Online Statistical Inference for Stochastic Optimization via Kiefer-Wolfowitz Methods

Abstract

This paper 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 trade-off 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.

Keywords: Asymptotic normality, Kiefer-Wolfowitz stochastic approximation, online inference, stochastic optimization

1 Introduction

Stochastic optimization algorithms, introduced by Robbins and Monro (1951); 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 et al., 2009; Nesterov and Spokoiny, 2017). We briefly highlight a few of them to motivate our paper.

- In some bandit problems, one may only have black-box access to individual objective values but not to their gradients (Flaxman et al., 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.

- In some statistical problems such as quantile regression and its variants (Koenker, 2005), the objective function is not differentiable. Extending the gradient definition to nonsmooth functions is generally nontrivial, and techniques of defining sets of local differential characteristics suffer from the incompleteness of chain rule in complex problems (Nesterov, 2005).

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

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

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,

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

where $\{\eta_n\}$ is a positive non-increasing step-size sequence, and $g(\boldsymbol{\theta}; \boldsymbol{\zeta})$ denotes the stochastic gradient, i.e., $g(\boldsymbol{\theta}; \boldsymbol{\zeta}) = \nabla f(\boldsymbol{\theta}; \boldsymbol{\zeta})$. In the scenarios that direct gradient measurements are inaccessible to practitioners, the Kiefer and Wolfowitz (1952) algorithm (KW) becomes the natural choice, as

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

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, 1978); Hall and Heyde (1980); Ruppert (1982); Chen (1988); Polyak and Tsybakov (1990); Spall (1992); Chen et al. (1999); Spall (2000); Hall and Molchanov (2003); Dippon (2003); Mokkadem and Pelletier (2007); Broadie et al. (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 et al., 2010, 2011; Jamieson et al., 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* $\overline{\boldsymbol{\theta}}_n^{(\mathtt{RM})} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\theta}_i^{(\mathtt{RM})}$ by

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

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

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

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 paper, we study the (KW) algorithm (3) with random search directions $\{\boldsymbol{v}_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{P}_{\boldsymbol{v}}$, i.e., 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 utilize 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 is inspired by the "random scaling" method proposed in Lee et al. (2022) 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 et al. (2021) further extended the batch-means method in Chen et al. (2020) to a fully online covariance estimator. Lee et al. (2022) extended the results in Polyak and Juditsky (1992) to a functional central limit theorem and utilize it to propose a novel online inference procedure that allows for efficient implementation. Fang et al. (2017) 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 ℓ_1-penalized problems. Shi et al. (2021) developed an online estimation procedure for high-dimensional statistical inference. Chen et al. (2021) studied statistical inference of online decision-making problems via SGD in a contextual bandit setting.

1.1 Notations and organization of the paper

We write vectors in boldface letters (e.g., $\boldsymbol{\theta}$ and $\boldsymbol{v}$) and scalers in lightface letters (e.g., η). For any positive integer n, we use $[n]$ as a shorthand for the discrete set $\{1, 2, \cdots, n\}$. 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,ℓ)-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 semi-definite. 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 $\widehat{\boldsymbol{\theta}}^{(\mathtt{ERM})}$ for the offline empirical risk minimizer, i.e., $\widehat{\boldsymbol{\theta}}^{(\mathtt{ERM})} = \mathrm{argmin}_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{\theta}; \boldsymbol{\zeta}_i)$. As we focus on the `(KW)` scheme in this paper, 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, i.e.,

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

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

The remainder of the paper 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. 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. Further discussions are provided in Section 6 and 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

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

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 et al. (1999); Spall (2000); He et al. (2003). These random direction methods can reduce the bias in gradient estimates as compared to their non-random counterparts. In the following, we write the (KW) algorithm with general random search directions, as in (7),

$$\begin{aligned} &\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} - \eta_n \widehat{g}_{h_n, \boldsymbol{v}_n}(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n), \\ &\quad \text{where } \widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta}; \boldsymbol{\zeta}) := \frac{1}{h} \Delta_{h,\boldsymbol{v}} f(\boldsymbol{\theta}; \boldsymbol{\zeta}) \boldsymbol{v} = \frac{f(\boldsymbol{\theta} + h\boldsymbol{v}; \boldsymbol{\zeta}) - f(\boldsymbol{\theta}; \boldsymbol{\zeta})}{h} \boldsymbol{v}. \end{aligned} \tag{10}$$

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

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

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 paper. 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 i.i.d. 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*

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

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*

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

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

Example 2.3 (Quantile Regression)**.** *Consider a quantile 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* i.i.d. *sample of $\boldsymbol{\zeta} = (\boldsymbol{x}, y)$ and the noise satisfies $\Pr(\epsilon_i \leq 0 | \boldsymbol{x}_i) = \tau$. 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 ρ_τ is non-differentiable, the* `(KW)` *gradient estimator $\widehat{g}_{h,\boldsymbol{v}}$ is well-defined and takes the following form,*

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

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, i.e., $\mathbb{E}_{\boldsymbol{\zeta}} g(\boldsymbol{\theta}; \boldsymbol{\zeta}) = \nabla F(\boldsymbol{\theta})$. 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 2.4. *We assume that the population loss function $F(\cdot)$ is twice continuously differentiable and L_f-smooth, i.e., $\nabla^2 F(\boldsymbol{\theta}) \preceq L_f I_d$ 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*

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

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 2.4 is provided in Section A of the supplementary material. To reduce the bias in the `(KW)` gradient, Lemma 2.4 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 $f(\boldsymbol{\theta}; \boldsymbol{\zeta})$. To simplify the illustration, we present the following assumptions first, even though the

main results remain to hold under some weakened assumptions. The relaxation is discussed below, and detailed theoretical derivation is relegated to Section A.2 of the supplementary material.

Assumption 1. *The population loss function $F(\boldsymbol{\theta})$ is twice continuously differentiable. Moreover, there exists $L_f > \lambda > 0$, such that, $\lambda I_d \preceq \nabla^2 F(\boldsymbol{\theta}) \preceq L_f I_d$ for any $\boldsymbol{\theta} \in \mathbb{R}^d$.*

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

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

$$\mathbb{E}\left\|\nabla^2 f(\boldsymbol{\theta}; \boldsymbol{\zeta}_n) - \nabla^2 f(\boldsymbol{\theta'}; \boldsymbol{\zeta}_n)\right\|^2 \leq L_h \|\boldsymbol{\theta} - \boldsymbol{\theta'}\|^2, \qquad \mathbb{E}\left\|[\nabla^2 f(\boldsymbol{\theta}^\star; \boldsymbol{\zeta}_n)]^2 - H^2\right\| \leq L_p,$$

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

Assumption 4. *We adopt i.i.d. 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 requires the population loss function $F(\cdot)$ to be λ-strongly convex and L_f-smooth, which is widely assumed in the existing literature of statistical inference on stochastic optimization (Polyak and Juditsky, 1992; Chen et al., 2020). Note that it is possible to replace this assumption with a weaker one that only assumes local strong convexity in the neighborhood of the true parameter $\boldsymbol{\theta}^\star$(Su and Zhu, 2018; Duchi and Ruan, 2021). This weaker condition satisfies the setting of logistic regression (Example 2.2). To highlight the main contributions of the paper, we discuss its relaxed form of Assumption 1 in Section A.2 of the supplementary material. 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. Relaxation to this assumption can be made to handle nonsmooth loss functions $f(\boldsymbol{\theta}; \boldsymbol{\zeta})$, such as the quantile regression as described in Example 2.3. We defer the discussion of this weaker assumption to Section A.2 in the supplementary material. The statements in Assumption 3 introduce the Lipschitz continuity condition and the concentration condition on the Hessian matrix. Assumption 4 guarantees that

the (KW) gradient $\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\boldsymbol{\zeta})$ is an asymptotically unbiased estimator of $\nabla F(\boldsymbol{\theta})$ when the spacing parameter h_n decreases to 0, as suggested by Lemma 2.4. The moment condition of $\boldsymbol{v}$ in Assumption 4 is imposed for technical simplicity and could be possibly weakened. 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 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 some 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. Furthermore, for sufficiently large n, we have for $0 < \delta \leq 2$,*

$$\mathbb{E}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}^\star\|^{2+\delta} \leq C n^{-\alpha(2+\delta)/2}, \tag{12}$$

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 in (12) on the parameters and the initial value $\boldsymbol{\theta}_0$ are provided in Remark A.1. A similar error bound is given by Duchi et al. (2015) in terms of the function values for $\delta = 0$. We generalize the result to the $(2+\delta)$-moment error bound on the parameter $\boldsymbol{\theta}$, where $\delta \in (0, 2]$ is assumed in Assumption 2 for the purpose of derivation of asymptotic normality. 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 $\boldsymbol{\theta}^\star$, i.e., $S := \mathbb{E}\left[\nabla f(\boldsymbol{\theta}^\star;\boldsymbol{\zeta})\nabla f(\boldsymbol{\theta}^\star;\boldsymbol{\zeta})^\top\right]$. 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 to 4, we have*

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

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 to 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,*

$$\sqrt{n}\ \left(\overline{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star\right) \Longrightarrow \mathcal{N}\left(\mathbf{0}, H^{-1}QH^{-1}\right), \qquad as \quad n \to \infty, \tag{13}$$

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

$$H^{-1}QH^{-1} - H^{-1}SH^{-1} = H^{-1}\mathbb{E}_{\boldsymbol{v}}\left[(\boldsymbol{v}\boldsymbol{v}^\top - I_d)S(\boldsymbol{v}\boldsymbol{v}^\top - I_d)\right]H^{-1} \succ 0, \tag{14}$$

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 $H^{-1}SH^{-1}$.

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 Λ is a diagonal matrix. Using the proof in Fabian (1968), we establish the following asymptotic distribution for $\boldsymbol{\theta}_n^{(\mathrm{KW})}$,*

$$n^{\alpha/2}(\boldsymbol{\theta}_n^{(\mathrm{KW})} - \boldsymbol{\theta}^\star) \Longrightarrow \mathcal{N}(0, \Sigma), \tag{15}$$

where each (k, ℓ)-th entry of the covariance matrix Σ is,

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

[1] Note that the asymptotic covariance $H^{-1}SH^{-1}$ 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.

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 $\boldsymbol{\theta}_n^{(\mathtt{KW})}$ without averaging converges at a slower rate $n^{-\alpha/2}$ than that of `(AKW)` *in Theorem 3.3.*

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 $\{\sqrt{d}\boldsymbol{e}_1, \sqrt{d}\boldsymbol{e}_2, \ldots, \sqrt{d}\boldsymbol{e}_d\}$ 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 $\{\sqrt{d}\boldsymbol{u}_1, \sqrt{d}\boldsymbol{u}_2, \ldots, \sqrt{d}\boldsymbol{u}_d\}$, where $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_d\}$ is an arbitrary *orthonormal basis* of $\mathbb{R}^d$, i.e., 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)` Non-uniform 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 $\mathcal{P}_{\boldsymbol{v}}$, we have*

`(G)` *Gaussian:* $Q^{(\mathtt{G})} = (2S + \mathrm{tr}(S) I_d)$.

(S) *Spherical:* $Q^{(\mathtt{S})} = \frac{d}{d+2}(2S + \mathrm{tr}(S)I_d)$.

(I) *Uniform in the canonical basis:* $Q^{(\mathtt{I})} = d\,\mathrm{diag}(S)$.

(U) *Uniform in an arbitrary orthonormal basis* U*:* $Q^{(\mathtt{U})} = dU\mathrm{diag}(U^\top S U)U^\top$.

(P) *Non-uniform in a natural coordinate basis:* $Q^{(\mathtt{P})} = \mathrm{diag}(S_{11}/p_1, S_{22}/p_2, \ldots, S_{dd}/p_d)$.

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 $Q^{(\mathtt{G})} = (d+2)I_d$ and $Q^{(\mathtt{S})} = Q^{(\mathtt{I})} = Q^{(\mathtt{U})} = dI_d$ and $Q^{(\mathtt{P})} = \mathrm{diag}(p_1^{-1}, p_2^{-1}, \ldots, p_d^{-1}) \succ I_d$ where $p_1 + p_2 + \cdots + p_d = 1$. Several additional findings and implications of Proposition 3.5 are discussed in Section A.3 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),

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

Here we first consider using the same sampling distribution across m queries and n iterations. In other words, $\boldsymbol{v}_i^{(j)}$ is sampled *i.i.d.* 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 the 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* $n \to \infty$*,*

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

Theorem 3.6 illustrates a trade-off 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 $\{\boldsymbol{v}_i^{(1)}, \boldsymbol{v}_i^{(2)}, \ldots, \boldsymbol{v}_i^{(m)}\}$ 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 of its asymptotic distribution.

Theorem 3.7. *Under the assumptions in Theorem 3.3, and the direction vectors in all iterations* $\{\widetilde{V}_i\}_{i=1}^n$ *are* i.i.d. *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* $\overline{\boldsymbol{\theta}}_n^{(m,\mathtt{WOR})}$*, has the following asymptotic distribution, as* $n \to \infty$*,*

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

By comparing the asymptotic covariance matrices in Theorems 3.6 and 3.7, $Q_m^{(\mathtt{WOR})}$ 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 $Q_m^{(\mathtt{WOR})} = S$. Therefore, the $(d+1)$-query (AKW) estimator $\overline{\boldsymbol{\theta}}_n^{(m,\mathtt{WOR})}$ achieves the same limiting covariance as that of the averaged (RM) estimator. Furthermore, when the model is well-specified, the limiting covariance matrix $H^{-1}SH^{-1} = H^{-1}$ achieves the Cramér-Rao lower bound. This result indicates that the $(d+1)$-query (AKW) estimator $\overline{\boldsymbol{\theta}}_n$ is asymptotically efficient (van der Vaart, 2000).

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 (13). A direct way is to construct a pair of consistent estimators $\widehat{H}$ and $\widehat{Q}$ of H and Q, respectively, and estimate the asymptotic covariance by the *plug-in* estimator

$\widehat{H}^{-1}\widehat{Q}\widehat{H}^{-1}$. 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:

$$\widetilde{G}_n = \sum_{k=1}^{d}\sum_{\ell=1}^{d} \widetilde{G}_{n,kl}\boldsymbol{e}_k\boldsymbol{e}_\ell^\top = \frac{1}{h_n^2}\sum_{k=1}^{d}\sum_{\ell=1}^{d}\left[\Delta_{h_n,\boldsymbol{e}_k} f(\boldsymbol{\theta}_{n-1}+h_n\boldsymbol{e}_\ell;\boldsymbol{\zeta}_n) - \Delta_{h_n,\boldsymbol{e}_k} f(\boldsymbol{\theta}_{n-1};\boldsymbol{\zeta}_n)\right]\boldsymbol{e}_k\boldsymbol{e}_\ell^\top, \quad (16)$$

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

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

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. *Under the assumptions in Theorem 3.3, we have*

$$\mathbb{E}\|\widetilde{H}_n - H\|^2 \le C_1 n^{-\alpha} + C_2 p^{-1} n^{-1}. \quad (18)$$

From Lemma 4.1, as $n \to \infty$, the error rate is dominated by the $C_1 n^{-\alpha}$ term, where α 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*

$$\widehat{H}_n = U\widehat{\Lambda}_n U^\top, \quad \widehat{\Lambda}_{n,kk} = \max\left\{\kappa_1, \widetilde{\Lambda}_{n,kk}\right\}, \ k = 1, 2, \ldots, d, \tag{19}$$

for any positive constant $\kappa_1 < \lambda$ where λ 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

$$\widehat{Q}_n := \frac{1}{n}\sum_{i=1}^{n} \widehat{g}_{h_i,\boldsymbol{v}_i}(\boldsymbol{\theta}_{i-1};\boldsymbol{\zeta}_i)\,\widehat{g}_{h_i,\boldsymbol{v}_i}(\boldsymbol{\theta}_{i-1};\boldsymbol{\zeta}_i)^\top, \tag{20}$$

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 (19) and $\widehat{Q}_n$ in (20) 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 to 4 hold for $\delta = 2$. 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 $\gamma \in \left(\frac{1}{2}, 1\right)$. 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 can obtain an asymptotic exact confidence interval as $n \to \infty$,

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

[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 only with one-pass over the sequential data. In particular, we could compute $\widehat{Q}_n$ sequentially as $\widehat{Q}_n = \frac{1}{n}((n-1)\widehat{Q}_{n-1} + Q_i)$.

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 utilized to construct the estimator of $H^{-1}QH^{-1}$.

In this paper, 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., 2022). Such an estimator is often referred to as the Fixed Bandwidth Heteroskedasticity and Autocorrelation Robust estimator (*fixed-b* HAR) in the econometrics literature. The *fixed-b* HAR estimator is able to overcome the series correlation and heteroskedasticity in the error terms for the OLS estimates of the linear regression (e.g. Kiefer et al. (2000)). For the `(RM)` scheme, Lee et al. (2022) utilized and generalized this technique to construct an online statistical inference procedure, and refer to this method as the *random scaling* method.

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 Theorem 3.3, we have*

$$\sqrt{n}\frac{\boldsymbol{w}^\top(\overline{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^\star)}{\sqrt{\boldsymbol{w}^\top V_n\boldsymbol{w}}}\Longrightarrow\frac{W_1}{\sqrt{\int_0^1(W_r-rW_1)^2\,\mathrm{d}r}},\tag{21}$$

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

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

Table 1: Cumulative probability table of the limiting distribution.

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,

$$\frac{\sqrt{n}(\overline{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}^\star_k)}{\sqrt{V_{n,kk}}} \Longrightarrow \frac{W_1}{\sqrt{\int_0^1 (W_r - rW_1)^2 \, \mathrm{d}r}}, \tag{22}$$

For the asymptotic distribution defined on the right hand side in (22), we repeat the quantiles of the distribution published by Abadir and Paruolo (1997) in Table 1[3]. Combining the asymptotic results in (22) and Table 1, we can construct coordinate-wise confidence intervals for the true parameter $\boldsymbol{\theta}^\star$. In addition, as

$$V_n = \frac{1}{n^2}\sum_{i=1}^n i^2(\overline{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_n)(\overline{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_n)^\top = \frac{1}{n^2}\sum_{i=1}^n i^2\overline{\boldsymbol{\theta}}_i\overline{\boldsymbol{\theta}}_i^\top - \frac{2}{n^2}\overline{\boldsymbol{\theta}}_n\sum_{i=1}^n i^2\overline{\boldsymbol{\theta}}_i^\top + \frac{1}{n^2}\sum_{i=1}^n i^2\overline{\boldsymbol{\theta}}_n\overline{\boldsymbol{\theta}}_n^\top \tag{23}$$

can be constructed in an online fashion via the iterative updates of the matrix $\sum_{i=1}^n i^2\overline{\boldsymbol{\theta}}_i\overline{\boldsymbol{\theta}}_i^\top$ and the vector $\sum_{i=1}^n i^2\overline{\boldsymbol{\theta}}_i$, 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 models (Examples 2.1–2.2) where $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ is an *i.i.d.* sample with the covariate $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$ and the response $y \in \mathbb{R}$. The true model parameter $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ is selected uniformly from the unit sphere before the experiments. For both models, we consider two different structures of the covariance matrices Σ: identity matrix I_d and equicorrelation covariance matrix (Equicorr in the tables), i.e., $\Sigma_{k\ell} = 0.2$ for all $k \neq \ell$ and $\Sigma_{kk} = 1$. The parameter α in the step size is specified to $\alpha = 0.501$. The variance of noise ϵ in the linear regression model (Example 2.1) is set to $\sigma^2 = 0.2$.

[3]Since the distribution on the right hand side of (21) is symmetric, we only provide one-side quantiles in the table.

d	Σ	Estimation error		Average coverage rate			Average length		
		Parameter	Plug-in Cov.	Plug-in	Fixed-b	Oracle	Plug-in	Fixed-b	Oracle
Linear regression									
5	Identity	0.0031	0.0384	0.9448	0.9464	0.9436	1.7555	2.1915	1.7533
		(0.0010)	(0.0106)	(0.1035)	(0.1174)	(0.1040)	(0.0082)	(0.4184)	-
	Equicorr	0.0035	0.0342	0.9428	0.9488	0.9412	2.0109	2.4895	2.0078
		(0.0012)	(0.0092)	(0.1096)	(0.1195)	(0.1102)	(0.0097)	(0.5323)	-
20	Identity	0.0135	0.1126	0.9319	0.9039	0.9288	3.5337	3.7424	3.5065
		(0.0023)	(0.0190)	(0.0594)	(0.0657)	(0.0616)	(0.0164)	(0.4292)	-
	Equicorr	0.0172	0.1124	0.9194	0.9014	0.9170	4.3140	4.5582	4.2753
		(0.0029)	(0.0199)	(0.0644)	(0.0674)	(0.0656)	(0.0207)	(0.5681)	-
100	Identity	0.0748	0.5707	0.9309	0.7501	0.9012	8.6675	7.6819	7.8397
		(0.0062)	(0.0648)	(0.0261)	(0.0397)	(0.0336)	(0.1081)	(0.4924)	-
	Equicorr	0.0921	0.5615	0.9331	0.7435	0.9044	10.7701	9.0193	9.7508
		(0.0076)	(0.0647)	(0.0250)	(0.0418)	(0.0320)	(0.1400)	(0.6161)	-
Logistic regression									
5	Identity	0.0265	0.0587	0.9432	0.9360	0.9440	3.1136	3.3775	3.1078
		(0.0115)	(0.0434)	(0.1219)	(0.1685)	(0.1148)	(0.0936)	(0.2074)	-
	Equicorr	0.0299	0.0697	0.9440	0.9364	0.9464	3.3620	3.8028	3.2580
		(0.0131)	(0.0514)	(0.1196)	(0.1566)	(0.1207)	(0.1057)	(0.2558)	-
20	Identity	0.0728	0.1030	0.9418	0.8956	0.9403	4.8751	5.1763	4.8374
		(0.0124)	(0.0250)	(0.0532)	(0.1156)	(0.0540)	(0.1973)	(0.4362)	-
	Equicorr	0.0799	0.1213	0.9383	0.8949	0.9369	5.6873	5.7532	5.6356
		(0.0146)	(0.0359)	(0.0577)	(0.1106)	(0.0561)	(0.1715)	(0.4064)	-
100	Identity	0.2440	0.5236	0.9673	0.7022	0.9082	12.0661	8.7892	10.3175
		(0.0211)	(0.1646)	(0.0193)	(0.0838)	(0.0295)	(0.4642)	(0.6000)	-
	Equicorr	0.2867	0.7685	0.9608	0.6950	0.9041	12.7375	9.4884	10.4868
		(0.0253)	(0.2933)	(0.0185)	(0.0728)	(0.0314)	(0.6870)	(0.6170)	-

Table 2: Estimation errors, averaged coverage rates, and average lengths of the proposed algorithm with search direction `(I)` and two function queries ($m = 1$). 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 (22).

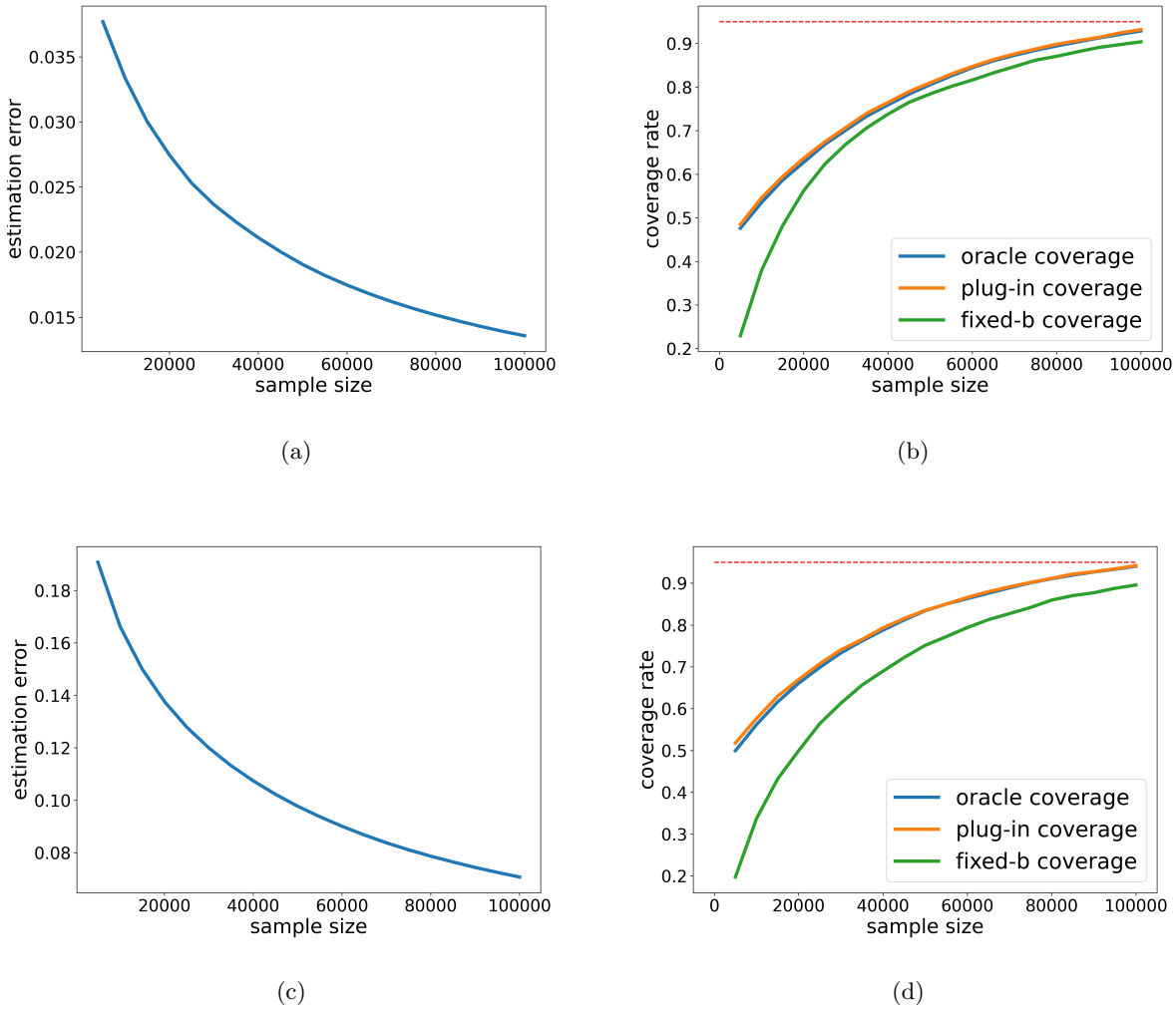


Figure 3: Convergence of the parameter estimation error $\|\overline{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star\|$ and coverage rates v.s. the sample size n when $d = 20$ and the population design matrix $\Sigma = I$. Plots (a) to (b) show the cases of linear regression and plots (c) to (d) show the cases of logistic regression. Dashed lines in plots (b) and (d) correspond to the nominal 95% coverage.

5.1 Estimation errors of (AKW) and the performance of inference procedures

We set the sample size $n = 10^5$ and the parameter dimension $d = 5, 20, 100$. 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 estimation error for the parameter $\boldsymbol{\theta}^\star$ in the Euclidean norm and the relative error of the plug-in covariance estimator in the spectral norm (see the first two columns), with 100 Monte-Carlo simulations,

$$\frac{\|\overline{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star\|}{\|\boldsymbol{\theta}^\star\|}, \quad \frac{\|\widehat{H}_n^{-1}\widehat{Q}_n\widehat{H}_n^{-1} - H^{-1}QH^{-1}\|}{\|H^{-1}QH^{-1}\|}. \tag{24}$$

Next, we set the nominal coverage probability as 95% and we project $\boldsymbol{\theta} \in \mathbb{R}^d$ onto $\boldsymbol{w} = (1, 1, \ldots, 1)^\top/\sqrt{d}$ to construct confidence intervals. In particular, we report the performance of the confidence interval with the average coverage rate and the average length of the intervals for (1) the plug-in covariance matrix estimator[4] (16) and (2) the fixed-b HAR procedure in (23). 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 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 $d = 100$. The average lengths of both methods are comparable to the lengths derived by the true limiting covariance.

Then, we fix $d = 20$ and the identity design matrix $\Sigma = I$. We present in Figure 3 the parameter estimation error $\|\overline{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star\|$ as the sample size n grows (see subplots (a) and (c) in Figure 3, for linear and logistic regression, respectively). In subplots (b) and (d) of Figure 3, we show the coverage rates for the plug-in and fixed-b HAR approaches as the sample size n increases. As one can see, coverage rates of the plug-in approach almost match the oracle case using the true asymptotic covariance matrix $H^{-1}QH^{-1}$. For the linear regression case, the plug-in and fixed-b HAR approaches are comparable. For the logistic regression case, the coverage rate of the fixed-b HAR inference procedure is slightly inferior than that of the plug-in method. On the other hand, the fixed-b HAR approach does not require additional function queries for the explicit estimation

[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. In Section 5.2 below, we extend the comparison for different p.

of the Hessian matrix. Additional simulation results for the equicorrelation design are relegated to Figure C.1 in the supplementary material.

Σ	Estimator	Comp. time	Estimation error		Average coverage rate		Average length	
			Hessian	Cov.	Estimator	Oracle	Estimator	Oracle
Identity								
	Plug-in $p = 1/400$	4.74s	0.1780	0.3179	0.8965	0.9288	3.6570	3.5065
			(0.0115)	(0.0423)	(0.0696)	(0.0616)	(0.0195)	-
	Plug-in $p = 1/20$	25.42s	0.0393	0.1503	0.9244	0.9288	3.5511	3.5065
			(0.0043)	(0.0282)	(0.0665)	(0.0616)	(0.0169)	-
	Plug-in $p = 1$	510.53s	0.0271	0.1126	0.9319	0.9288	3.5337	3.5065
			(0.0021)	(0.0190)	(0.0594)	(0.0616)	(0.0164)	-
	Fixed-b	2.82s	-	-	0.9039	0.9288	3.7424	3.5065
			-	-	(0.0657)	(0.0616)	(0.4292)	-
Equicorr								
	Plug-in $p = 1/400$	4.78s	0.0381	0.4211	0.8815	0.9170	4.4547	4.2753
			(0.0043)	(0.0421)	(0.0753)	(0.0656)	(0.0304)	-
	Plug-in $p = 1/20$	25.60s	0.0117	0.1540	0.9122	0.9170	4.3489	4.2753
			(0.0025)	(0.0271)	(0.0691)	(0.0656)	(0.0293)	-
	Plug-in $p = 1$	512.07s	0.0082	0.1124	0.9194	0.9170	4.3140	4.2753
			(0.0018)	(0.0199)	(0.0644)	(0.0656)	(0.0207)	-
	Fixed-b	2.85s	-	-	0.9014	0.9170	4.5582	4.2753
			-	-	(0.0674)	(0.0656)	(0.5681)	-

Table 4: Computation time, estimation errors, averaged coverage rates, and average lengths of the proposed algorithm with search direction `(I)` and two function queries ($m = 1$). Sample size $n = 10^5$, dimension $d = 20$ under the linear regression model. Corresponding standard errors are reported in the brackets. We compare the plug-in covariance estimator (plug-in) based inference (19) using $p = 1, 1/20, 1/400$ and fixed-b HAR (fixed-b) based inference (23).

5.2 Comparison of the inference procedures

In this subsection, we provide detailed comparisons of different inference procedures. Specifically, we fix dimension $d = 20$, and compare the performance for the plug-in and fixed-b HAR schemes.

Σ Estimator	Comp. time	Estimation error		Average coverage rate		Average length	
		Hessian	Cov.	Estimator	Oracle	Estimator	Oracle
Identity							
Plug-in $p = 1/400$	5.70s	0.1812	0.3293	0.9039	0.9403	5.0715	4.8374
		(0.0281)	(0.0792)	(0.0501)	(0.0540)	(0.2071)	-
Plug-in $p = 1/20$	32.32s	0.0737	0.1636	0.9330	0.9403	4.9599	4.8374
		(0.0114)	(0.0393)	(0.0593)	(0.0540)	(0.1833)	-
Plug-in $p = 1$	643.86s	0.0597	0.1030	0.9418	0.9403	4.8751	4.8374
		(0.0093)	(0.0250)	(0.0532)	(0.0540)	(0.1973)	-
Fixed-b	3.13s	-	-	0.8956	0.9403	5.1763	4.8374
		-	-	(0.1156)	(0.0540)	(0.4362)	-
Equicorr							
Plug-in $p = 1/400$	5.75s	0.0993	0.3620	0.8880	0.9369	5.9456	5.6356
		(0.0287)	(0.0992)	(0.0540)	(0.0561)	(0.1716)	-
Plug-in $p = 1/20$	32.53s	0.0356	0.1441	0.9288	0.9369	5.7766	5.6356
		(0.0120)	(0.0440)	(0.0599)	(0.0561)	(0.1556)	-
Plug-in $p = 1$	645.56s	0.0240	0.1213	0.9383	0.9369	5.6873	5.6356
		(0.0101)	(0.0359)	(0.0577)	(0.0561)	(0.1715)	-
Fixed-b	3.17s	-	-	0.8949	0.9369	5.7532	5.6356
		-	-	(0.1106)	(0.0561)	(0.4064)	-

Table 5: Computation time, estimation errors, averaged coverage rates, and average lengths of the proposed algorithm with search direction `(I)` and two function queries ($m = 1$). Sample size $n = 10^5$, dimension $d = 20$ under the logistic regression model. Corresponding standard errors are reported in the brackets. We compare the plug-in covariance estimator (plug-in) based inference (19) using $p = 1, 1/20, 1/400$ and fixed-b HAR (fixed-b) based inference (23).

For plug-in estimators, at each iteration, we update the Hessian estimator $\widehat{H}_n$ in (19) using (16) with probability p chosen from 1, d^{-1}, d^{-2}. The fixed-b scheme is updated by (23). We report the computation time, the estimation error of the Hessian matrix, and the average coverage rate and length of these candidates based on 100 replications. The computation time is recorded in a simulation environment running Python 3.8 with a single 10-core Apple M1 Max chip.

The simulation for linear regression and logistic regression is given below in Tables 4–5, respectively. As can be referred from the two tables, the fixed-b HAR approach gives the fastest execution, due to the fact that no additional function queries are required for Hessian matrix computation. The fixed-b HAR method is even faster than the case of $p = d^{-2}$ where we update only one entry (in expectation) for the Hessian matrix in each `(KW)` step. Among plug-in cases, the performance of inference improves as p increases, and it achieves a relatively more reliable coverage for $p \geq d^{-1}$, (i.e., at least d entries (in expectation) receive updates for the Hessian estimator per `(KW)` step), with a significant cost of the computation time. In practice, we would recommend the fixed-b HAR method for those computation-sensitive tasks, and the plug-in method with Hessian sampling probability $p \geq d^{-1}$ in less computation-sensitive tasks.

5.3 Choices of the search direction distribution

In this subsection, we compare the results for different directions $\mathcal{P}_{\boldsymbol{v}}$. We report the results for the logistic regression model with the identity design matrix $\Sigma = I$ in Table 6. Table 6 suggests the `(AKW)` algorithms with search directions `(I)`, `(S)`, `(G)` achieve similar performance for parameter estimation error and average coverage rates, while the average confidence intervals of `(G)` are generally larger. The observations in the numerical experiments match our Proposition 3.5. Additional simulation results of the linear regression model and the equicorrelation design are relegated to Tables C.2, C.3, C.4 in the supplementary material.

5.4 Multi-query `(AKW)` estimator

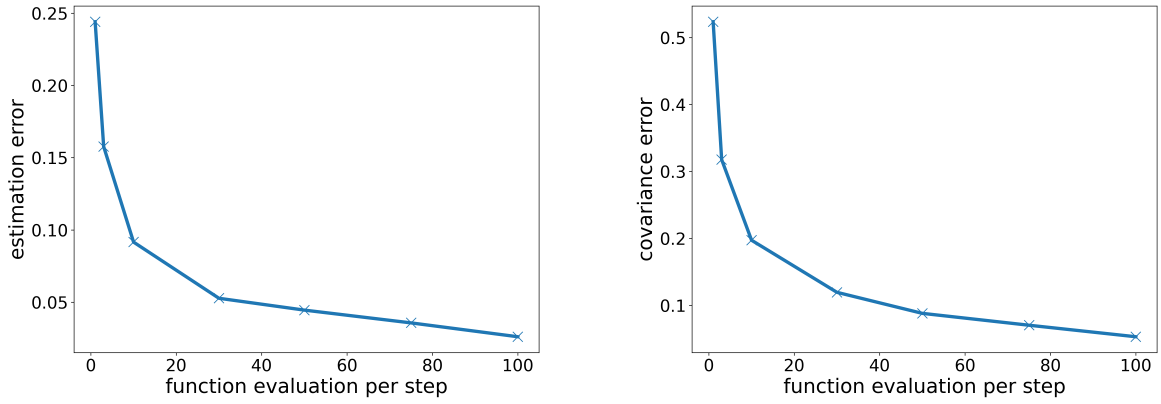
We further conduct experiments for the `(KW)` algorithm with multiple function-value queries ($m > 1$) and compare the performance of $m = 10, 100$ using different search directions with sampling

d	$\mathcal{P}_{\boldsymbol{v}}$	Estimation error		Average coverage rate		Average length	
		Parameter	Plug-in Cov.	Plug-in	Oracle	Plug-in	Oracle
5	(I)	0.0265	0.0587	0.9432	0.9440	3.1136	3.1078
		(0.0115)	(0.0434)	(0.1219)	(0.1148)	(0.8648)	-
	(S)	0.0264	0.0599	0.9396	0.9376	3.0639	3.0625
		(0.0124)	(0.0453)	(0.1276)	(0.1250)	(0.8211)	-
	(G)	0.0312	0.0718	0.9412	0.9420	3.6304	3.6237
		(0.0139)	(0.0498)	(0.1193)	(0.1176)	(0.9770)	-
20	(I)	0.0728	0.1030	0.9418	0.9403	4.8751	4.8374
		(0.0124)	(0.0250)	(0.0532)	(0.0540)	(0.6441)	-
	(S)	0.0711	0.1017	0.9438	0.9419	4.8414	4.8156
		(0.0116)	(0.0246)	(0.0523)	(0.0524)	(0.6322)	-
	(G)	0.0749	0.1054	0.9427	0.9423	5.0873	5.0507
		(0.0121)	(0.0248)	(0.0563)	(0.0523)	(0.6654)	-
100	(I)	0.2440	0.5236	0.9673	0.9082	12.0661	10.3175
		(0.0211)	(0.1646)	(0.0193)	(0.0295)	(1.0106)	-
	(S)	0.2353	0.5122	0.9605	0.9145	13.1366	11.1788
		(0.0205)	(0.1530)	(0.0201)	(0.0358)	(1.0891)	-
	(G)	0.2357	0.5147	0.9614	0.9161	13.2836	11.2901
		(0.0202)	(0.1531)	(0.0205)	(0.0380)	(1.0929)	-

Table 6: Comparison among different direction distributions $\mathcal{P}_{\boldsymbol{v}}$ (Detailed specification of `(I),(S),(G)` can be referred to Section 3.1). We consider the logistic regression model with design matrix $\Sigma = I$, and the `(AKW)` estimators are computed based on the case of two function queries ($m = 1$). Corresponding standard errors are reported in the brackets.

$m;\Sigma$	$\mathcal{P}_{\boldsymbol{v}}$	Estimation error		Average coverage rate		Average length	
		Parameter	Plug-in Cov.	Plug-in	Oracle	Plug-in	Oracle
10; Identity	(I+WOR)	0.0916	0.1972	0.9547	0.9342	3.7013	3.4794
		(0.0103)	(0.1053)	(0.0225)	(0.0330)	(0.2970)	-
	(I+WR)	0.0947	0.2004	0.9551	0.9353	3.8800	3.6383
		(0.0106)	(0.1025)	(0.0215)	(0.0310)	(0.3053)	-
	(S)	0.0958	0.2134	0.9552	0.9320	3.8893	3.6352
		(0.0118)	(0.1172)	(0.0219)	(0.0368)	(0.3054)	-
10; Equicorr	(I+WOR)	0.1184	0.2581	0.9404	0.9126	3.6432	3.3700
		(0.0122)	(0.1278)	(0.0252)	(0.0382)	(0.2240)	-
	(I+WR)	0.1235	0.2828	0.9431	0.9125	3.8352	3.5234
		(0.0145)	(0.1573)	(0.0266)	(0.0437)	(0.2498)	-
	(S)	0.1224	0.2753	0.9435	0.9135	3.8225	3.5165
		(0.0144)	(0.1501)	(0.0259)	(0.0422)	(0.2614)	-
100; Identity	(I+WOR)	0.0261	0.0531	0.9455	0.9438	0.8978	0.8938
		(0.0022)	(0.0135)	(0.0297)	(0.0305)	(0.0290)	-
	(I+WR)	0.0333	0.0568	0.9455	0.9441	1.4037	1.3948
		(0.0030)	(0.0196)	(0.0253)	(0.0262)	(0.0803)	-
	(S)	0.0334	0.0556	0.9458	0.9439	1.4034	1.3941
		(0.0028)	(0.0199)	(0.0231)	(0.0247)	(0.0816)	-
100; Equicorr	(I+WOR)	0.0328	0.0664	0.9490	0.9441	0.9056	0.8971
		(0.0035)	(0.0199)	(0.0339)	(0.0356)	(0.0493)	-
	(I+WR)	0.0453	0.0823	0.9494	0.9444	1.4183	1.3946
		(0.0046)	(0.0322)	(0.0240)	(0.0270)	(0.0766)	-
	(S)	0.0451	0.0821	0.9497	0.9449	1.4157	1.3930
		(0.0048)	(0.0321)	(0.0249)	(0.0270)	(0.0777)	-

Table 7: Comparison among different sampling schemes for multi-query algorithms under logistic regression model with dimension $d = 100$ and $m = 10, 100$, respectively (Detailed specification of `(I+WOR),(I+WR),(S)` can be referred to Section 3.1). Corresponding standard errors are reported in the brackets.



(a) (b)

Figure 8: The parameter estimation error and the relative covariance estimation error (see (24)) for multiple function-value evaluations. The x-axis is the number of function evaluations per step (i.e., $m+1$). Here, we consider the logistic regression model with $n = 10^5$ and $d = 100$.

schemes `(I+WR)`, `(I+WOR)`, and `(S)`. We note that `(I+WR)` and `(I+WOR)` refer to the uniform sampling from natural basis with and without replacement, respectively; and `(S)` refers to the uniform sampling from the sphere. We report the results of the logistic regression model in Table 7 and relegate the results of the linear regression to Table C.5 in the supplementary material.

When $m = 10$, the `(KW)` algorithm using all three sampling schemes achieves similar performance in both estimation and inference. When $m = 100$, the algorithm with `(I+WOR)` achieves better performance than the other two sampling schemes by constructing around 30% shorter confidence intervals on average while achieving comparable coverage rates.

We further present in Figure 8 the estimation error of the parameters and covariance matrices when we increase the function-query complexity m. The numerical results matches the magnitudes of Q with regard to different m in Theorems 3.6–3.7, which could help practitioners choose an appropriate m to balance the accuracy and computational cost. We report the logistic regression results with the identity design matrix $\Sigma = I$ in Figure 8 and relegate the equicorrelation design and the linear regression results to Figures C.6–C.7 in the supplementary material.

6 Conclusion and Future Work

In this paper, we investigate the statistical inference problem for the Kiefer-Wolfowitz stochastic optimization algorithm with random search directions. We show the asymptotic normality for the `(KW)`-type estimators and provide consistent estimators of the asymptotic covariance matrix to facilitate the inference. Our theoretical analysis provides a comprehensive comparison on the impact of different random search directions, the number of multi-query evaluations, and sampling schemes. Our findings are validated by numerical experiments.

For future works, our results and estimation methods may be potentially useful to understand asymptotic behaviors of other gradient-free variants of stochastic optimization algorithms, e.g. moment-adjusted stochastic gradients (Liang and Su, 2019), stochastic optimization under constraints (Duchi and Ruan, 2021), high dimensional stochastic algorithms (Chao and Cheng, 2019; Shi et al., 2021), and SGD in contextual bandit settings Chen et al. (2021).

References

Abadir, Karim M and Paolo Paruolo (1997). Two mixed normal densities from cointegration analysis. *Econometrica 65*(3), 671–680.

Agarwal, Alekh, Ofer Dekel, and Lin Xiao (2010). Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Conference on Learning Theory*, pp. 28–40.

Agarwal, Alekh, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Alexander Rakhlin (2011). Stochastic convex optimization with bandit feedback. In *Advances in Neural Information Processing Systems*, pp. 1035–1043.

Blum, Julius R (1954). Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics 25*(4), 737–744.

Broadie, Mark, Deniz Cicek, and Assaf Zeevi (2011). General bounds and finite-time improvement for the kiefer-wolfowitz stochastic approximation algorithm. *Operations Research 59*(5), 1211–1224.

Chao, Shih-Kang and Guang Cheng (2019). A generalization of regularized dual averaging and its dynamics. *arXiv preprint arXiv:1909.10072*.

Chen, Hung (1988). Lower rate of convergence for locating a maximum of a function. *The Annals of Statistics*, 1330–1334.

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

Chen, Han-Fu, Tyrone E Duncan, and Bozenna Pasik-Duncan (1999). A kiefer-wolfowitz algorithm with randomized differences. *IEEE Transactions on Automatic Control 44*(3), 442–453.

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.

Conn, Andrew R, Katya Scheinberg, and Luis N Vicente (2009). *Introduction to derivative-free optimization*. Society for Industrial and Applied Mathematics.

Dippon, Jürgen (2003). Accelerated randomized stochastic optimization. *The Annals of Statistics 31*(4), 1260–1281.

Duchi, John C, Michael I Jordan, Martin J Wainwright, and Andre Wibisono (2015). Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory 61*(5), 2788–2806.

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

Fabian, Vaclav (1967). Stochastic approximation of minima with improved asymptotic speed. *The Annals of Mathematical Statistics*, 191–200.

Fabian, Vaclav (1968). On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics 39*(4), 1327–1332.

Fabian, Väclav (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer.

Fang, Yixin, Jinfeng Xu, and Lei Yang (2017). On scalable inference with stochastic gradient descent. *arXiv preprint arXiv:1707.00192*.

Flaxman, Abraham D, Adam Tauman Kalai, and H Brendan McMahan (2005). Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 385–394.

Ghadimi, Saeed and Guanghui Lan (2013). Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization 23*(4), 2341–2368.

Hall, Peter and Christopher C Heyde (1980). *Martingale limit theory and its application*. Academic press.

Hall, Peter and Ilya Molchanov (2003). Sequential methods for design-adaptive estimation of discontinuities in regression curves and surfaces. *The Annals of Statistics 31*(3), 921–941.

He, Ying, Michael C Fu, and Steven I Marcus (2003). Convergence of simultaneous perturbation stochastic approximation for nondifferentiable optimization. *IEEE Transactions on Automatic Control 48*(8), 1459–1463.

Jamieson, Kevin G, Robert Nowak, and Ben Recht (2012). Query complexity of derivative-free optimization. In *Advances in Neural Information Processing Systems*, pp. 2672–2680.

Joshi, Siddharth and Stephen Boyd (2008). Sensor selection via convex optimization. *IEEE Transactions on Signal Processing 57*(2), 451–462.

Kiefer, Jack and Jacob Wolfowitz (1952). Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics 23*(3), 462–466.

Kiefer, Nicholas M, Timothy J Vogelsang, and Helle Bunzel (2000). Simple robust testing of regression hypotheses. *Econometrica 68*(3), 695–714.

Koenker, Roger (2005). *Quantile Regression*. Cambridge University Press.

Koronacki, Jacek (1975). Random-seeking methods for the stochastic unconstrained optimization. *International Journal of Control 21*(3), 517–527.

Lee, Sokbae, Yuan Liao, Myung Hwan Seo, and Youngki Shin (2022). 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.

Liang, Tengyuan and Weijie Su (2019). Statistical inference for the population landscape via moment-adjusted stochastic gradients. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 81*(2), 431–456.

Mokkadem, Abdelkader and Mariane Pelletier (2007). A companion for the kiefer–wolfowitz–blum stochastic approximation algorithm. *The Annals of Statistics 35*(4), 1749–1772.

Moulines, Eric and Francis Bach (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *24*, 451–459.

Nesterov, Yurii (2005). Lexicographic differentiation of nonsmooth functions. *Mathematical programming 104*(2), 669–700.

Nesterov, Yurii and Vladimir Spokoiny (2017). Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics 17*(2), 527–566.

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

Polyak, B. T. and A. B. Tsybakov (1990). Optimal order of accuracy of search algorithms in stochastic optimization. *Problemy Peredachi Informatsii*, 126–133.

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

Ruppert, David (1982). Almost sure approximations to the robbins-monro and kiefer-wolfowitz processes with dependent noise. *The Annals of Probability 10*(1), 178–187.

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

Shamir, Ohad (2017). An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research 18*(1), 1703–1713.

Shi, Chengchun, Rui Song, Wenbin Lu, and Runze Li (2021). Statistical inference for high-dimensional models via recursive online-score estimation. *Journal of the American Statistical Association 116*(535), 1307–1318.

Spall, James C (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control 37*(3), 332–341.

Spall, James C (2000). Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Transactions on Automatic Control 45*(10), 1839–1853.

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*.

Toulis, Panos and Edoardo M Airoldi (2017). Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics 45*(4), 1694–1727.

van der Vaart, Aad W (2000). *Asymptotic statistics*, Volume 3. Cambridge University Press, Cambridge.

Wainwright, Martin J and Michael Irwin Jordan (2008). *Graphical models, exponential families, and variational inference*. Now Publishers Inc.

Wang, Yining, Simon Du, Sivaraman Balakrishnan, and Aarti Singh (2018). Stochastic zeroth-order optimization in high dimensions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 1356–1365. PMLR.

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