$. Let $\{\boldsymbol{e}_k\}_{k=1}^d$ be the standard basis in $\mathbb{R}^d$ with the $k$th coordinate as 1 and the other coordinates as 0. Denote $I_d$ as the identity matrix in $\mathbb{R}^{d\times d}$. Let $\|\cdot\|$ denote the standard Euclidean norm for vectors and the spectral norm for matrices. We use $A_{k\ell}$ and $A_{n,k\ell}$ to denote the $(k,\ell)$th element of matrices $A, A_n \in \mathbb{R}^{d\times d}$, respectively, for all $k, \ell \in [d]$. Furthermore, we denote by $\mathrm{diag}(\boldsymbol{v})$ a matrix in $\mathbb{R}^{d\times d}$ whose main diagonal is the same as the vector $\boldsymbol{v}$ and off-diagonal elements are zero, for some vector $\boldsymbol{v} \in \mathbb{R}^d$. With a slight abuse of notation, for a matrix $M \in \mathbb{R}^{d\times d}$, we also let $\mathrm{diag}(M)$ denote a $\mathbb{R}^{d\times d}$ diagonal matrix with same diagonal elements as matrix $M$. We use the standard Loewner order notation $A \succeq 0$ if a matrix $A$ is positive semidefinite. We use $\boldsymbol{\theta}^{(\mathtt{RM})}$ and $\boldsymbol{\theta}^{(\mathtt{KW})}$ to denote the iterates generated by the (RM) scheme and the (KW) scheme, respectively. We use $\widehat{\boldsymbol{\theta}}^{(\mathrm{ERM})}$ for the offline empirical risk minimizer, that is, $\widehat{\boldsymbol{\theta}}^{(\mathrm{ERM})} = \mathrm{argmin}_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{\theta}; \boldsymbol{\zeta}_i)$. As we focus on the (KW) scheme in this article, we sometimes omit the superscript (KW) in the estimator to make room for the other notations. In derivations of the (KW) estimator, we denote the finite difference of $f(\cdot)$ as,

$$\Delta_{h,\boldsymbol{v}} f(\boldsymbol{\theta}; \boldsymbol{\zeta}) = f(\boldsymbol{\theta} + h\boldsymbol{v}; \boldsymbol{\zeta}) - f(\boldsymbol{\theta}; \boldsymbol{\zeta}), \tag{8}$$

for some spacing parameter $h \in \mathbb{R}_+$ and search vector $\boldsymbol{v} \in \mathbb{R}^d$. We use $\mathbb{E}_n$ to denote the conditional expectation with respect to the natural filtration, that is,

$$\mathbb{E}_n[\boldsymbol{\theta}_{n+1}] := \mathbb{E}[\boldsymbol{\theta}_{n+1}|\mathcal{F}_n], \quad \mathcal{F}_n := \sigma\{\boldsymbol{\theta}_k, \boldsymbol{\zeta}_k | k \le n\}.$$

We use the $\mathcal{O}(\cdot)$ notation to hide universal constants independent of the sample size $n$.

The remainder of the article is organized as follows. In Section 2, we describe the Kiefer-Wolfowitz algorithm with random search directions along with three illustrative examples of the classical regression problems. We also provide a technical lemma to characterize the limiting behavior of the (KW) gradient, which leads to the distributional constraint of the random direction vector. In Section 3, we first introduce the technical assumptions before we present the finite-sample rate of convergence of the (KW) estimator. We further provide the asymptotic distribution of the (AKW) estimator, accompanied by discussions on the statistical (in)efficiency. We highlight a comparison of the choices of the direction distributions in Section 3.1, and further extend the theoretical analysis to multi-query settings of the (KW) algorithm in Section 3.2. Section 3.3 generalizes our analysis to some specific nonsmooth loss functions, such as the quantile regression. Based on the established asymptotic distribution results, we propose two types of online statistical inference procedures in Section 4. A functional extension of the distributional analysis of (KW) as a stochastic process is also provided. Numerical experiments in Section 5 lend empirical support to our theory. All proofs are relegated to the supplementary material.

## 2. Kiefer-Wolfowitz Algorithm

In this section, we introduce the general form of the Kiefer-Wolfowitz (KW) gradient estimator and the corresponding iterative algorithm $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} - \eta_n \widehat{g}(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n)$. In the seminal work by Blum (1954), the (KW) gradient estimator $\widehat{g}(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n)$ is constructed by approximating the stochastic gradient $g(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n)$ using the canonical basis of $\mathbb{R}^d$, $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d\}$, as search directions. In particular, given any $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{\zeta} \sim \mathcal{P}_{\boldsymbol{\zeta}}$, the $k$th coordinate of the (KW) gradient estimator

$$\big(\widehat{g}_{h,\boldsymbol{e}}(\boldsymbol{\theta}; \boldsymbol{\zeta})\big)_k = \frac{f(\boldsymbol{\theta} + h\boldsymbol{e}_k; \boldsymbol{\zeta}) - f(\boldsymbol{\theta}; \boldsymbol{\zeta})}{h}, \quad \text{for} \quad k = 1, 2, \ldots, d, \tag{9}$$

where $h$ is a spacing parameter for approximation. At each iteration, (9) queries $d+1$ function values from $d$ fixed directions $\{\boldsymbol{e}_k\}_{k=1}^d$. To reduce the query complexity, a random difference becomes a natural choice. Koronacki (1975) introduced a random version of the (KW) algorithm using a sequence of random unit vectors that are independent and uniformly distributed on the unit sphere or unit cube. Spall (1992) also provided a random direction version of the (KW) algorithm, named as the simultaneous perturbation stochastic approximation (SPSA) algorithm and later extended to several variants (Chen, Duncan, and Pasik-Duncan 1999; Spall 2000; He, Fu, and Marcus 2003). These random direction methods can reduce the bias in gradient estimates as compared to their nonrandom counterparts. In the following, we write the (KW) algorithm with general random search directions, as in (7),

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} - \eta_n \widehat{g}_{h_n,\boldsymbol{v}_n}(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n), \qquad \text{where} \tag{10}$$

$$\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta}; \boldsymbol{\zeta}) := \frac{1}{h}\Delta_{h,\boldsymbol{v}} f(\boldsymbol{\theta}; \boldsymbol{\zeta})\boldsymbol{v} = \frac{f(\boldsymbol{\theta} + h\boldsymbol{v}; \boldsymbol{\zeta}) - f(\boldsymbol{\theta}; \boldsymbol{\zeta})}{h}\boldsymbol{v}.$$

Here $\{\boldsymbol{v}_n\}$ is sampled from an underlying distribution $\mathcal{P}_{\boldsymbol{v}}$ satisfying certain conditions (see Assumption 4 in Section 3). At each iteration $n$, the algorithm samples a direction vector $\boldsymbol{v}_n$ independently from $P_{\boldsymbol{v}}$, and makes two solitary function-value queries, $f(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n)$ and $f(\boldsymbol{\theta}_{n-1} + h_n\boldsymbol{v}_n; \boldsymbol{\zeta}_n)$. We refer to the (KW) gradient estimator $\widehat{g}_{h_n,\boldsymbol{v}_n}(\boldsymbol{\theta}_{n-1}, \boldsymbol{\zeta}_n)$ in (10) as a *two-query* finite-difference approximation of the stochastic gradient. If one

is allowed to make additional function-value queries, an averaging of the function values from multiple directions generates a *multi-query* stochastic gradient estimator with reduced variance. In particular, at each iteration $n$, the practitioner makes $m+1$ queries $\{f(\boldsymbol{\theta}_{n-1};\boldsymbol{\zeta}_n), f(\boldsymbol{\theta}_{n-1}+h_n\boldsymbol{v}_n^{(j)};\boldsymbol{\zeta}_n)\}_{1\le j\le m}$ via $m$ random directions $\{\boldsymbol{v}_n^{(j)}\}$ sampled from $\mathcal{P}_{\boldsymbol{v}}$. If $\mathcal{P}_{\boldsymbol{v}}$ is a finite distribution, practitioners may choose to sample *with* or *without replacement*. In summary, an $(m+1)$-*query* (KW) algorithm constructs a stochastic gradient estimator

$$\begin{aligned}\overline{g}_n^{(m)}(\boldsymbol{\theta}_{n-1};\boldsymbol{\zeta}_n) &= \frac{1}{m}\sum_{j=1}^{m}\widehat{g}_{h_n,\boldsymbol{v}_n^{(j)}}(\boldsymbol{\theta}_{n-1};\boldsymbol{\zeta}_n) \\ &= \frac{1}{mh_n}\sum_{j=1}^{m}\Delta_{h_n,\boldsymbol{v}_n^{(j)}}f(\boldsymbol{\theta}_{n-1};\boldsymbol{\zeta}_n)\boldsymbol{v}_n^{(j)},\end{aligned} \tag{11}$$

at each iteration $n$, and updates $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} - \eta_n\overline{g}_n^{(m)}(\boldsymbol{\theta}_{n-1};\boldsymbol{\zeta}_n)$. Here we restrict the procedure to sampling from the same distribution $\mathcal{P}_{\boldsymbol{v}}$ independently across different iterations. We use $\boldsymbol{\theta}_n^{(m)}$ to denote the final (KW) estimator using the above $(m+1)$-query finite-difference approximation.

We now provide some illustrative examples of the two-query (KW) estimator $\widehat{g}_{h_n,\boldsymbol{v}_n}$ in (10) used in popular statistical models, and we will refer to these examples throughout the article. A multi-query extension of the examples can be constructed accordingly.

*Example 2.1 (Linear Regression).* Consider a linear regression model $y_i = \boldsymbol{x}_i^\top\boldsymbol{\theta}^\star + \epsilon_i$ where $\{\boldsymbol{\zeta}_i = (\boldsymbol{x}_i, y_i),\ i = 1,2,\ldots,n\}$ is an iid sample of $\boldsymbol{\zeta} = (\boldsymbol{x}, y)$ and the noise $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$. We use a quadratic loss function $f(\boldsymbol{\theta};\boldsymbol{\zeta}) = (y - \boldsymbol{x}^\top\boldsymbol{\theta})^2$. Therefore, the stochastic gradient $\nabla f(\boldsymbol{\theta};\boldsymbol{\zeta}) = \big(\boldsymbol{x}^\top\boldsymbol{\theta} - y\big)\boldsymbol{x}$, and the (KW) gradient estimator $\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\{\boldsymbol{x},y\})$ in (10) becomes

$$\begin{aligned}\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\{\boldsymbol{x},y\}) &= \frac{1}{h}\Big[\big(y - \boldsymbol{x}^\top(\boldsymbol{\theta}+h\boldsymbol{v})\big)^2 - \big(y-\boldsymbol{x}^\top\boldsymbol{\theta}\big)^2\Big]\boldsymbol{v} \\ &= 2\boldsymbol{v}\boldsymbol{v}^\top\big(\boldsymbol{x}^\top\boldsymbol{\theta} - y\big)\boldsymbol{x} + h(\boldsymbol{x}^\top\boldsymbol{v})^2\boldsymbol{v}.\end{aligned}$$

*Example 2.2 (Logistic Regression).* Consider a logistic regression model with a binary response $y_i \in \{-1,1\}$ generated by $\Pr(y_i|\boldsymbol{x}_i) = \big(1+\exp\big(-y_i\boldsymbol{x}_i^\top\boldsymbol{\theta}^\star\big)\big)^{-1}$. The individual loss function $f(\boldsymbol{\theta};\boldsymbol{\zeta}) = \log\big(1+\exp(-y\boldsymbol{x}^\top\boldsymbol{\theta})\big)$. The stochastic gradient $\nabla f(\boldsymbol{\theta};\boldsymbol{\zeta}) = -y\boldsymbol{x}\big(1+\exp(y\boldsymbol{x}^\top\boldsymbol{\theta})\big)^{-1}$, and the (KW) gradient estimator $\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\{\boldsymbol{x},y\})$ in (10) becomes

$$\begin{aligned}\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\{\boldsymbol{x},y\}) &= \frac{\boldsymbol{v}}{h}\Big[\log\big(1+\exp(-y\boldsymbol{x}^\top(\boldsymbol{\theta}+h\boldsymbol{v}))\big) \\ &\qquad - \log\big(1+\exp(-y\boldsymbol{x}^\top\boldsymbol{\theta})\big)\Big] \\ &= \frac{-y\boldsymbol{v}\boldsymbol{v}^\top\boldsymbol{x}}{1+\exp(y\boldsymbol{x}^\top\boldsymbol{\theta})} + \frac{y^2(\boldsymbol{x}^\top\boldsymbol{v})^2\exp(y\boldsymbol{x}^\top\boldsymbol{\theta})h\boldsymbol{v}}{2(1+\exp(y\boldsymbol{x}^\top\boldsymbol{\theta}))^2} \\ &\quad + \mathcal{O}(h^2), \quad \text{as } h\to 0_+,\end{aligned}$$

under some regularity conditions on $\boldsymbol{\theta}$ and the distribution of $\boldsymbol{x}$.

We note that for the (RM) scheme with differentiable loss functions, the stochastic gradient is an unbiased estimator of the population gradient under very mild assumption, that is, $\mathbb{E}_{\boldsymbol{\zeta}}g(\boldsymbol{\theta};\boldsymbol{\zeta}) = \nabla F(\boldsymbol{\theta})$. In contrast, the (KW) gradient estimator is no longer an unbiased estimator of $\nabla F(\boldsymbol{\theta})$. In the following lemma, we precisely quantifies the bias incurred by the (KW) gradient estimator.

*Lemma 2.3.* We assume that the population loss function $F(\cdot)$ is twice continuously differentiable and $L_f$-smooth, that is, $\nabla^2F(\boldsymbol{\theta}) \preceq L_fI_d$ for any $\boldsymbol{\theta}\in\mathbb{R}^d$. Given any fixed parameter $\boldsymbol{\theta}\in\mathbb{R}^d$, suppose the random direction vector $\boldsymbol{v}$ is independent from $\boldsymbol{\zeta}$, we have

$$\big\|\mathbb{E}\,\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\boldsymbol{\zeta}) - \nabla F(\boldsymbol{\theta})\big\| \le \Big\|\mathbb{E}\big(\boldsymbol{v}\boldsymbol{v}^\top - I_d\big)\nabla F(\boldsymbol{\theta})\Big\| + \frac{h}{2}L_f\mathbb{E}\|\boldsymbol{v}\|^3,$$

where the expectation in $\mathbb{E}\,\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\boldsymbol{\zeta})$ takes over both the randomness in $\boldsymbol{v}$ and $\boldsymbol{\zeta}$.

The proof of Lemma 2.3 is provided in Section A of the supplementary material. To reduce the bias in the (KW) gradient, Lemma 2.3 indicates that one should choose the random direction $\boldsymbol{v}_n$ that satisfies the distributional constraint $\mathbb{E}[\boldsymbol{v}_n\boldsymbol{v}_n^\top] = I_d$ (see Assumption 4 in Section 3). We will further conduct a comprehensive analysis in Section 3.1 on different choices of distributions $\mathcal{P}_{\boldsymbol{v}}$ satisfying the condition $\mathbb{E}[\boldsymbol{v}_n\boldsymbol{v}_n^\top] = I_d$. Despite the existence of the bias, as the spacing parameter $h_n\to 0$, the bias convergences to zero asymptotically.

## 3. Theoretical Results

We first introduce some regularity assumptions on the population loss $F(\boldsymbol{\theta})$ and the individual loss $f(\boldsymbol{\theta};\boldsymbol{\zeta})$.

*Assumption 1.* The population loss function $F(\boldsymbol{\theta})$ is twice continuously differentiable, convex, and $L_f$-smooth, that is, $0 \preceq \nabla^2F(\boldsymbol{\theta}) \preceq L_fI_d$ for any $\boldsymbol{\theta}\in\mathbb{R}^d$. In addition, there exist $\delta_1,\lambda > 0$, such that, $\nabla^2F(\boldsymbol{\theta}) \succeq \lambda I_d$ for any $\boldsymbol{\theta}$ in the $\delta_1$-ball centered at $\boldsymbol{\theta}^\star$.

*Assumption 2.* Assume $\mathbb{E}\big[\nabla f(\boldsymbol{\theta};\boldsymbol{\zeta}_n)\big] = \nabla F(\boldsymbol{\theta})$ for any $\boldsymbol{\theta}\in\mathbb{R}^d$. Moreover, for some $0 < \delta \le 2$, there exists $M > 0$ such that $\mathbb{E}\|\nabla f(\boldsymbol{\theta};\boldsymbol{\zeta}_n) - \nabla F(\boldsymbol{\theta})\|^{2+\delta} \le M\big(\|\boldsymbol{\theta}-\boldsymbol{\theta}^\star\|^{2+\delta}+1\big)$.

*Assumption 3.* There are constants $L_h, L_p > 0$ such that for any $\boldsymbol{\theta},\boldsymbol{\theta}'\in\mathbb{R}^d$,

$$\begin{aligned}&\mathbb{E}\,\big\|\nabla^2 f(\boldsymbol{\theta};\boldsymbol{\zeta}_n) - \nabla^2 f(\boldsymbol{\theta}';\boldsymbol{\zeta}_n)\big\|^2 \le L_h\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^2, \\ &\mathbb{E}\,\big\|[\nabla^2 f(\boldsymbol{\theta}^\star;\boldsymbol{\zeta}_n)]^2 - H^2\big\| \le L_p,\end{aligned}$$

where $H$ is the Hessian matrix of the population loss function $F(\cdot)$, that is, $H = \nabla^2F(\boldsymbol{\theta}^\star)$.

*Assumption 4.* We adopt iid random direction vectors $\{\boldsymbol{v}_n\}$ from some common distribution $\boldsymbol{v} \sim \mathcal{P}_{\boldsymbol{v}}$ such that $\mathbb{E}[\boldsymbol{v}\boldsymbol{v}^\top] = I_d$. Moreover, assume that the $(6+3\delta)$th moment of $\boldsymbol{v}$ is bounded.

We discuss the above assumptions and compare them with the standard conditions in the literature of (RM)-type SGD inference. Assumption 1 assumes the population loss function $F(\cdot)$ to be $\lambda$-strongly convex in a $\delta_1$ neighborhood of the true parameter $\boldsymbol{\theta}^\star$, which is often referred to as *local strong convexity*

assumption widely used in many existing literature of statistical inference on (RM)-type stochastic optimization (e.g., Polyak and Juditsky 1992). This condition is satisfied in the settings of linear and logistic regression (Examples 2.1–2.2) with classical design assumptions on the covariates $\boldsymbol{x}$. Assumption 2 introduces the unbiasedness condition on the stochastic gradient $\nabla f(\boldsymbol{\theta};\boldsymbol{\zeta})$ when the individual loss function $f(\boldsymbol{\theta};\boldsymbol{\zeta})$ is smooth. The $(2+\delta)$th moment condition is the classical Lyapunov condition used in the derivation of asymptotic normality. The statements in Assumption 3 introduce the Lipschitz continuity condition and the concentration condition on the Hessian matrix. Relaxation to Assumptions 2 and 3 can be made to handle some nonsmooth loss functions $f(\boldsymbol{\theta};\boldsymbol{\zeta})$, such as the quantile regression as described in Section 3.3. Assumption 4 guarantees that the (KW) gradient $\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\boldsymbol{\zeta})$ is an asymptotically unbiased estimator of $\nabla F(\boldsymbol{\theta})$ as the spacing parameter $h_n \to 0$, as suggested by Lemma 2.3. We provide several examples of $\mathcal{P}_{\boldsymbol{v}}$ in Section 3.1.

Before we derive the asymptotic distribution for (AKW), we first provide a consistency result and finite sample error bound for the final (KW) iterate $\boldsymbol{\theta}_n$:

*Proposition 3.1.* Assume Assumptions 1, 2, and 4 hold. Set the step size as $\eta_n = \eta_0 n^{-\alpha}$ for some constant $\eta_0 > 0$ and $\alpha \in \left(\frac{1}{2}, 1\right)$ and the spacing parameter as $h_n = h_0 n^{-\gamma}$ for constant $h_0 > 0$, and $\gamma \in \left(\frac{1}{2}, 1\right)$. The (KW) iterate $\boldsymbol{\theta}_n$ converges to $\boldsymbol{\theta}^\star$ almost surely.

Additionally, assume Assumptions 1 holds with $\delta_1 = +\infty$, that is, the population loss function $F(\boldsymbol{\theta})$ is globally $\lambda$-strongly convex and $L_f$-smooth. For sufficiently large $n$, we have, $\mathbb{E}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}^\star\|^{2+\delta} \le C n^{-\alpha(2+\delta)/2}$, where the constant $C$ depends on $d, \lambda, L_f, \alpha, \gamma, \eta_0, h_0$.

The proof of Proposition 3.1 and the explicit dependency of the constant $C$ on the parameters and the initial value $\boldsymbol{\theta}_0$ are provided in Remark A.1 of the supplementary material. A similar error bound on the parameter $\boldsymbol{\theta}$ is given by Duchi et al. (2015) in terms of the function values for $\delta = 0$. We provide an error bound for the $(2+\delta)$-moment under our assumption, where $\delta \in (0, 2]$ is assumed in Assumption 2. Proposition 3.1 suggests that the asymptotic rate of the (KW) estimator matches the best convergence rate of the (RM) estimator (Moulines and Bach 2011) when the spacing parameter $h_n = h_0 n^{-\gamma}$ is a decreasing sequence with $\gamma \in (\frac{1}{2}, 1)$.

Recall that to characterize the asymptotic behavior of (RM) iterates, we denote by $S$, the Gram matrix of $\nabla f(\boldsymbol{\theta};\boldsymbol{\zeta})$ at the true parameter $\boldsymbol{\theta}^\star$, that is, $S := \mathbb{E}\left[\nabla f(\boldsymbol{\theta}^\star;\boldsymbol{\zeta}) \nabla f(\boldsymbol{\theta}^\star;\boldsymbol{\zeta})^\top\right]$. Analogously, we define the limiting Gram matrix of the (KW) gradient estimator $\widehat{g}_{h,\boldsymbol{v}}$ at $\boldsymbol{\theta}^\star$ as $h \to 0$ to be $Q$. The following lemma proves that the limiting Gram matrix takes the form of $Q = \mathbb{E}\left[\boldsymbol{v}\boldsymbol{v}^\top S \boldsymbol{v}\boldsymbol{v}^\top\right]$, and it quantifies the distance between $\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta}^\star;\boldsymbol{\zeta}) \widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta}^\star;\boldsymbol{\zeta})^\top$ and $Q$, as the spacing parameter $h \to 0$.

*Lemma 3.2.* Under Assumptions 1, 2, 3, and 4, we have

$$\left\| \mathbb{E}\left[\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta}^\star;\boldsymbol{\zeta}) \widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta}^\star;\boldsymbol{\zeta})^\top\right] - Q \right\| \le C h (1 + h^2),$$
$$Q = \mathbb{E}\left[\boldsymbol{v}\boldsymbol{v}^\top S \boldsymbol{v}\boldsymbol{v}^\top\right].$$

where $S = \mathbb{E}\left[\nabla f(\boldsymbol{\theta}^\star;\boldsymbol{\zeta}) \nabla f(\boldsymbol{\theta}^\star;\boldsymbol{\zeta})^\top\right]$ is defined in Assumption 2.

With Lemma 3.2 in place, we state our first main result that characterizes the limiting distribution of the averaged (AKW) iterates defined in (6).

*Theorem 3.3.* Let Assumptions 1, 2, 3, and 4 hold. Set the step size as $\eta_n = \eta_0 n^{-\alpha}$ for some constant $\eta_0 > 0$ and $\alpha \in \left(\frac{1}{2}, 1\right)$, and the spacing parameter as $h_n = h_0 n^{-\gamma}$ for some constant $h_0 > 0$, and $\gamma \in \left(\frac{1}{2}, 1\right)$. The averaged (KW) estimator $\overline{\boldsymbol{\theta}}_n$ satisfies,

$$\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star\right) \Longrightarrow \mathcal{N}\left(\mathbf{0}, H^{-1} Q H^{-1}\right), \qquad \text{as} \quad n \to \infty, \tag{12}$$

where $H = \nabla^2 F(\boldsymbol{\theta}^\star)$ is the population Hessian matrix and $Q = \mathbb{E}\left[\boldsymbol{v}\boldsymbol{v}^\top S \boldsymbol{v}\boldsymbol{v}^\top\right]$ is defined in Lemma 3.2. Here $\Longrightarrow$ represents the convergence in distribution.

We now compare the asymptotic covariance matrix of $\overline{\boldsymbol{\theta}}_n$ with that of the (RM) counterpart in (5).[1] As one can see, the asymptotic covariance matrix of (AKW) estimator $\overline{\boldsymbol{\theta}}_n$ exhibits a similar sandwich form as the covariance matrix of (RM), but strictly dominates the latter, regardless of the choice of random direction vectors $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\}$. In fact, it is easy to check that

$$H^{-1} Q H^{-1} - H^{-1} S H^{-1} \tag{13}$$
$$= H^{-1} \mathbb{E}_{\boldsymbol{v}}\left[(\boldsymbol{v}\boldsymbol{v}^\top - I_d) S (\boldsymbol{v}\boldsymbol{v}^\top - I_d)\right] H^{-1} \succ 0,$$

which suggests the (AKW) estimator suffers an inevitable loss of efficiency compared to the $\widehat{\boldsymbol{\theta}}^{(\mathrm{RM})}$. In Section 3.2, we analyze (AKW) with multiple function-value queries at each iteration. With the price of additional per-iteration computational complexity, one is able to improve the statistical efficiency of (AKW) and achieve the optimal asymptotic variance $H^{-1} S H^{-1}$.

*Remark 3.4.* To complete the distributional analysis on (KW) iterates, we also provide the asymptotic distribution of the $n$th iterate $\boldsymbol{\theta}_n^{(\mathrm{KW})}$ of (3) without averaging. Assume the Hessian matrix has decomposition $H = P \Lambda P^\top$, where $P$ is an orthogonal matrix and $\Lambda$ is a diagonal matrix. Using the proof in Fabian (1968), we establish the following asymptotic distribution for $\boldsymbol{\theta}_n^{(\mathrm{KW})}$,

$$n^{\alpha/2}(\boldsymbol{\theta}_n^{(\mathrm{KW})} - \boldsymbol{\theta}^\star) \Longrightarrow \mathcal{N}(0, \Sigma), \tag{14}$$

where each $(k, \ell)$th entry of the covariance matrix $\Sigma$ is,

$$\Sigma_{k\ell} = \eta_0 \left(P^\top Q P\right)_{kl} \left(\Lambda_{kk} + \Lambda_{\ell\ell}\right)^{-1}, \quad 1 \le k, \ell \le d.$$

Here $\eta_0 > 0$ and $\alpha \in (\frac{1}{2}, 1)$ are specified in the step size $\eta_n = \eta_0 n^{-\alpha}$. As $\alpha < 1$, the $n$th iterate $\boldsymbol{\theta}_n^{(\mathrm{KW})}$ without averaging converges at a slower rate $n^{-\alpha/2}$ than that of (AKW) in Theorem 3.3.

[1] Note that the asymptotic covariance $H^{-1} S H^{-1}$ in (5) is "optimal" in the sense that it matches the asymptotic covariance for the empirical risk minimizer $\widehat{\boldsymbol{\theta}}^{(\mathrm{ERM})}$ without online computation and gradient information constraint.

### 3.1. Examples: Choices of Direction Distribution

By Theorem 3.3, the asymptotic covariance matrix of (AKW) estimator, $H^{-1}QH^{-1}$, depends on the distribution of search direction $\mathcal{P}_{\boldsymbol{v}}$ via $Q = \mathbb{E}[\boldsymbol{v}\boldsymbol{v}^\top S\boldsymbol{v}\boldsymbol{v}^\top]$. In this section, we compare the asymptotic covariance matrices of the (AKW) estimator when the random directions $\{\boldsymbol{v}_i\}_{i=1}^n$ are sampled from different $\mathcal{P}_{\boldsymbol{v}}$'s. Several popular choices of $\mathcal{P}_{\boldsymbol{v}}$ are listed as follows,

- (G) Gaussian: $\boldsymbol{v} \sim \mathcal{N}(0, I)$.
- (S) Spherical: $\boldsymbol{v}$ is sampled from the uniform distribution on the sphere $\|\boldsymbol{v}\|^2 = d$.
- (I) Uniform in the canonical basis: $\boldsymbol{v}$ is sampled from $\left\{\sqrt{d}\boldsymbol{e}_1, \sqrt{d}\boldsymbol{e}_2, \ldots, \sqrt{d}\boldsymbol{e}_d\right\}$ with equal probability, where $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d\}$ is the canonical basis of $\mathbb{R}^d$.

It is easy to verify that the above three classical choices of $\mathcal{P}_{\boldsymbol{v}}$ satisfy Assumption 4, among which (G) and (S) are continuous distributions, while (I) is a discrete distribution. In particular, (I) is a discrete uniform distribution with equal probability among the $d$ vectors of the standard basis of Euclidean space $\mathbb{R}^n$, which can be generalized in the following two forms.

- (U) Uniform in an arbitrary orthonormal basis $U$: $\boldsymbol{v}_i$ is sampled uniformly from $\left\{\sqrt{d}\boldsymbol{u}_1, \sqrt{d}\boldsymbol{u}_2, \ \ldots, \sqrt{d}\boldsymbol{u}_d\right\}$, where $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_d\}$ is an arbitrary *orthonormal basis* of $\mathbb{R}^d$, that is, the matrix $U = (\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_d)$ is a $d \times d$ orthonormal matrix such that $UU^\top = U^\top U = I$.
- (P) Nonuniform in the canonical basis with probability $(p_1, p_2, \ldots, p_d)$: $\boldsymbol{v} = \sqrt{1/p_k}\,\boldsymbol{e}_k$ with probability $p_k > 0$, for $k \in [d]$ and $\sum_{k=1}^d p_k = 1$.

The following proposition provides expressions of the matrix $Q$ for the above five choices of $\mathcal{P}_{\boldsymbol{v}}$.

*Proposition 3.5.* Under the assumptions in Theorem 3.3, for above examples of $\mathcal{P}_{\boldsymbol{v}}$,

- (G) Gaussian: $Q^{(\mathrm{G})} = (2S + \mathrm{tr}(S)I_d)$.
- (S) Spherical: $Q^{(\mathrm{S})} = \frac{d}{d+2}(2S + \mathrm{tr}(S)I_d)$.
- (I) Uniform in the canonical basis: $Q^{(\mathrm{I})} = d\,\mathrm{diag}(S)$.
- (U) Uniform in an arbitrary orthonormal basis $U$: $Q^{(\mathrm{U})} = d\,U\mathrm{diag}(U^\top SU)U^\top$.
- (P) Nonuniform in a natural coordinate basis: $Q^{(\mathrm{P})} = \mathrm{diag}(S_{11}/p_1, S_{22}/p_2, \ldots, S_{dd}/p_d)$.

From Proposition 3.5, one can see that any of the above choices of $\mathcal{P}_{\boldsymbol{v}}$ leads to a $Q^{(\cdot)}$ that strictly dominates $S$. Take $S = I_d$ as an example, we have $Q^{(\mathrm{G})} = (d+2)I_d$ and $Q^{(\mathrm{S})} = Q^{(\mathrm{I})} = Q^{(\mathrm{U})} = dI_d$ and $Q^{(\mathrm{P})} = \mathrm{diag}(p_1^{-1}, p_2^{-1}, \ldots, p_d^{-1}) \succ I_d$ where $p_1 + p_2 + \cdots + p_d = 1$. Several additional findings and implications of Proposition 3.5 are discussed in Section A.2 of the supplementary material. To briefly mention a few, the Gaussian direction (G) is always inferior to the spherical direction (S). Among the rest of the choices, there is *no domination* relationship, and different optimality criterion in the experimental design leads to different optimal choices of $\mathcal{P}_{\boldsymbol{v}}$.

### 3.2. Multi-Query Extension and Statistical Efficiency

We now consider the (AKW) estimator using $(m+1)$ function queries $\overline{\boldsymbol{\theta}}_n^{(m)}$ in (11),

$$\begin{aligned}\overline{\boldsymbol{\theta}}_n^{(m)} &= \frac{1}{n}\sum_{i=1}^n \boldsymbol{\theta}_i^{(m)}, \qquad \text{where}\\ \boldsymbol{\theta}_i^{(m)} &= \boldsymbol{\theta}_{i-1}^{(m)} - \eta_i \overline{g}_n^{(m)}(\boldsymbol{\theta}_{i-1}; \boldsymbol{\zeta}_i)\\ &= \boldsymbol{\theta}_{i-1}^{(m)} - \frac{\eta_i}{m}\sum_{j=1}^m \widehat{g}_{h_i, \boldsymbol{v}_i^{(j)}}(\boldsymbol{\theta}_{i-1}; \boldsymbol{\zeta}_i).\end{aligned}$$

Here we first consider using the same sampling distribution across $m$ queries and $n$ iterations. In other words, $\boldsymbol{v}_i^{(j)}$ is sampled iid from $\mathcal{P}_{\boldsymbol{v}}$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$. Analogous to Theorem 3.3, we present the asymptotic distribution of multi-query (AKW),

*Theorem 3.6.* Under the assumptions in Theorem 3.3, the $(m+1)$-query (AKW) estimator has the following asymptotic distribution, as $n \to \infty$,

$$\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n^{(m)} - \boldsymbol{\theta}^\star\right) \Longrightarrow \mathcal{N}\left(\mathbf{0}, H^{-1}Q_m H^{-1}\right), \qquad \text{where}$$
$$Q_m = \frac{1}{m}Q + \frac{m-1}{m}S.$$

Theorem 3.6 illustrates a tradeoff effect between the statistical efficiency and computational efficiency. When $m = 1$ and only two queries of function evaluations are available, Theorem 3.6 reduces to Theorem 3.3, and $Q_m = Q$. Conversely, as $m \to \infty$, we have $Q_m \to S$. Therefore, the asymptotic covariance of $(m+1)$-query (AKW) estimator $\overline{\boldsymbol{\theta}}_n^{(m)}$ approaches the optimal covariance $H^{-1}SH^{-1}$ as $m$ approaches infinite. Nevertheless, the algorithm requires $m$ function-value queries at each iteration, which significantly increases the computation complexity.

For a finite $m$, a slight revision of the sampling scheme of the direction vectors $\{\boldsymbol{v}_i^{(j)}\}_{j=1,2,\ldots,m}$ provides a remedy to achieve a smaller and indeed optimal asymptotic covariance matrix. Particularly at the $i$th iteration, one may sample $m$ direction vectors $\{\boldsymbol{v}_i^{(j)}\}_{j=1,2,\ldots,m}$ from a discrete distribution (such as (I) and (U)) *without replacement*. In such settings, the direction vectors $\left\{\boldsymbol{v}_i^{(1)}, \boldsymbol{v}_i^{(2)}, \ldots, \boldsymbol{v}_i^{(m)}\right\}$ are no longer independent but they have the same marginal distribution. The asymptotic distribution of the multi-query (KW) algorithm sampling without replacement is provided in the following theorem.

*Theorem 3.7.* Under the assumptions in Theorem 3.3, and the direction vectors in all iterations $\left\{\widetilde{V}_i\right\}_{i=1}^n$ are iid from $\mathcal{P}_{\boldsymbol{v}}$ such that $\widetilde{V}_i = \left(\boldsymbol{v}_i^{(1)}, \boldsymbol{v}_i^{(2)}, \ldots, \boldsymbol{v}_i^{(m)}\right)$ follows discrete sampling scheme in (I) and (U) WithOut Replacement (WOR), the $(m+1)$-query (AKW) estimator, referred to as $\overline{\boldsymbol{\theta}}_n^{(m,\mathrm{WOR})}$, has the following asymptotic distribution, as $n \to \infty$,

$$\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n^{(m,\mathrm{WOR})} - \boldsymbol{\theta}^\star\right) \Longrightarrow \mathcal{N}\left(\mathbf{0}, H^{-1}Q_m^{(\mathrm{WOR})}H^{-1}\right), \qquad \text{where}$$
$$Q_m^{(\mathrm{WOR})} = \frac{(d-m)}{m(d-1)}Q + \frac{d(m-1)}{m(d-1)}S.$$

By comparing the asymptotic covariance matrices in Theorems 3.6 and 3.7, $Q_m^{(\mathrm{WOR})}$ for sampling without replacement case is strictly smaller than $Q_m$ in Theorems 3.6 when we consider multi-query evaluation ($m \geq 2$). Moreover, when $m = d$, it is easy to see that $Q_m^{(\mathrm{WOR})} = S$. Therefore, the $(d+1)$-query (AKW) estimator $\overline{\boldsymbol{\theta}}_n^{(m,\mathrm{WOR})}$ achieves the same limiting covariance as that of the averaged (RM) estimator. Under a well-specified parametric model, the limiting covariance matrix $H^{-1}SH^{-1}$ achieves the Cramér-Rao lower bound. This result indicates that the $(d+1)$-query $\overline{\boldsymbol{\theta}}_n^{(d,\mathrm{WOR})}$ is asymptotically efficient (van der Vaart 2000).

### 3.3. Asymptotic Behavior of (AKW) Estimator for Nonsmooth Losses

The analysis of the asymptotic distribution of the (AKW) estimator remains valid naturally for some nonsmooth loss functions $F(\boldsymbol{\theta})$ including quantile regression, specifically:

*Example 3.8 (Quantile Regression).* Consider a quantile regression model $y_i = \boldsymbol{x}_i^\top \boldsymbol{\theta}^\star + \varepsilon_i$ where $\{\boldsymbol{\zeta}_i = (\boldsymbol{x}_i, y_i),\ i = 1, 2, \ldots, n\}$ is an iid sample of $\boldsymbol{\zeta} = (\boldsymbol{x}, y)$ and the noise $\varepsilon_i$ is independent from $\boldsymbol{x}_i$ and $\Pr(\varepsilon_i \leq 0) = \tau$. The individual loss $f(\boldsymbol{\theta}; \zeta) = \rho_\tau(y - \boldsymbol{x}^\top \boldsymbol{\theta})$, where $\rho_\tau(z) = z(\tau - 1_{\{z<0\}})$. Although $\rho_\tau$ is nondifferentiable, the (KW) gradient estimator $\widehat{g}_{h,\boldsymbol{v}}$ is well-defined and takes the following form,

$$\begin{aligned}\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta}; \{\boldsymbol{x}, y\}) =& \frac{\boldsymbol{v}}{h}\left[\rho_\tau\big(y - \boldsymbol{x}^\top(\boldsymbol{\theta} + h\boldsymbol{v})\big) - \rho_\tau\big(y - \boldsymbol{x}^\top\boldsymbol{\theta}\big)\right] \\ =& \boldsymbol{v}\boldsymbol{v}^\top\boldsymbol{x}\big(\tau - 1_{\{y - \boldsymbol{x}^\top\boldsymbol{\theta} < 0\}}\big), \\ & \text{for } 0 < h < \left|\frac{y - \boldsymbol{x}^\top\boldsymbol{\theta}}{\boldsymbol{x}^\top\boldsymbol{v}}\right|.\end{aligned}$$

We next state modeling assumptions for some nonsmooth losses including Example 3.8.

*Assumption 5.* Assume that $f(\boldsymbol{\theta}; \zeta) = \rho(y - \boldsymbol{x}^\top\boldsymbol{\theta})$ where $\rho(u)$ is a convex function with a subgradient $\psi(u)$, and $|\psi(u)| \leq C(|u| + 1)$ for some constant $C > 0$. The covariates $\boldsymbol{x}$ is independent of $\epsilon$ and $\boldsymbol{x}$ has finite eighth moments and nondegenerate covariance matrices. Assume the probability density function $p(x)$ of $\varepsilon$ is in $C^3$, its derivatives up to third order are all integrable, and $\varepsilon$ has finite fourth moment. Assume $\phi(u) = \mathbb{E}[\psi(u + \varepsilon)]$ is differentiable, $\phi(0) = 0$, and $u\phi(u) > 0$ for any $u \neq 0$. We further assume $\phi'(0) > 0$, and there exist constants $C > 0$ such that $|\phi'(u)| \leq C$ and $|\phi'(u) - \phi'(v)| \leq C|u - v|$.

Assumption 5 essentially guarantees that the population loss function $F(\boldsymbol{\theta})$ is smooth and locally strongly convex, and the distribution of the noise $\epsilon$ is smooth enough such that, the empirical loss (averaged individual loss) well approximates the population loss asymptotically. We now restate Theorem 3.3 for certain nonsmooth losses under Assumption 5. The proof of Theorem 3.9 is relegated to Section A.3 of the supplementary material.

*Theorem 3.9.* Let Assumptions 4 and 5 hold. Under the stepsize and spacing parameter conditions specified in Theorem 3.3, the averaged estimator $\overline{\boldsymbol{\theta}}_n$ satisfies,

$$\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star\right) \Longrightarrow \mathcal{N}\left(\mathbf{0}, H^{-1}QH^{-1}\right), \qquad \text{as} \quad n \to \infty, \tag{15}$$

where $Q = \mathbb{E}[\boldsymbol{v}\boldsymbol{v}^\top S\boldsymbol{v}\boldsymbol{v}^\top]$, $S = \mathbb{E}[\psi^2(\varepsilon)\boldsymbol{x}\boldsymbol{x}^\top]$, and $H = \mathbb{E}[\phi'(0)\boldsymbol{x}\boldsymbol{x}^\top]$.

From Theorem 3.9, we know that the (AKW) estimator of the above quantile regression model is asymptotically normal with asymptotic covariance matrix $H^{-1}QH^{-1}$ where $Q$ depends on the sampling directions (see Proposition 3.5). In an quantile regression Example 3.8 when the noise $\epsilon$ follows the normal distribution with standard deviation 1 and $\Pr(\epsilon \leq 0) = \tau$, a direct computation shows $H = \varphi\big(\Phi^{-1}(\tau)\big)\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$, where $\varphi$ and $\Phi$ are the probability and cumulative distribution functions of a standard normal distribution. Furthermore, if $\boldsymbol{v}$ is sampled uniformly from the canonical basis with two function queries, we can see from Proposition 3.5 that $Q = Q^{(\mathrm{I})} = d\,\mathrm{diag}(S)$ where $S = \tau(1 - \tau)\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$.

## 4. Online Statistical Inference

In the previous section, we provide the asymptotic distribution for the (AKW) estimator. For the purpose of conducting statistical inference of $\boldsymbol{\theta}^\star$, we need a consistent estimator of the limiting covariance $H^{-1}QH^{-1}$ in (12). A direct way is to construct a pair of consistent estimators $\hat{H}$ and $\hat{Q}$ of $H$ and $Q$, respectively, and estimate the asymptotic covariance by the *plug-in* estimator $\hat{H}^{-1}\hat{Q}\hat{H}^{-1}$. Offline construction of those estimators is generally straightforward. However, as the (KW) scheme typically applies to sequential data, it is ideal to estimate the asymptotic covariance in an online fashion without storing the data. Therefore, one cannot simply replace the true parameter $\boldsymbol{\theta}^\star$ by its estimate $\overline{\boldsymbol{\theta}}_n$ in $Q$ and $H$ in an online setting, since we can no longer access the data stream $\{\boldsymbol{\zeta}_i\}_{i=1}^n$ after the estimator $\overline{\boldsymbol{\theta}}_n$ is obtained. To address this challenge, we first propose the following finite-difference Hessian estimator at each iteration $n$:

$$\begin{aligned}\widetilde{G}_n &= \sum_{k=1}^d \sum_{\ell=1}^d \widetilde{G}_{n,kl}\boldsymbol{e}_k\boldsymbol{e}_\ell^\top \\ &= \frac{1}{h_n^2}\sum_{k=1}^d\sum_{\ell=1}^d\Big[\Delta_{h_n,\boldsymbol{e}_k}f(\boldsymbol{\theta}_{n-1} + h_n\boldsymbol{e}_\ell; \boldsymbol{\zeta}_n) \\ &\qquad - \Delta_{h_n,\boldsymbol{e}_k}f(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n)\Big]\boldsymbol{e}_k\boldsymbol{e}_\ell^\top.\end{aligned} \tag{16}$$

This construction can be viewed as a multi-query (with $d^2 + 1$ queries of function values at each iteration) (KW) scheme with the (I) choice of the random directions. Other choices of the search directions can be used as well, and discussions are provided in Section B.1 of the supplementary material. Each additional function-value query beyond the first one provides an estimate $\widetilde{G}_{n,kl}$ for the $(k,l)$th entry of the matrix $\widetilde{G}_n$. To reduce the computational cost in $\widetilde{G}_n$, at each iteration, the algorithm may compute a random subset of entries of $\widetilde{G}_n$ and partially inhere the remaining entries from the previous estimator $\widetilde{G}_{n-1}$. For example, each entry $\widetilde{G}_{n,k\ell}$ is updated with probability $p \in (0, 1]$. The procedure thus requires $\mathcal{O}(pd^2)$

function-value queries at each step. If we set $p = \mathcal{O}(1/d^2)$, then the query complexity is reduced to $\mathcal{O}(1)$ per step. Since the construction of (16) does not guarantee symmetry, an additional symmetrization step needs to be conducted, as

$$\widetilde{H}_n = \frac{1}{n}\sum_{i=1}^{n}\frac{\widetilde{G}_i + \widetilde{G}_i^\top}{2}. \tag{17}$$

The next lemma quantifies the estimation error of the Hessian estimator $\widetilde{H}_n$ in (17) and the proof is provided in Section B of the supplementary material.

*Lemma 4.1.* Assume Assumptions 1, 2, 3, 4 hold, or Assumptions 4, 5 hold, then $\widetilde{H}_n$ converges in probability to $H$. Additionally, if Assumptions 1, 2, 3, 4 hold with $\delta_1 = +\infty$ and $\delta = 2$, we have $\mathbb{E}\|\widetilde{H}_n - H\|^2 \le C_1 n^{-\alpha} + C_2 p^{-1} n^{-1}$.

From Lemma 4.1, as $n \to \infty$, the error rate is dominated by the $C_1 n^{-\alpha}$ term, where $\alpha$ is the parameter of the decaying step sizes.

*Remark 4.2.* In construction of the estimator of the limiting covariance matrix $H^{-1}QH^{-1}$, it is necessary to avoid the possible singularity of $\widetilde{H}_n$. A common practice is to adopt a thresholding version of $\widetilde{H}_n$ in (17). Let $U\widetilde{\Lambda}_n U^\top$ be the eigenvalue decomposition of $\widetilde{H}_n$, and define

$$\begin{aligned}\widehat{H}_n &= U\widehat{\Lambda}_n U^\top, \\ \widehat{\Lambda}_{n,kk} &= \max\left\{\kappa_1, \widetilde{\Lambda}_{n,kk}\right\},\ k = 1, 2, \ldots, d,\end{aligned} \tag{18}$$

for any positive constant $\kappa_1 < \lambda$ where $\lambda$ is defined in Assumption 1. It is guaranteed by construction that $\widehat{H}_n$ is strictly positive definite and thus invertible.

On the other hand, the estimator of Gram matrix $Q$ can be naturally constructed as

$$\widehat{Q}_n := \frac{1}{n}\sum_{i=1}^{n}\widehat{g}_{h_i,\boldsymbol{v}_i}(\boldsymbol{\theta}_{i-1};\boldsymbol{\zeta}_i)\,\widehat{g}_{h_i,\boldsymbol{v}_i}(\boldsymbol{\theta}_{i-1};\boldsymbol{\zeta}_i)^\top, \tag{19}$$

where $\widehat{g}_{h_i,\boldsymbol{v}_i}(\boldsymbol{\theta}_{i-1};\boldsymbol{\zeta}_i)$ is the (KW) update in the $i$th iteration obtained by (10). As both $\widehat{H}_n$ in (18) and $\widehat{Q}_n$ in (19) can be constructed sequentially without storing historical data,[2] the final plug-in estimator $\widehat{H}_n^{-1}\widehat{Q}_n\widehat{H}_n^{-1}$ can also be constructed in an online fashion. Based on Lemma 4.1, we obtain the following consistency result of the covariance matrix estimator $\widehat{H}_n^{-1}\widehat{Q}_n\widehat{H}_n^{-1}$.

*Theorem 4.3.* Assume Assumptions 1, 2, 3, 4 hold, or Assumptions 4, 5 hold. Under the stepsize and spacing parameter conditions specified in Theorem 3.3 or Theorem 3.9, we have $\widehat{H}_n^{-1}\widehat{Q}_n\widehat{H}_n^{-1}$ converges in probability to $H^{-1}QH^{-1}$.

Furthermore, if Assumptions 1, 2, 3, 4 hold with $\delta_1 = +\infty$ and $\delta = 2$, we have $\mathbb{E}\left\|\widehat{H}_n^{-1}\widehat{Q}_n\widehat{H}_n^{-1} - H^{-1}QH^{-1}\right\| \le Cn^{-\alpha/2}$.

We defer the technical proof to Section B of the supplementary material. Theorem 4.3 establishes the consistency and the rate of the convergence of our proposed covariance matrix estimator $\widehat{H}_n^{-1}\widehat{Q}_n\widehat{H}_n^{-1}$. Given Theorems 3.3 and 4.3, a confidence interval of the projected true parameter $\boldsymbol{w}^\top\boldsymbol{\theta}^\star$ for any $\boldsymbol{w} \in \mathbb{R}^d$ can be constructed via a projection of $\overline{\boldsymbol{\theta}}_n$ and $\widehat{H}_n^{-1}\widehat{Q}_n\widehat{H}_n^{-1}$ onto $\boldsymbol{w}$. Specifically, for a pre-specified confidence level $q$ and the corresponding $z$-score $z_{q/2}$, we obtain an asymptotic exact confidence interval as $n \to \infty$,

$$\mathbb{P}\left\{\boldsymbol{w}^\top\boldsymbol{\theta}^\star \in \left[\boldsymbol{w}^\top\overline{\boldsymbol{\theta}}_n - \frac{z_{q/2}}{\sqrt{n}}\sqrt{\boldsymbol{w}^\top\widehat{H}_n^{-1}\widehat{Q}_n\widehat{H}_n^{-1}\boldsymbol{w}},\ \boldsymbol{w}^\top\overline{\boldsymbol{\theta}}_n + \frac{z_{q/2}}{\sqrt{n}}\sqrt{\boldsymbol{w}^\top\widehat{H}_n^{-1}\widehat{Q}_n\widehat{H}_n^{-1}\boldsymbol{w}}\right]\right\} \to 1 - q.$$

### 4.1. Online Inference Without Additional Function-Value Queries

Despite the simplicity of the plug-in approach, the proposed estimator $\widehat{H}_n^{-1}\widehat{Q}_n\widehat{H}_n^{-1}$ incurs additional computational and storage cost as it requires additional function-value queries for constructing $\widehat{H}_n$. It raises a natural question: *is it possible to conduct inference only based on* (KW) *iterates* $\{\boldsymbol{\theta}_i\}_{i=1,2,\ldots}$ *without additional function-value queries*?

In this section, we provide an affirmative answer to this question, and propose an alternative online statistical inference procedure using the intermediate (KW) iterates only, without requiring any additional function-value query. Intuitively, the (AKW) estimator in (6) is constructed as the average of all intermediate (KW) iterates $\{\boldsymbol{\theta}_i\}_{i=1}^n$. If all iterates were independent and identically distributed, the asymptotic covariance could have been directly estimated by the sample covariance matrix of the iterates $\frac{1}{n}\sum_{i=1}^n(\boldsymbol{\theta}_i - \overline{\boldsymbol{\theta}})(\boldsymbol{\theta}_i - \overline{\boldsymbol{\theta}})^\top$. Unfortunately, the (KW) iterates are far from independent and indeed highly correlated. Nevertheless, the autocorrelation structure of the iterates can be carefully analyzed and used to construct the estimator of $H^{-1}QH^{-1}$.

In this article, we adopt an alternative approach to take more advantage of the autocorrelation structure by leveraging the techniques from robust testing literature (Abadir and Paruolo 1997; Kiefer et al. 2000; Lee et al. 2022a). Such an estimator is often referred to as the Fixed Bandwidth Heteroscedasticity and Autocorrelation Robust estimator (*fixed-b* HAR) in the econometrics literature to overcome the series correlation and heteroscedasticity in the error terms for the OLS estimates of the linear regression. For stochastic approximation, Lee et al. (2022a) first used and generalized this technique to construct an online statistical inference procedure, and refer to this method as *random scaling*, followed by Lee et al. (2022b) and Chen et al. (2023) for extension to quantile regression and generalized method of moments.

In particular, we present the following theorem based on a functional extension of the distributional analysis of the intermediate (KW) iterates $\{\boldsymbol{\theta}_t\}$ as a stochastic process.

*Theorem 4.4.* For any $\boldsymbol{w} \in \mathbb{R}^d$, under the assumptions in Theorems 3.3 or 3.9, we have

$$\sqrt{n}\frac{\boldsymbol{w}^\top(\overline{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star)}{\sqrt{\boldsymbol{w}^\top V_n \boldsymbol{w}}} \Longrightarrow \frac{W_1}{\sqrt{\int_0^1 (W_r - rW_1)^2\,\mathrm{d}r}}, \tag{20}$$

[2] The sequence $\widehat{Q}_n := \frac{1}{n}\sum_{i=1}^n Q_i$ with $Q_i = \widehat{g}_{h_i,\boldsymbol{v}_i}(\boldsymbol{\theta}_{i-1};\boldsymbol{\zeta}_i)\,\widehat{g}_{h_i,\boldsymbol{v}_i}(\boldsymbol{\theta}_{i-1};\boldsymbol{\zeta}_i)^\top$ can be constructed in one pass over the sequential data. In particular, we could compute $\widehat{Q}_n = \frac{1}{n}((n-1)\widehat{Q}_{n-1} + Q_i)$ sequentially.

**Table 1.** Cumulative probability table of the limiting distribution.

| Quantile | 90% | 95% | 97.5% | 99% |
|---|---|---|---|---|
| Abadir and Paruolo (1997) Table 1 | 3.875 | 5.323 | 6.747 | 8.613 |

where $V_n = \frac{1}{n^2}\sum_{i=1}^n i^2(\bar{\boldsymbol{\theta}}_i - \bar{\boldsymbol{\theta}}_n)(\bar{\boldsymbol{\theta}}_i - \bar{\boldsymbol{\theta}}_n)^\top$, and $\bar{\boldsymbol{\theta}}_i = \frac{1}{i}\sum_{\ell=1}^i \boldsymbol{\theta}_\ell$ is the average of iterates up to the $i$th iteration, and $\{W_t\}_{t\geq 0}$ is the standard one-dimensional Brownian motion.

As an important special case, when $\boldsymbol{w} = \boldsymbol{e}_k$ for $k = 1, 2, \ldots, d$, we have the convergence in each coordinate to the following pivotal limiting distribution,

$$\frac{\sqrt{n}(\bar{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}^\star_k)}{\sqrt{V_{n,kk}}} \Longrightarrow \frac{W_1}{\sqrt{\int_0^1 (W_r - rW_1)^2 \, \mathrm{d}r}}. \tag{21}$$

For the asymptotic distribution defined on the right hand side in (21), we repeat the quantiles of the distribution published by Abadir and Paruolo (1997) in Table 1.[3] Combining the asymptotic results in (21) and Table 1, we can construct coordinate-wise confidence intervals for the true parameter $\boldsymbol{\theta}^\star$. In addition, as

$$\begin{aligned} V_n &= \frac{1}{n^2}\sum_{i=1}^n i^2(\bar{\boldsymbol{\theta}}_i - \bar{\boldsymbol{\theta}}_n)(\bar{\boldsymbol{\theta}}_i - \bar{\boldsymbol{\theta}}_n)^\top \\ &= \frac{1}{n^2}\sum_{i=1}^n i^2\bar{\boldsymbol{\theta}}_i\bar{\boldsymbol{\theta}}_i^\top - \frac{\bar{\boldsymbol{\theta}}_n}{n^2}\sum_{i=1}^n i^2\bar{\boldsymbol{\theta}}_i^\top - \frac{1}{n^2}\Big(\sum_{i=1}^n i^2\bar{\boldsymbol{\theta}}_i\Big)\bar{\boldsymbol{\theta}}_n \\ &\quad + \frac{1}{n^2}\sum_{i=1}^n i^2\bar{\boldsymbol{\theta}}_n\bar{\boldsymbol{\theta}}_n^\top \end{aligned} \tag{22}$$

can be constructed in an online fashion via the iterative updates of the matrix $\sum_{i=1}^n i^2\bar{\boldsymbol{\theta}}_i\bar{\boldsymbol{\theta}}_i^\top$ and the vector $\sum_{i=1}^n i^2\bar{\boldsymbol{\theta}}_i$, the proposed online inference procedure only requires one pass over the data.

## 5. Numerical Experiments

In this numerical section, we first investigate the empirical performance of the proposed inference procedures and their corresponding coverage rates. We consider linear regression and logistic regression in this section (Examples 2.1–2.2) and conduct simulations for quantile regression (Example 3.8) in the next section. Here $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ is an iid sample with the covariate $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$ and the response $y \in \mathbb{R}$. We set the sample size $n = 10^5$ and the parameter dimension $d = 5, 20, 50$. The true model parameter $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ is selected uniformly from the unit sphere. For both models, we consider two different structures of the covariance matrices $\Sigma$: identity matrix $I_d$ and equicorrelation covariance matrix (Equicorr in the tables), that is, $\Sigma_{k\ell} = 0.2$ for all $k \neq \ell$ and $\Sigma_{kk} = 1$ for $k = 1, 2, \ldots, d$. The stepsize and spacing parameters for the first $50 \times d$ iterations are set flat to avoid a sharp change in the learning rate. Particularly, the stepsize $\eta_n = \eta_0(\max\{n, 50d\})^{-\alpha}$ and spacing parameters $h_n = h_0(\max\{n, 50d\})^{-\gamma}$ where the exponents are set to $\alpha = \gamma = 0.501$ to satisfy the assumptions in Theorem 3.3. The constant $h_0$ is set to 0.01 for both examples, and $\eta_0$ is a tunable hyperparameter. The variance of noise $\varepsilon$ in the linear regression model (Example 2.1) is set to $\sigma^2 = 1$. For both examples, the algorithm initialized from $\boldsymbol{\theta}_0$ randomly sampled spherically with radius 0.01.

**Table 2.** Estimation errors, averaged coverage rates, and average lengths of the proposed algorithm with search direction `(I)` and two function queries ($m = 1$).

| $d$ | $\Sigma$ | Estimation error (standard error) | Average coverage rate | | | Average length | | |
|---|---|---|---|---|---|---|---|---|
| | | | Plug-in | Oracle | Fixed-$b$ | Plug-in | Oracle | Fixed-$b$ |
| Linear | | | | | | | | |
| 5 | Identity | 0.015 ( 0.005 ) | 0.944 | 0.938 | 0.940 | 0.028 | 0.028 | 0.036 |
| | Equicorr | 0.017 ( 0.006 ) | 0.958 | 0.954 | 0.946 | 0.032 | 0.032 | 0.041 |
| 20 | Identity | 0.066 ( 0.010 ) | 0.943 | 0.938 | 0.928 | 0.058 | 0.056 | 0.074 |
| | Equicorr | 0.082 ( 0.014 ) | 0.938 | 0.931 | 0.923 | 0.071 | 0.068 | 0.087 |
| 50 | Identity | 0.180 ( 0.018 ) | 0.947 | 0.917 | 0.881 | 0.097 | 0.089 | 0.108 |
| | Equicorr | 0.227 ( 0.026 ) | 0.937 | 0.912 | 0.860 | 0.121 | 0.110 | 0.126 |
| Logistic | | | | | | | | |
| 5 | Identity | 0.037 ( 0.011 ) | 0.946 | 0.938 | 0.916 | 0.065 | 0.065 | 0.075 |
| | Equicorr | 0.042 ( 0.015 ) | 0.934 | 0.932 | 0.908 | 0.073 | 0.073 | 0.085 |
| 20 | Identity | 0.152 ( 0.025 ) | 0.943 | 0.937 | 0.862 | 0.128 | 0.125 | 0.136 |
| | Equicorr | 0.177 ( 0.030 ) | 0.939 | 0.935 | 0.848 | 0.154 | 0.150 | 0.158 |
| 50 | Identity | 0.404 ( 0.040 ) | 0.914 | 0.912 | 0.688 | 0.199 | 0.197 | 0.140 |
| | Equicorr | 0.495 ( 0.051 ) | 0.920 | 0.917 | 0.620 | 0.245 | 0.241 | 0.142 |

NOTE: Sample size $n = 10^5$. Corresponding standard errors are reported in the brackets. We compare the plug-in covariance estimator (plug-in) based inference (17) and fixed-$b$ HAR (fixed-$b$) based inference (21).

We first report the performance of `(AKW)` with the search direction uniformly sampled from the natural basis, referred to as `(I)` in Section 3.1. In Table 2, we present the mean and standard error of the estimation errors in the Euclidean norm (i.e., $\|\bar{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star\|$, see the first column), with 100 Monte Carlo simulations. Next, we set the nominal coverage probability as 95% and we project $\boldsymbol{\theta} \in \mathbb{R}^d$ onto $\boldsymbol{e}_k$ to construct confidence intervals, where $\boldsymbol{e}_k$ is the standard basis in $\mathbb{R}^d$ with the $k$th coordinate as 1 and the other coordinates as 0. We record the average coverage rate and the average length of the intervals among the $d$ coordinates for (1) the plug-in covariance matrix estimator[4] (16) and (2) the fixed-$b$ HAR procedure in (22), for each simulation and report the mean coverage and median interval lengths. As an oracle benchmark, we also report the length of the confidence interval with respect to the true covariance matrix $H^{-1}QH^{-1}$ of the plug-in approach and the corresponding mean coverage rate. As shown from Table 2, the coverage rate of the plug-in covariance estimator and the oracle coverage rates are very close to the desired 95% coverage, while the fixed-$b$ HAR approach is comparable in small dimension $d = 5, 20$ but has lower coverage rates for the large dimension $d = 50$. The average lengths of the plug-in method are comparable to the lengths derived from the true limiting covariance. Due to the space constraints, we relegate the additional simulation results for other choices of direction distributions and multi-query methods to Section C of the supplementary material.

[3] Since the distribution on the right hand side of (20) is symmetric, we provide one-side quantiles only.

[4] Here we use updating probability $p = 1$ for the plug-in estimation. In other words, $d^2 + 1$ queries of function values are obtained at each iteration.

**Table 3.** Estimation errors, averaged coverage rates (Coverage), median interval lengths (Length), and computation time (Time) of the proposed algorithm with search direction (I), (S), (G) defined in Section 3.1 and two function queries ($m = 1$), under quantile regression model.

| $\tau$ | Search direction | Estimation error (standard error) | Plug-in Coverage | Plug-in Length | Plug-in Time | Fixed-$b$ Coverage | Fixed-$b$ Length | Fixed-$b$ Time |
|---|---|---|---|---|---|---|---|---|
| | (I) | 0.041 ( 0.007 ) | 0.923 | 0.033 | 318.9 | 0.893 | 0.041 | 99.7 |
| | (S) | 0.042 ( 0.007 ) | 0.950 | 0.040 | 306.4 | 0.885 | 0.041 | 89.0 |
| 0.1 | (G) | 0.043 ( 0.007 ) | 0.896 | 0.032 | 344.4 | 0.873 | 0.043 | 83.8 |
| | (I) | 0.027 ( 0.004 ) | 0.903 | 0.020 | 303.7 | 0.915 | 0.028 | 94.8 |
| | (S) | 0.026 ( 0.004 ) | 0.904 | 0.020 | 295.0 | 0.919 | 0.027 | 85.4 |
| 0.5 | (G) | 0.027 ( 0.004 ) | 0.928 | 0.022 | 268.5 | 0.911 | 0.028 | 64.2 |
| | (I) | 0.041 ( 0.006 ) | 0.934 | 0.034 | 299.8 | 0.900 | 0.041 | 93.7 |
| | (S) | 0.041 ( 0.007 ) | 0.949 | 0.040 | 293.1 | 0.904 | 0.042 | 84.9 |
| 0.9 | (G) | 0.043 ( 0.007 ) | 0.892 | 0.032 | 267.3 | 0.897 | 0.043 | 64.0 |

NOTE: Sample size $n = 10^6$, dimension $d = 20$. Corresponding standard errors are reported in the brackets. We compare the plug-in covariance estimator (plug-in) based inference (17) and fixed-$b$ HAR (fixed-$b$) based inference (21).

### 5.1. Numerical Experiments on Non-smooth Loss Function

In this section, we provide simulation studies to illustrate the performance of the (AKW) estimator and inference procedures on quantile regression. Our data is generated from a linear model, $y_i = \boldsymbol{x}_i^\top \boldsymbol{\theta}^\star + \epsilon_i$, where $\{\boldsymbol{\zeta}_i = (\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is an iid sample with the covariate $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ and the noise $\{\epsilon_i\}$ follows an iid normal distribution such that $\epsilon_i \sim \mathcal{N}(-\sigma\Phi^{-1}(\tau), \sigma^2)$, $\Pr(\epsilon_i \le 0 \mid \boldsymbol{x}_i) = \tau$. Here $\Phi(\cdot)$ is the cumulative density function of standard normal distribution and $\Phi^{-1}(\cdot)$ is its inverse function. For each quantile level $\tau \in (0,1)$, the individual loss is $f(\boldsymbol{\theta}; \zeta) = \rho_\tau\left(y - \boldsymbol{x}^\top\boldsymbol{\theta}\right)$, where $\rho_\tau(z) = z(\tau - 1\{z < 0\})$. In this example, we have $H = \frac{1}{\sigma}\phi\big(\Phi^{-1}(\tau)\big)\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$ and $Q = \mathbb{E}[\boldsymbol{v}\boldsymbol{v}^\top S \boldsymbol{v}\boldsymbol{v}^\top]$ where $S = \tau(1-\tau)\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$, according to Theorem 3.9. The explicit form of $Q$ is provided in Proposition 3.5, for example, if we sample uniformly from the canonical basis with two function queries ($m = 1$), then $Q^{(\mathrm{I})} = d\,\mathrm{diag}(S)$.

In the numerical experiments below, we fix sample size $n = 10^6$, dimension $d = 20$, and the noise variance $\sigma^2 = 1$. The step-sizes and spacing parameters are set with the same specifications as the previous experiment except for $h_0 = 1$ and $\eta_0 = 0.03$. We present our results below in Table 3 with three quantile levels $\tau = 0.1, 0.5, 0.9$ and three searching direction schemes (I), (S), (G), detailed descriptions of which are presented in Proposition 3.5 of Section 3.1. Reported numbers contain the mean and standard error of the estimation error for the plug-in, and mean coverage rate and median interval lengths for plug-in and fixed-$b$ procedures of constructing confidence intervals, based on 100 Monte Carlo simulations. We further report the mean computation time for the two procedures recorded on compute nodes equipped with dual CPU sockets of 24-core Intel Cascade Lake Platinum 8268 chips. As can be inferred from the table, both procedures have good coverage rates. The fixed-$b$ HAR inference structure generates slightly larger confidence intervals with computes four times faster since it does not need additional function queries to compute the Hessian estimator.

## Supplementary Materials

The supplementary materials include: (1) all the proof details for the lemmas and main theorems, (2) additional results of numerical experiments.

## Acknowledgments

We thank Professor Yuan Liao and the anonymous reviewers and editors for constructive comments that improved our article.

## Disclosure Statement

The authors report there are no competing interests to declare.

## Funding

Xi Chen thanks for support from NSF via IIS-1845444.

## References

Abadir, K. M., and Paruolo, P. (1997), "Two Mixed Normal Densities from Cointegration Analysis," *Econometrica*, 65, 671–680. [8,9]

Agarwal, A., Dekel, O., and Xiao, L. (2010), "Optimal Algorithms for Online Convex Optimization with Multi-Point Bandit Feedback," in *Conference on Learning Theory*, pp. 28–40. [2]

Agarwal, A., Foster, D. P., Hsu, D. J., Kakade, S. M., and Rakhlin, A. (2011), "Stochastic Convex Optimization with Bandit Feedback," in *Advances in Neural Information Processing Systems*, pp. 1035–1043. [2]

Blum, J. R. (1954), "Multidimensional Stochastic Approximation Methods," *The Annals of Mathematical Statistics*, 25, 737–744. [2,3]

Broadie, M., Cicek, D., and Zeevi, A. (2011), "General Bounds and Finite-Time Improvement for the Kiefer-Wolfowitz Stochastic Approximation Algorithm," *Operations Research*, 59, 1211–1224. [2]

Chao, S.-K., and Cheng, G. (2019), "A Generalization of Regularized Dual Averaging and its Dynamics," arXiv preprint arXiv:1909.10072. [3]

Chen, H. (1988), "Lower Rate of Convergence for Locating a Maximum of a Function," *The Annals of Statistics*, 16, 1330–1334. [2]

Chen, H., Lu, W., and Song, R. (2021), "Statistical Inference for Online Decision Making via Stochastic Gradient Descent," *Journal of the American Statistical Association*, 116, 708–719. [3]

Chen, H.-F., Duncan, T. E., and Pasik-Duncan, B. (1999), "A Kiefer-Wolfowitz Algorithm with Randomized Differences," *IEEE Transactions on Automatic Control*, 44, 442–453. [2,3]

Chen, X., Lee, J. D., Tong, X. T., and Zhang, Y. (2020), "Statistical Inference for Model Parameters in Stochastic Gradient Descent," *The Annals of Statistics*, 48, 251–273. [2]

Chen, X., Lee, S., Liao, Y., Seo, M. H., Shin, Y., and Song, M. (2023), "SGMM: Stochastic Approximation to Generalized Method of Moments," *Journal of Financial Econometrics*, forthcoming. [3,8]

Conn, A. R., Scheinberg, K., and Vicente, L. N. (2009), *Introduction to Derivative-Free Optimization*, Philadelphia, PA: Society for Industrial and Applied Mathematics. [1]

Dippon, J. (2003), "Accelerated Randomized Stochastic Optimization," *The Annals of Statistics*, 31, 1260–1281. [2]

Duchi, J. C., Jordan, M. I., Wainwright, M. J., and Wibisono, A. (2015), "Optimal Rates for Zero-Order Convex Optimization: The Power of Two Function Evaluations," *IEEE Transactions on Information Theory*, 61, 2788–2806. [2,5]

Duchi, J. C., and Ruan, F. (2021), "Asymptotic Optimality in Stochastic Optimization," *The Annals of Statistics*, 49, 21–48. [3]

Fabian, V. (1967), "Stochastic Approximation of Minima with Improved Asymptotic Speed," *The Annals of Mathematical Statistics*, 38, 191–200. [2]

——— (1968), "On Asymptotic Normality in Stochastic Approximation," *The Annals of Mathematical Statistics*, 39, 1327–1332. [5]

——— (1980), "Stochastic Approximation Methods for Constrained and Unconstrained Systems," *SIAM Review*, 22, 382–384. [2]

Fang, Y., Xu, J., and Yang, L. (2018), "Online Bootstrap Confidence Intervals for the Stochastic Gradient Descent Estimator," *Journal of Machine Learning Research*, 19, 1–21. [3]

Flaxman, A. D., Kalai, A. T., and Brendan McMahan, H. (2005), "Online Convex Optimization in the Bandit Setting: Gradient Descent Without

Gradient," in *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 385–394. [1]

Ghadimi, S., and Lan, G. (2013), "Stochastic First-and Zeroth-Order Methods for Nonconvex Stochastic Programming," *SIAM Journal on Optimization*, 23, 2341–2368. [2]

Hall, P., and Heyde, C. C. (1980), *Martingale Limit Theory and its Application*, New York: Academic press. [2]

Hall, P., and Molchanov, I. (2003), "Sequential Methods for Design-Adaptive Estimation of Discontinuities in Regression Curves and Surfaces," *The Annals of Statistics*, 31, 921–941. [2]

He, Y., Fu, M. C., and Marcus, S. I. (2003), "Convergence of Simultaneous Perturbation Stochastic Approximation for Nondifferentiable Optimization," *IEEE Transactions on Automatic Control*, 48, 1459–1463. [3]

Jamieson, K. G., Nowak, R., and Recht, B (2012), "Query Complexity of Derivative-Free Optimization," in *Advances in Neural Information Processing Systems*, pp. 2672–2680. [2]

Joshi, S., and Boyd, S. (2008), "Sensor Selection via Convex Optimization," *IEEE Transactions on Signal Processing*, 57, 451–462. [1]

Kiefer, J., and Wolfowitz, J. (1952), "Stochastic Estimation of the Maximum of a Regression Function," *The Annals of Mathematical Statistics*, 23, 462–466. [1,2]

Kiefer, N. M., Vogelsang, T. J., and Bunzel, H. (2000), "Simple Robust Testing of Regression Hypotheses," *Econometrica*, 68, 695–714. [8]

Koronacki, J. (1975), "Random-Seeking Methods for the Stochastic Unconstrained Optimization," *International Journal of Control*, 21, 517–527. [3]

Lee, S., Liao, Y., Seo, M. H., and Shin, Y. (2022a), "Fast and Robust Online Inference with Stochastic Gradient Descent via Random Scaling," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36), pp. 7381–7389. [2,8]

——— (2022b), "Fast Inference for Quantile Regression with Tens of Millions of Observations," Available at SSRN 4263158. [3,8]

Liang, T., and Su, W. (2019), "Statistical Inference for the Population Landscape via Moment-Adjusted Stochastic Gradients," *Journal of the Royal Statistical Society*, Series B, 81, 431–456. [3]

Mokkadem, A., and Pelletier, M. (2007), "A Companion for the Kiefer–Wolfowitz–Blum Stochastic Approximation Algorithm," *The Annals of Statistics*, 35, 1749–1772. [2]

Moulines, E., and Bach, F. (2011), "Non-asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning," in *Advances in Neural Information Processing Systems* (Vol. 24), pp. 451–459. [5]

Nesterov, Y., and Spokoiny, V. (2017), "Random Gradient-Free Minimization of Convex Functions," *Foundations of Computational Mathematics*, 17, 527–566. [1,2]

Polyak, B. T., and Juditsky, A. B. (1992), "Acceleration of Stochastic Approximation by Averaging," *SIAM Journal on Control and Optimization*, 30, 838–855. [2,5]

Polyak, B. T., and Tsybakov, A. B. (1990), "Optimal Order of Accuracy of Search Algorithms in Stochastic Optimization," *Problemy Peredachi Informatsii*, 26, 126–133. [2]

Robbins, H., and Monro, S. (1951), "A Stochastic Approximation Method," *The Annals of Mathematical Statistics*, 22, 400–407. [1]

Ruppert, D. (1982), "Almost Sure Approximations to the Robbins-Monro and Kiefer-Wolfowitz Processes with Dependent Noise," *The Annals of Probability*, 10, 178 – 187. [2]

——— (1988), "Efficient Estimations from a Slowly Convergent Robbins-Monro Process," Technical Report, Cornell University Operations Research and Industrial Engineering. [2]

Shamir, O. (2017), "An Optimal Algorithm for Bandit and Zero-Order Convex Optimization with Two-Point Feedback," *Journal of Machine Learning Research*, 18, 1703–1713. [1,2]

Shi, C., Song, R., Lu, W., and Li, R. (2021), "Statistical Inference for High-Dimensional Models via Recursive Online-Score Estimation," *Journal of the American Statistical Association*, 116, 1307–1318. [3]

Spall, J. C. (1992), "Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation," *IEEE Transactions on Automatic Control*, 37, 332–341. [2,3]

——— (2000), "Adaptive Stochastic Approximation by the Simultaneous Perturbation Method," *IEEE Transactions on Automatic Control*, 45, 1839–1853. [2,3]

Su, W. J., and Zhu, Y. (2018), "Uncertainty Quantification for Online Learning and Stochastic Approximation via Hierarchical Incremental Gradient Descent," arXiv preprint arXiv:1802.04876. [3]

Toulis, P., and Airoldi, E. M. (2017), "Asymptotic and Finite-Sample Properties of Estimators based on Stochastic Gradients," *The Annals of Statistics*, 45, 1694–1727. [3]

van der Vaart, A. W. (2000). *Asymptotic Statistics* (Vol. 3), Cambridge: Cambridge University Press. [7]

Wainwright, M. J., and Jordan, M. I. (2008), *Graphical Models, Exponential Families, and Variational Inference*, Hanover, MA: Now Publishers Inc. [1]

Wang, Y., Du, S., Balakrishnan, S., and Singh, A. (2018), "Stochastic Zeroth-Order Optimization in High Dimensions," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 1356–1365, PMLR. [2]

Zhu, W., Chen, X., and Wu, W. B. (2023), "Online Covariance Matrix Estimation in Stochastic Gradient Descent," *Journal of the American Statistical Association*, 118, 393–404. [2]