Hessian Inverse Approximation as Covariance for Random Perturbation in Black-Box Problems

Jingyi Zhu
JINGYI.ZHU@ALIBABA-INC.COM
DAMO Academy, Alibaba Group, USA

Abstract

In stochastic optimization problems using noisy zeroth-order (ZO) oracles only, the randomized counterpart of Kiefer-Wolfowitz-type method is widely used to estimate the gradient. Existing algorithms generate the randomized perturbation from a zero-mean and unit-covariance distribution. In contrast, this work considers the generalization where the perturbations have a potentially non-identity covariance constructed from the history of the ZO queries. We propose to feed the second-order approximation into the covariance matrix of the random perturbation, so it is dubbed as Hessian-aided random perturbation (HARP). HARP collects four zeroth-order queries per iteration to form approximations for both the gradient and the Hessian. We show the convergence (in an almost surely sense) and derive the convergence rate for HARP under standard assumptions. We demonstrate, with theoretical guarantees and numerical experiments, that HARP is less sensitive to ill-conditioning and more query-efficient than other gradient approximation schemes using unit-covariance random perturbation.

Keywords:
stochastic optimization, simultaneous perturbation, gradient-free methods, Hessian approximation

1. Introduction

Stochastic approximation (SA) is a class of recursive procedures to locate roots of equations in the presence of noisy measurements. When only noisy zeroth-order (ZO) information is available, it is common practice to generate deterministic perturbation [4, 16] or random perturbation [10, 15, 29] in finding extrema. SA methods using ZO information have regained their popularity in evolutionary strategy (as an alternative to reinforcement learning) [22, 28] and adversarial image attack [5, 17]. To the best of our knowledge, all the existing random-perturbation-based methods generate the perturbation from a distribution with zero-mean and unit-covariance, which enforce that every component of the perturbation vector is independent with all other components. The resulting gradient estimate may not be robust to scaling and correlation of different parameters. Therefore, this paper establishes the theoretical guarantee for the SA procedure using random perturbation with non-identity covariance. Specifically, we feed the Hessian inverse approximation into the perturbation covariance, so the newly-proposed method is dubbed as Hessian-aided random perturbation (HARP). HARP exhibits faster and more stable convergence performance other SA algorithms in ill-conditioned problems, for which we provide both the theoretical analysis and the numerical illustration (via universal image attack).

We now describe the problem setting. Let $\theta \in \mathbb{R}^d$ concatenate all the adjustable model parameters. Let the random variable $\omega \in \Omega$ represent the (generally uncontrollable) stochasticity of the underlying system. Consider

$$\min_{\theta \in \mathbb{R}^d} L(\theta) \equiv \mathbb{E}_{\omega \sim \mathbb{P}}[\ell(\theta, \omega)], \tag{1}$$

where the loss function $L(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$ measures the underlying system performance, and the random variable $\ell(\cdot, \cdot) : \mathbb{R}^d \times \Omega \mapsto \mathbb{R}$ evaluated at (θ, ω) represents a noisy observation of $L(\theta)$ corrupted by ω. Under the context that *only* noisy zeroth-order (ZO) information $\ell(\cdot, \omega)$ for some $\omega \sim \mathbb{P}$ is available at certain values of θ and that these noisy ZO queries may be *expensive*, this work considers the generic stochastic approximation (SA) algorithm:

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{\boldsymbol{g}}_k(\hat{\theta}_k), \ k \geq 1, \tag{2}$$

where $\hat{\theta}_k$ denotes the recursive estimate at the kth iteration, $\hat{\boldsymbol{g}}_k(\hat{\theta}_k)$ represents the estimate for the gradient $\boldsymbol{g}(\hat{\theta}_k)$, and $a_k > 0$ is the stepsize. This work focuses on the following gradient estimation scheme using two ZO queries per iteration:

$$\hat{\boldsymbol{g}}_k(\hat{\theta}_k) = \frac{\ell(\hat{\theta}_k + c_k \boldsymbol{\Delta}_k, \omega_k^+) - \ell(\hat{\theta}_k - c_k \boldsymbol{\Delta}_k, \omega_k^-)}{2c_k} \boldsymbol{m}_k(\boldsymbol{\Delta}_k), \tag{3}$$

where c_k represents the differencing magnitude, the d-dimensional random perturbation vectors $\boldsymbol{\Delta}_k$ is assumed to be drawn from a distribution with $\mathbf{0}$-mean and $\boldsymbol{\Sigma}_k^{-1}$-covariance, and the mapping $\boldsymbol{m}_k(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}^d$ is odd. The details will be discussed momentarily.

As for the statistical structure between ω_k^+ and ω_k^-, two classical scenarios are considered. The first one where ω_k^+ and ω_k^- are independent and identically distributed will be termed as IID. The antithesis of IID, where $\omega_k^+ = \omega_k^-$, will be referred to as "common random number" (CRN). The CRN scenario are useful *simulation-based* optimization.

1.1. Prior Work and Our Contribution

The generic form for gradient estimate in (3) subsumes random direction stochastic approximation (RDSA) [10, 11] with $\boldsymbol{\Delta}_k$ being uniformly distributed on the unit spherical surface and $\boldsymbol{m}_k(\boldsymbol{\Delta}_k) = d\boldsymbol{\Delta}_k$, smoothed functional stochastic approximation (SFSA) [15] with $\boldsymbol{\Delta}_k$ being standard multivariate normally distributed and $\boldsymbol{m}_k(\boldsymbol{\Delta}_k) = \boldsymbol{\Delta}_k$, simultaneous perturbation stochastic approximation (SPSA) [29] with each component of $\boldsymbol{\Delta}_k$ being Rademacher distributed and $\boldsymbol{m}_k(\boldsymbol{\Delta}_k) = \boldsymbol{\Delta}_k$. Although the randomized scheme (3) exists for a long time and demonstrates numerical advantages over FDSA [16], theoretical foundation regarding the *optimal* choices of $\boldsymbol{\Delta}_k$ is lacking and extra caution is required in its implementation.

We propose a new algorithm called "Hessian-aided random perturbation" (HARP). The choice of feeding Hessian approximation into $\boldsymbol{\Sigma}_k$ is motivated by overcoming the shortcomings of $\boldsymbol{\Sigma}_k = \boldsymbol{I}$ in Section 2.1, analyzed theoretically through almost surely convergence and convergence rate in Section 4.2, and demonstrated through two numerical experiments in Section 5. Previously, in both stochastic optimization [31] and deterministic optimization, the Hessian is applied in parameter update *only*. HARP adaptively changes the covariance $\boldsymbol{\Sigma}_k^{-1}$ of the perturbation $\boldsymbol{\Delta}_k$ using Hessian approximation, so that one can conveniently handle the issues pertaining to the scaling and correlation of different parameters, see Section 2.1. Compared with prior algorithms using unit-covariance random perturbation, HARP exhibits faster and more stable convergence performance, especially in ill-conditioned problems.

What sets our work different from the prior work [14, 24] is discussed in Subsection 3.3. In short, [24] is applicable for additive[1] CRN noise, and the corresponding analysis cannot be generalized to the general CRN noise discussed in Section 4.2, not to mention the IID noise discussed in Section 4.1. Additionally, [32] also considers leveraging the Hessian estimates to achieve faster convergence. The results therein have to be interpreted with extra caution: the random perturbation $\boldsymbol{\Delta}_k$ impacts both the gradient and the Hessian estimates at each iteration, yet the proofs therein ignore the randomness in the Hessian estimate.

1.2. Notation Convention

Matrix and vector operations Let $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ be a matrix and let $\boldsymbol{x} \in \mathbb{R}^d$ be a vector. $\|\boldsymbol{x}\|$ returns the Euclidean norm of $\boldsymbol{x}$, and $\|\boldsymbol{A}\|$ returns the spectral norm of $\boldsymbol{A}$. If $\boldsymbol{A}$ is real-symmetric, $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ return the smallest and the largest eigenvalues of $\boldsymbol{A}$. The binary operator $\otimes$ represents the Kronecker product.

Probability and SA conventions For a sequence of random variables X_k, $X_k = o_P(1)$ means that X_k converges to zero in probability as $k \to \infty$, and $X_k = O_P(1)$ means that X_k is stochastically bounded.

Let $\mathcal{F}_k$ represent the history of the recursion (2) until the kth iteration, and let $\mathbb{E}_k(\cdot)$ denote the conditional expectation $\mathbb{E}[\cdot|\,\mathcal{F}_k]$. To facilitate later discussion, we rewrite $\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)$ as:

$$\begin{aligned}\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k) &= \boldsymbol{g}(\hat{\boldsymbol{\theta}}_k) + \mathbb{E}_k[\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k) - \boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)] + \{\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k) - \mathbb{E}_k[\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)]\} \\ &\equiv \boldsymbol{g}(\hat{\boldsymbol{\theta}}_k) + \boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k) + \boldsymbol{\xi}_k(\hat{\boldsymbol{\theta}}_k)\,, \end{aligned} \tag{4}$$

where $\boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k)$ represents the bias of $\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)$ as an estimator of $\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)$, and $\boldsymbol{\xi}_k(\hat{\boldsymbol{\theta}}_k)$ represents the noise term. The decomposition (4) is useful in asymptotic normality of $\hat{\boldsymbol{\theta}}_k$ later, where $\boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k)$ affects the asymptotic mean and $\boldsymbol{\xi}_k(\hat{\boldsymbol{\theta}}_k)$ affects the asymptotic variance.

Miscellaneous notation $\mathbb{I}_E$ represents the indicator function of a logical expression E. For $L(\cdot)$ that is three-times continuously differentiable, let $L^{(3)}(\boldsymbol{\theta}) \in \mathbb{R}^{1\times d^3}$ represent the third-order derivative of $L(\cdot)$ evaluated at $\boldsymbol{\theta}$; moreover, let $L^{(3)}_{i_1,i_2,i_3}(\boldsymbol{\theta}) \in \mathbb{R}$ represent the third-order derivative of $L(\cdot)$ with respect to (w.r.t.) the i_1th, i_2th, and i_3th elements of $\boldsymbol{\theta}$.

2. Motivation and Description of HARP

This section motivates HARP and lists the pseudo code.

2.1. Motivation Behind HARP

Prior work summarized Section 1.1 enforce $\boldsymbol{\Sigma}_k = \boldsymbol{I}$. We take SPSA as an example.

(i) The estimate $\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)$ using SPSA may not be robust to scaling, as every component of $\hat{\boldsymbol{\theta}}_k$ is perturbed by the *same* magnitude of perturbation c_k.

1. This is a special CRN case, which completely remove the dependency on the observation noise from the entire gradient estimate $\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)$ in (3). Namely, additive CRN noise, the numerator in (3) involves the randomness in $\boldsymbol{\Delta}_k$ only.

(ii) The estimate $\hat{\boldsymbol{g}}_k(\hat{\theta}_k)$ may not be robust to various correlations between different components of the parameter, as the perturbations along all components of θ are *independent* with each other.

As it turns out later in Subsection 4.3, a sensible choice of $\boldsymbol{\Sigma}_k$ is $\boldsymbol{H}(\hat{\theta}_k)$. In this way, (i) can be resolved: say, $L(\theta) = (100\theta_1^2 + \theta_2^2)/2$, then a $\boldsymbol{\Delta}_k$ with zero mean and a covariance of $\mathrm{diag}(0.01, 1)$ will *on average* impose 10% of the change magnitude in θ_2 onto that of θ_1. Meanwhile, (ii) can be handled: say, $L(\theta) = (\theta_1^2 + \theta_2^2 + \theta_1\theta_2)/2$. When the direction of the gradient estimate, $\boldsymbol{m}_k(\boldsymbol{\Delta}_k)$, has a covariance of $\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ will have heavier probability mass on $(1,\, 1)^T$ and $(-1,\, -1)^T$ than on $(-1,\, 1)^T$ and $(1,\, -1)^T$.

2.2. Algorithm Description

Section 2.1 explains the shortcomings (i)–(ii) of using $\boldsymbol{\Sigma}_k = \boldsymbol{I}$ and illustrates the benefit of using $\boldsymbol{\Sigma}_k = \boldsymbol{H}(\hat{\theta}_k)$. Obviously, we cannot access $\boldsymbol{H}(\hat{\theta}_k)$ in the black-box problem (1). We consider constructing estimate for $\boldsymbol{H}(\hat{\theta}_k)$ using ZO queries gathered in $\mathcal{F}_k$. To form a $\mathcal{F}_k$-measurable second-order approximation, HARP is comprised of two recursions, one for θ as in (2) and one for the Hessian $\boldsymbol{H}(\theta)$ as below:

$$\begin{cases} \hat{\boldsymbol{H}}_k = \boldsymbol{f}_k\left(\overline{\boldsymbol{H}}_k\right) , \\ \overline{\boldsymbol{H}}_{k+1} = (1-w_k)\overline{\boldsymbol{H}}_k + w_k \left\{ \boldsymbol{m}_k(\widetilde{\boldsymbol{\Delta}}_k)[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)]^T + \boldsymbol{m}_k(\boldsymbol{\Delta}_k)[\boldsymbol{m}_k(\widetilde{\boldsymbol{\Delta}}_k)]^T \right\} \bar{\ell}_k/(4c_k\widetilde{c}_k) \,. \end{cases} \tag{5}$$

Here, c_k and $\widetilde{c}_k$ are the differencing magnitudes, the d-dimensional random perturbation vectors $\boldsymbol{\Delta}_k$ and $\widetilde{\boldsymbol{\Delta}}_k$ are assumed to be drawn from a distribution with $\boldsymbol{0}$-mean and $\boldsymbol{\Sigma}_k^{-1}$-covariance, the mapping $\boldsymbol{m}_k(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}^d$ is odd, and $\bar{\ell}_k = \ell(\hat{\theta}_k + c_k\boldsymbol{\Delta}_k + \widetilde{c}_k\widetilde{\boldsymbol{\Delta}}_k, \omega_k^{+,+}) - \ell(\hat{\theta}_k + c_k\boldsymbol{\Delta}_k, \omega_k^{+}) - \ell(\hat{\theta}_k - c_k\boldsymbol{\Delta}_k + \widetilde{c}_k\widetilde{\boldsymbol{\Delta}}_k, \omega_k^{-,+}) + \ell(\hat{\theta}_k - c_k\boldsymbol{\Delta}_k, \omega_k^{+})$. The current estimate $\hat{\boldsymbol{H}}_k$ and the smoothing (moving average) estimate $\overline{\boldsymbol{H}}_k$ can be initialized as the identity/scalar matrix. The mapping $\boldsymbol{f}_k : \mathbb{R}^{d\times d} \mapsto \{\text{positive definite matrices in } \mathbb{R}^{d\times d}\}$ copes the potential nonpositive-definiteness of $\overline{\boldsymbol{H}}_k$. A valid choice for $\boldsymbol{f}_k(\cdot)$ is $\boldsymbol{f}_k(\boldsymbol{H}) = (\boldsymbol{H}^T\boldsymbol{H} + \varepsilon_k\boldsymbol{I})^{1/2}$ with $\varepsilon_k \to 0$, which can be implemented in $O(d^2)$ FLOPs [37]. When $\widetilde{c}_k = O(c_k)$, and other gain sequence conditions are met, $\hat{\boldsymbol{H}}_k$ approaches the Hessian evaluated at the optimum at a rate no slower than $O(c_k^2)$.

The detailed pseudo code for HARP is summarized in Algorithm 1. Readers are referred to our code hyperlinks in Section 5 and [37] for implementation guidance.

3. Performance Metric

Before analyzing HARP listed in Algorithm 1, let us discuss the metric that evaluates the algorithm performance.

3.1. Convergence Mode

Now that all randomness in $\hat{\theta}_k$ stemming from $\Omega \times \Omega_{\boldsymbol{\Delta}}$, it is standard practice to measure the algorithmic performance of the recursions (2) by showing

$$\hat{\theta}_k \text{ converges almost surely (strongly) to } \theta^* \,,\, (\hat{\theta}_k \xrightarrow{\text{a.s.}} \theta^*)\,, \tag{6}$$

Algorithm 1: Hessian-Amended Random Perturbation (GitHub)

Result: terminal estimate $\hat{\theta}_K$
initialization $\hat{\theta}_0$, $\hat{\boldsymbol{H}}_0 = \boldsymbol{I}$, $\overline{\boldsymbol{H}}_0 = \boldsymbol{I}$, and coefficients $a_k, c_k, \widetilde{c}_k, w_k$ for $0 \le k \le K$;
for $k = 0, 1, \cdots K$ **do**
  generate $\boldsymbol{\Delta}_k$ from a distribution with a mean of $\boldsymbol{0}$ and a covariance of $\hat{\boldsymbol{H}}_k^{-1}$ and compute $\boldsymbol{m}_k(\boldsymbol{\Delta}_k) = \hat{\boldsymbol{H}}_k \boldsymbol{\Delta}_k$;
  collect two ZO queries and estimate $\hat{\boldsymbol{g}}_k(\hat{\theta}_k)$ via (3);
  update $\hat{\theta}_k$ using (2);
  generate $\widetilde{\boldsymbol{\Delta}}_k$ from a distribution with a mean of $\boldsymbol{0}$ and a covariance of $\hat{\boldsymbol{H}}_k^{-1}$ and compute $\boldsymbol{m}_k(\widetilde{\boldsymbol{\Delta}}_k) = \hat{\boldsymbol{H}}_k \widetilde{\boldsymbol{\Delta}}_k$;
  collect two additional ZO queries and estimate $\hat{\boldsymbol{H}}_k$. $\triangleright$ [37, Algorithms 1–2] provides a way to achieve $O(d^2)$ FLOPs. Other forms of $\boldsymbol{f}_k(\cdot)$ satisfying conditions in [31] also work.
end

or

$$\hat{\theta}_k \text{ converges to } \theta^* \text{ in mean-squared sense}\,, \; (\hat{\theta}_k \xrightarrow{\text{m.s.}} \theta^*)\,. \tag{7}$$

Robbins and Monro [25] gave conditions for (6) whereas Blum [4] for (7)[2]. We will prove (6) in Section 4.

3.2. Rate of Convergence

When either (6) or (7) is shown, finding the rate of convergence naturally follows. The asymptotic root-mean-squared (RMS) error $[\mathbb{E}(\|\hat{\theta}_k - \theta^*\|)^2]^{1/2}$ of the underlying estimate $\hat{\theta}_k$ is a sensible measure of the distance between the $\hat{\theta}_k$ and θ^* average across all sample paths. Therefore, we aim to find the smallest upper bound τ^* such that $k^{\tau_0/2}(\hat{\theta}_k - \theta^*) = O_P(1)$ for all $\tau_0 \le \tau^*$, which is formalized as:

$$\begin{cases} \max_{\mathcal{S}} \; \tau\,, \\ \text{s.t. random sequence } (\hat{\theta}_k - \theta^*) \text{ is } O_P(k^{-\tau/2})\,, \end{cases} \tag{8}$$

where the hyperparameter set $\mathcal{S}$ includes all the controllable stepsizes, and both τ and $O_P(1)$ are functions of $\mathcal{S}$. Thanks to the algorithmic form (2), the decomposition (4), and [2, Sect. 27], the constraint in (8) always takes the following form:

$$k^{\tau/2}(\hat{\theta}_k - \theta^*) \xrightarrow{\text{dist.}} \mathcal{N}(\mu, \boldsymbol{B}) \text{ for finite } \mu, \boldsymbol{B} \succ \boldsymbol{0}\,, \tag{9}$$

where $\xrightarrow{\text{dist.}}$ represents "convergence in distribution," and $(\tau, \mu, \boldsymbol{B})$ are functions of $\mathcal{S}$. When (9) holds and $[k^{\tau/2}(\hat{\theta}_k - \theta^*)]$ is uniformly integrable for any $\tau \le \tau^*$, (7) holds. The RMS error is asymptotic to $\lim_{k\to\infty}[\mathbb{E}(\|\hat{\theta}_k - \theta^*\|^2)]^{1/2} = k^{-\tau/2}[\|\mu\|^2 + \text{tr}(\boldsymbol{B})]$.

2. Neither (6) nor (7) implies the other [3, Chap. 5]. Both (6) and (7) imply convergence in probability and convergence in distribution.

3.2.1. FURTHER REMARKS ON RMS

To minimize the RMS, it makes more sense to perform

$$\min_{\mathcal{S}} \left\{ k^{-\tau/2} \left[\|\mu\|^2 + \operatorname{tr}(\boldsymbol{B}) \right] \right\} , \tag{10}$$

as opposed to (8). When k is small, the finite constant $[\|\mu\|^2 + \operatorname{tr}(\boldsymbol{B})]$ that are hidden from the big-O notation $O(k^{-\tau/2})$ can be dominating. For sufficiently large k, the effect of the scaling coefficients dies down, and (10) reduces to (8). Sections 4.1–4.2 show that the solution to (8) is

$$\tau^* = \begin{cases} 2/3\,, & \text{for IID noise}\,, \quad (11) \\ 1\,, & \text{for CRN noise}\,, \quad (12) \end{cases}$$

when $L(\cdot)$ is is non-quadratic[3] and three-times[4] continuously differentiable.

3.2.2. ITERATION AND QUERY COMPLEXITY

The complexity analysis for (2) is straightforward when the RMS metric (10) is in use. To achieve

$$\epsilon\text{-accurate estimate } \hat{\theta}_k \text{ s.t. } [\mathbb{E}(\|\hat{\theta}_k - \theta^*\|^2)]^{1/2} \leq \epsilon\,, \tag{13}$$

the the *average* desired number of iteration is

$$\left\{ [\|\mu\|^2 + \operatorname{tr}(\boldsymbol{B})]/\epsilon \right\}^{2/\tau^*} = \begin{cases} O(\epsilon^{-3})\,, & \text{IID noise}\,, \\ O(\epsilon^{-2})\,, & \text{CRN noise}\,. \end{cases} \tag{14}$$

Remark 1 *When (2) takes a fixed number, say $2q$, of ZO queries, and q is independent from the parameter dimension d, then the corresponding query complexity is*

$$2q \left\{ [\|\mu\|^2 + \operatorname{tr}(\boldsymbol{B}/q)]/\epsilon \right\}^{\frac{2}{\tau^*}} = \begin{cases} O(\epsilon^{-3}), & \textit{IID noise,} \\ O(\epsilon^{-2}), & \textit{CRN noise.} \end{cases} \tag{15}$$

3.3. Other Forms of "Convergence" Rate

[24, Sect. 4] uses the following notion

$$\epsilon\text{-accurate estimate } \hat{\theta}_k \text{ s.t. } \mathbb{E}[L(\hat{\theta}_k) - L(\theta^*)] \leq \epsilon\,, \tag{16}$$

as opposed to (13), and (16) is popular for analyzing ZO algorithms [14]. Let us offer a few remarks on the differences between (13) and (16). First of all, the resultant "convergence" rate under the notion (16) require *non*-decaying rate. [33, Chap. 4] points out that $\hat{\theta}_k$ will *not* converge to θ^* in standard statistical sense (either a.s. or m.s. in Subsection 3.1) when $a_k \not\to 0$. In fact, there is no "convergence" per se [36], as $\hat{\theta}_k$ will be "random-walking" within a neighborhood of θ^* even for sufficiently large k [35]. Second, [14, 24] and all the subsequent work on ZO algorithms require *additive* CRN noise, and the corresponding analysis can *not* be generalized to the general CRN

3. For a quadratic function $L(\cdot)$, $\tau^* = 1$ for both IID and CRN noise.
4. For a function $L(\cdot)$ that is p-times continuously differentiable for odd p, the fastest rate for the RMS is $O(k^{-(p-1)/2p})$, which goes to $O(k^{-1/2})$ as $p \to \infty$ [12].

noise case discussed in Section 4.2, not to mention the IID noise case in Section 4.1. Third, the complexity result [24, Eq. (59)] does not reveal the eigen-structure of $\boldsymbol{H}(\cdot)$ under certain smoothness assumption. On the contrary, $\boldsymbol{B}$ in (9) conveys all the eigen-information of $\boldsymbol{H}(\theta^*)$, as we shall see momentarily. It makes more sense that the RMS should be larger for ill-conditioned problems compared with well-conditioned problems. Last but not least, $[\mathbb{E}(\|\hat{\theta}_k - \theta^*\|^2)]^{1/2} \le \epsilon$ implies $\mathbb{E}[L(\hat{\theta}_k) - L(\theta^*)] \le \epsilon'$, but generally not the other way around.

Overall, the notion (16) and the analysis in [14, 24] are useful when (i) *additive* CRN noise scenario is possible, and (ii) the experimenter aims to report an acceptable output within the neighborhood of θ^* given a limited iteration/query complexity. In fact, the non-decaying gain does provide better performance under a budget-limited context [34, 36]. Finally, it is advisable to use "concentration" and "concentration rate" [18, Chaps. 7–8].

3.4. Dependency on Dimensionality d

When dimensionality d varies as the recursion goes on, e.g., d plays an important role in structural optimization and etc., it is advisable to include the dependency on d in the constraint of (8) as "random vector sequence $(\hat{\theta}_k - \theta^*)$ is $O_P(dk^{-\tau/2})$." Nevertheless, we decide to omit d for clarity and for the reason that the problem dimension d is *generally* not an adjustable[5]. Moreover, reducing the order of d appearing in the convergence rate is only possible when certain sparsity conditions are imposed or certain sparsity-promoting regularization is added to the loss function.

4. Convergence Result

4.1. IID Scenario

Overall, the *fastest* rate of the RMS error $(\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2])^{1/2}$ under IID noise is $O(k^{-1/3})$, which is attained when $a_k = O(k^{-1})$ and $c_k = O(k^{-1/6})$. (11) is inherently slower than (12), due to the trade-off between the bias magnitude $\mathbb{E}\|\boldsymbol{\beta}_k(\hat{\theta}_k)\|$ and the variance $\mathbb{E}\|\boldsymbol{\xi}_k(\hat{\theta}_k)\|^2$ of the noise, which is summarized in Lemma 1 below.

As pointed out in Subsection 3.2.1, not only the rate itself but also the scaling coefficient play a role in the algorithmic performance. This section first show the a.s. convergence of the estimate $\hat{\theta}_k$ generated from (2) when the covariance of the perturbation sequence may be varied, and then discuss the impact of the perturbation covariance on the finite constant $[\|\boldsymbol{\mu}\|^2 + \mathrm{tr}(\boldsymbol{B})]$.

4.1.1. ORDER OF BIAS AND VARIANCE OF $\hat{\boldsymbol{g}}_k(\hat{\theta}_k)$

Let us first discuss the bias-variance trade-off in $\hat{\boldsymbol{g}}_k(\hat{\theta}_k)$ for IID noise. Several assumptions are imposed on the underlying loss function $L(\cdot)$, the procedure to generate random perturbation $\boldsymbol{\Delta}_k$, especially the $\mathcal{F}_k$-measurable covariance matrix $\boldsymbol{\Sigma}_k$, and the observation noise $\varepsilon_k^{\pm} \equiv \ell(\hat{\theta}_k \pm c_k\boldsymbol{\Delta}_k, \omega_k^{\pm}) - L(\hat{\theta}_k \pm c\boldsymbol{\Delta}_k)$.

Assumption A.1 (Loss Function) *Assume that there exists some K, such that for $k \ge K$, $L^{(3)}(\theta)$ evaluated for all θ in an open neighborhood of $\hat{\theta}_k$ exists continuously and $\|L^{(3)}(\theta)\|_\infty \le D_1$ almost surely (a.s.).*

5. This contrasts with k and ϵ, both of which can be selected by the experimenter.

Assumption A.2 (Perturbation) *Assume that the perturbation sequence $\{\boldsymbol{\Delta}_k\}$ are independently distributed with a mean of $\mathbf{0}$ and a covariance matrix $\boldsymbol{\Sigma}_k^{-1}$. Meanwhile, the mapping $\boldsymbol{m}_k(\cdot)$ is an odd function. Moreover, both $\boldsymbol{\Delta}_k$ and $\boldsymbol{m}_k(\boldsymbol{\Delta}_k)$ are independent of $\hat{\boldsymbol{\theta}}_k$. Finally, assume that $\mathbb{E}_k[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\boldsymbol{\Delta}_k] \overset{\text{a.s.}}{=} \boldsymbol{I}$ and $\mathbb{E}_k[\|\boldsymbol{\Delta}_k\|^6\|\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\|^2] \overset{\text{a.s.}}{\leq} D_2$ uniformly for all k.*

Assumption A.3 (IID) *Assume $\mathbb{E}[\,\varepsilon_k^+ - \varepsilon_k^- \,|\, \hat{\boldsymbol{\theta}}_k, \boldsymbol{\Delta}_k] \overset{\text{a.s.}}{=} 0$, and $\mathbb{E}[\,(\varepsilon_k^+ - \varepsilon_k^-)^2 \,|\, \hat{\boldsymbol{\theta}}_k, \boldsymbol{\Delta}_k] \overset{\text{a.s.}}{\leq} D_3$ uniformly for all k.*

Lemma 1 *When assumptions A.1, A.2, and A.3 hold,*

$$\boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k) \overset{\text{a.s.}}{=} \frac{c_k^2}{12}\mathbb{E}_k\left\{[L^{(3)}(\overline{\boldsymbol{\theta}}_k^+) + L^{(3)}(\overline{\boldsymbol{\theta}}_k^-)](\boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k)\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\right\}, \tag{17}$$

$$\begin{aligned}\boldsymbol{\xi}_k(\hat{\boldsymbol{\theta}}_k) \overset{\text{a.s.}}{=} &\frac{(\varepsilon_k^+ - \varepsilon_k^-)}{2c_k}\boldsymbol{m}_k(\boldsymbol{\Delta}_k) + [\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\boldsymbol{\Delta}_k^T - \boldsymbol{I}]\,\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)\\ &+ \frac{c_k^2}{12}[L^{(3)}(\overline{\boldsymbol{\theta}}_k^+) + L^{(3)}(\overline{\boldsymbol{\theta}}_k^-)](\boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k)\boldsymbol{m}_k(\boldsymbol{\Delta}_k) - \boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k),\end{aligned} \tag{18}$$

where $\overline{\boldsymbol{\theta}}_k^{\pm}$ is some convex combination of $\hat{\boldsymbol{\theta}}_k$ and $(\hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k)$. Overall, the magnitude of the bias term $\mathbb{E}_k\|\boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k)\|$ is $O(c_k^2)$, and the second-moment of the noise term $\mathbb{E}_k[\|\boldsymbol{\xi}_k\|^2]$ is $O(c_k^{-2})$.

The difficulty in tuning c_k stems from the trade-off between the bias term $O(c_k^2)$ and the variance term $O(c_k^{-2})$.

Discussion on A.1 The $O(c_k^2)$ bias and $O(c_k^{-2})$ variance in Lemma 1 remain valid when the "three-times continuously differentiablility" in A.1 is changed to "twice-continuously differentiablility and *Lipschitz* Hessian." Under such condition, we may still obtain $\mathbb{E}_k\|\boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k)\| = O(c_k^2)$ and $\mathbb{E}_k[\|\boldsymbol{\xi}_k(\hat{\boldsymbol{\theta}}_k)\|] = O(c_k^{-2})$.

4.1.2. ALMOST SURELY CONVERGENCE

Several additional assumptions are imposed to facilitate the strong convergence.

Assumption A.4 (Iterate Boundedness and ODE Condition) *Assume $\|\hat{\boldsymbol{\theta}}_k\| \overset{\text{a.s.}}{<} \infty$ for all k. Also assume that $\boldsymbol{\theta}^*$ is an asymptotically stable solution of the differential equation $\mathrm{d}\boldsymbol{x}(t)/\mathrm{d}t = -\boldsymbol{g}(\boldsymbol{x})$, whose solution under initial condition $\boldsymbol{x}_0$ will be denoted as $\boldsymbol{x}(t|\,\boldsymbol{x}_0)$. Moreover, let $D(\boldsymbol{\theta}^*) \equiv \{\boldsymbol{x}_0 : \lim_{t\to\infty}\boldsymbol{x}(t|\,\boldsymbol{x}_0) = \boldsymbol{\theta}^*\}$. Further assume that $\hat{\boldsymbol{\theta}}_k$ falls within some compact subset of $D(\boldsymbol{\theta}^*)$ infinitely often for almost all sample points.*

Assumption A.4' (Unique Minimum) *Assume that $\boldsymbol{\theta}^*$ is the unique minimizer such that $\sup\{\|\boldsymbol{\theta}\| : L(\boldsymbol{\theta}) \leq L(\boldsymbol{\theta}^*) + C_1\} < \infty$ for every $C_1 > 0$, $\inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|>C_2}[L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*)] > 0$ for every $C_2 > 0$, $\inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|>C_3}\|\boldsymbol{g}(\boldsymbol{\theta})\| > 0$ for every $C_3 > 0$. Moreover, there exists some K, such that for $k \geq K$, $\boldsymbol{H}(\cdot)$ satisfies $\|\boldsymbol{H}(\boldsymbol{\theta})\|_\infty < D_4$ for all $\boldsymbol{\theta}$ in an open neighborhood of $\hat{\boldsymbol{\theta}}_k$ a.s.*

Assumption A.5 (Stepsize) $a_k > 0$, $c_k > 0$, $a_k \to 0$, $c_k \to 0$, $\sum_k a_k = \infty$, $\sum_k a_k^2 c_k^{-2} < \infty$.

Theorem 1 (Almost Surely Convergence) *Under the assumptions A.1, A.2, A.3 (as in Lemma 1), along with A.4 and A.5, we have $\hat{\boldsymbol{\theta}}_k \overset{k\to\infty}{\longrightarrow} \boldsymbol{\theta}^*$ a.s.*

Theorem 1' (Almost Surely Convergence) *Under A.1, A.2, A.3, along with A.4' and A.5, we have*

i) $\|\hat{\theta}_k\| \overset{\text{a.s.}}{<} \infty$ *for all* k.

ii) $\hat{\theta}_k \overset{k\to\infty}{\longrightarrow} \theta^*$ *a.s.*

Discussion on A.4 and A.4' First of all, note that neither A.4 nor A.4' implies the other. Moreover, $\boldsymbol{H}(\cdot)$ being strongly convex is a *sufficient* condition for both A.4 and A.4'. Nonetheless, strong convexity is *not* a *necessary* condition for either A.4 and A.4'. Therefore, both Theorem 1 and Theorem 1' imply a.s. convergence when $L(\cdot)$ is strongly convex, but they also imply the a.s. convergence result for functions that are more complicated beyond strongly convex functions. [19, pp. 40–41] discusses why the iterate-boundedness in A.4 *may* not not a restrictive condition and could be expected to hold in most applications.

4.1.3. ASYMPTOTIC NORMALITY

Additional assumptions are needed to facilitate the weak convergence result.

Assumption A.6 (Additional Conditions on Perturbation and Noise) *Assume that there exists a* $\boldsymbol{\Sigma} \succ \boldsymbol{0}$ *such that* $\boldsymbol{\Sigma}_k \overset{k\to\infty}{\longrightarrow} \boldsymbol{\Sigma}$. *There exists some* $C_4 > 0$ *such that* $\mathbb{E}_k[\|\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\|^{2+C_4}] \overset{\text{a.s.}}{<} \infty$ *and* $\mathbb{E}[(\varepsilon_k^+ - \varepsilon_k^-)^{2+C_4} \big| \hat{\theta}_k, \boldsymbol{\Delta}_k] \overset{\text{a.s.}}{<} \infty$ *uniformly for all* k. *Finally,* $\boldsymbol{H}(\theta^*) \succ \boldsymbol{0}$.

Remark 2 *Note that under IID scenario for the observation noise, we have* $\mathbb{E}[(\varepsilon_k^+ - \varepsilon_k^-)^2 \big| \hat{\theta}_k, \boldsymbol{\Delta}_k] \to 2\mathrm{Var}(\ell(\theta^*, \omega))$ *a.s., where the variance is taken over* $\omega \in \Omega$. *This is due to* $\hat{\theta}_k \overset{\text{a.s.}}{\longrightarrow} \theta^*$ *shown Theorem 1 and* $c_k \to 0$ *assumed in A.5.*

Let us first show the property of our Hessian estimate described in Section 2.2.

Theorem 2 *Under aforementioned conditions, and assume* $\widetilde{c}_k = O(c_k)$, *we have* $\overline{\boldsymbol{H}}_k \overset{\text{a.s.}}{\longrightarrow} \boldsymbol{H}(\theta^*)$.

We now show the rate of convergence of HARP in Algorithm 1. According to A.5, we use $a_k = {}^{a}/{k^\alpha}$ and $c_k = {}^{c}/{k^\gamma}$ for $k \geq 0$, where

$$\alpha \in (1/2, 1] \ , \text{ and } \gamma \in (0, \alpha - 1/2) \, . \tag{19}$$

Granted, there are other forms for stepsizes (a_k, c_k). However, they do not necessarily provide improved rates [27]. Before stating Theorem 3, we introduce extra notations. Let $\tau = \alpha - 2\gamma$ and $\tau_+ = \tau \cdot \mathbb{I}_{\{\alpha=1\}}$. Let $\boldsymbol{\Gamma}_k = a\boldsymbol{H}(\overline{\theta}_k)$ with $\overline{\theta}_k$ being some convex combination of $\hat{\theta}_k$ and θ^*, $\boldsymbol{t}_k = -ak^{\tau/2}\beta_k(\hat{\theta}_k)$, and $\boldsymbol{v}_k \equiv -ak^{-\gamma}\boldsymbol{\xi}_k(\hat{\theta}_k)$.

Theorem 3 (Asymptotic Normality) *Assume A.1, A.2, A.3, A.4 or A.4', A.5, and A.6 hold. Pick* $a > \tau_+/[2\lambda_{\min}(\boldsymbol{H}(\theta^*))]$ *and* $\alpha \leq 6\gamma$, *we have*

$$k^{\tau/2}(\hat{\theta}_k - \theta^*) \overset{\text{dist.}}{\longrightarrow} \mathcal{N}(\mu, \boldsymbol{B}) \ , \tag{20}$$

where $(\mu, \boldsymbol{B})$ *satisfies the linear system (21) and the Lyapunov equation (22) respectively:*

$$\begin{cases} (\boldsymbol{\Gamma} - \tau_+\boldsymbol{I}/2)\mu = \boldsymbol{t} \, , & (21) \\ (\boldsymbol{\Gamma} - \tau_+\boldsymbol{I}/2)\boldsymbol{B} + \boldsymbol{B}(\boldsymbol{\Gamma}^T - \tau_+\boldsymbol{I}/2) = \dfrac{a^2\mathrm{Var}[\ell(\theta^*, \omega)]}{2c^2}\boldsymbol{\Sigma} \, . & (22) \end{cases}$$

In (21–22), $\boldsymbol{\Gamma} = \lim_{k\to\infty} \boldsymbol{\Gamma}_k = a\boldsymbol{H}(\theta^*)$*, the* $\mathrm{Var}[\ell(\theta^*, \omega)]$ *and* $\boldsymbol{\Sigma}$ *are defined in Remark 2 and A.6 respectively, and*

$$\boldsymbol{t} = \lim_{k\to\infty} \boldsymbol{t}_k = -\frac{ac^2}{6}\mathbb{I}_{\{\alpha=6\gamma\}}\mathbb{E}[L^{(3)}(\theta^*)\cdot(\boldsymbol{\Delta}\otimes\boldsymbol{\Delta}\otimes\boldsymbol{\Delta})\cdot\boldsymbol{m}(\boldsymbol{\Delta})]\,, \tag{23}$$

where $\boldsymbol{\Delta}$ *is* $\mathbf{0}$*-mean and* $\boldsymbol{\Sigma}^{-1}$*-covariance.*

Remark 3 *[1] provides the explicit solution to (22):*

$$\boldsymbol{B} = \frac{a^2\mathrm{Var}[\ell(\theta^*, \omega)]}{2c^2}\int_0^\infty e^{t(\tau+\boldsymbol{I}/2-\boldsymbol{\Gamma})}\boldsymbol{\Sigma}e^{t(\tau+\boldsymbol{I}/2-\boldsymbol{\Gamma}^T)}\mathrm{d}t\,. \tag{24}$$

4.2. CRN Scenario

This section considers the CRN noise scenario, where the *fastest* rate $O(k^{-1/2})$ for RMS is achieved when $\alpha = 1$ and $\gamma > 1/4$. Here, the bias-variance trade-off as arising in Lemma 1 no longer applies, see Lemma 2, whence Section 4.2 has a faster convergence rate compared to Section 4.1. The previous assumption on the noise is now changed for the CRN scenario.

Assumption A.3' (CRN) $\omega_k (= \omega_k^+ = \omega_k^-)$ *are i.i.d. and are independent from* $\mathcal{F}_k$*. Let* $\mathrm{g}(\cdot,\cdot) : \mathbb{R}^d \times \Omega \mapsto \mathbb{R}^d$ *be the partial derivative of* $\ell(\theta, \omega)$ *w.r.t.* θ*. Assume that* $\|\mathrm{g}(\theta, \omega)\|_\infty \le D_5$ *uniformly for all* θ *and a.s. for all* ω*.*

Lemma 2 (Second Moment of $\hat{\boldsymbol{g}}_k(\hat{\theta}_k)$) *When A.1, A.2, and A.3' hold,*

$$\mathbb{E}_k\{\|\hat{\boldsymbol{g}}_k(\hat{\theta}_k)^2\|\} \overset{\text{a.s.}}{=} \mathbb{E}\|\mathrm{g}(\hat{\theta}_k, \omega_k)\|^2 + o(1) \overset{\text{a.s.}}{=} \int_{\omega\in\Omega}\|\mathrm{g}(\hat{\theta}_k, \omega)\|^2\mathrm{d}\mathbb{P}(\omega) + o(1)\,. \tag{25}$$

The a.s. convergence result is similar to Theorem 1 or Theorem 1'. The corresponding proofs are similar using Lemma 2. We turn to finding the convergence rate directly. Before stating Theorem 4, we define some notations. Let $\alpha_+ \equiv \alpha\cdot\mathbb{I}_{\{\alpha=1\}}$. Let $\boldsymbol{\Gamma}_k = a\boldsymbol{H}(\overline{\theta}_k)$ with $\overline{\theta}_k$ being some convex combination of $\hat{\theta}_k$ and θ^*, $\boldsymbol{t}_k = -ak^{\alpha/2}\beta_k(\hat{\theta}_k)$, and $\boldsymbol{v}_k = -a\xi_k(\hat{\theta}_k)$.

Theorem 4 (Asymptotic Normality) *Assume A.1, A.2, A.3', A.4 or A.4', A.5, A.6. Pick* $a > \alpha_+/[2\lambda_{\min}(\boldsymbol{H}(\theta^*))]$ *and* $\alpha < 4\gamma$*, we have*

$$k^{\alpha/2}(\hat{\theta}_k - \theta^*) \xrightarrow{\text{dist.}} \mathcal{N}(\mathbf{0}, \boldsymbol{B})\,, \tag{26}$$

where $\boldsymbol{B}$ *satisfies*

$$(\boldsymbol{\Gamma} - \alpha_+\boldsymbol{I}/2)\boldsymbol{B} + \boldsymbol{B}(\boldsymbol{\Gamma}^T - \alpha_+\boldsymbol{I}/2) = a^2\Sigma\,. \tag{27}$$

Here, $\boldsymbol{\Gamma} = \lim_{k\to\infty}\boldsymbol{\Gamma}_k = a\boldsymbol{H}(\theta^*)$*, and* Σ *has elements*

$$\Sigma_{i,j} = \mathbb{I}_{\{i=j\}}\int_{\omega\in\Omega}\|\mathrm{g}(\theta^*, \omega)\|^2\mathrm{d}\mathbb{P}(\omega) + \mathbb{I}_{\{i\ne j\}}\int_{\omega\in\Omega}[\mathrm{g}(\theta^*, \omega)]_i[\mathrm{g}(\theta^*, \omega)]_j\mathrm{d}\mathbb{P}(\omega)\,, \tag{28}$$

where $[\mathrm{g}(\theta^*, \omega)]_i$ *denotes the ith component of* $\mathrm{g}(\theta^*, \omega)$*.*

Recall that in IID scenario, (20) involves a nonzero μ when the fastest rate $O(k^{-1/3})$ is achieved at $(\alpha, \gamma) = (1, 1/6)$. On the contrary, in the CRN scenario, the mean in (26) is zero when the fastest rate $O(k^{-1/2})$ is achieved whenever $(\alpha, \gamma) = (1, > 1/4)$.

Remark 4 *The asymptotic result shows that the covariance structure* $\boldsymbol{\Sigma}_k(\to \boldsymbol{\Sigma})$ *for* $\boldsymbol{\Delta}_k$ *no longer impacts the asymptotic normality (rate of convergence). Instead, the moments of* $\mathrm{g}(\theta^*, \omega)$ *takes over given the assumed differentiablility of the random function* $\ell(\theta, \omega)$ *in A.3'.*

4.3. Comparison Between HARP and SPSA

Let us see what happens when $\boldsymbol{\Sigma}_k \to \boldsymbol{\Sigma} = \boldsymbol{H}(\theta^*)$. Let us write out (24) in Remark 3 for $\alpha < 6\gamma$. Let the eigen-decomposition of $\boldsymbol{H}(\theta^*)$ be $\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T$, for orthogonal matrix $\boldsymbol{P}$ and diagonal matrix $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \cdots, \lambda_d)$. Then $\boldsymbol{B}$ in (22) equals $\boldsymbol{P}\boldsymbol{M}\boldsymbol{P}^T$, where the (i,j)th elements of $\boldsymbol{M}$ is

$$m_{i,j} = \frac{a^2 \operatorname{Var}(\ell(\theta^*, \omega))}{2c^2} (\boldsymbol{P}^T \boldsymbol{\Sigma} \boldsymbol{P})_{i,j} (a\lambda_i + a\lambda_j - \tau_+)^{-1} .$$

For all the algorithms listed in Subsection 1.1, with $\boldsymbol{\Sigma}_k = \boldsymbol{I}$, the trace of the covariance term is asymptotic to

$$\frac{a^2 \operatorname{Var}[\ell(\theta^*, \omega)]}{2c^2} \sum_{i=1}^{d} (2a\lambda_i - \tau_+)^{-1} , \tag{29}$$

whereas HARP in Algorithm 1, with $\boldsymbol{\Sigma}_k = \hat{\boldsymbol{H}}_k \to \boldsymbol{H}(\theta^*)$, gives

$$\frac{a^2 \operatorname{Var}[\ell(\theta^*, \omega)]}{2c^2} \sum_{i=1}^{d} \frac{1}{2a - \tau_+/\lambda_i} . \tag{30}$$

Note that both (29) and (30) diverge when *any* one of the eigenvalues of $\boldsymbol{H}(\theta^*)$ is close to zero. Nonetheless, (30) is smaller than (29) when $\lambda_i \ll 1$ for some $1 \le i \le d$, under which circumstance the iteration complexity (14) of HARP *can* be better than that of SPSA—at the cost of two additional ZO queries per iteration, see the last line in Algorithm 1.

5. Numerical Illustration

We now present two empirical examples to demonstrate the fast optimization and the wide applicability of HARP.

5.1. Synthetic Problem: Skew-Quartic Function

Section 4.3 demonstrates that HARP performs better under ill-conditioned problem. This synthetic example uses the skew-quartic function in [31] as the true loss $L(\cdot)$ in (1). The corresponding Hessian has one single large eigenvalue and $(d-1)$ close-to-zero eigenvalues. This loss function is poorly-conditioned. The noisy loss observation $\ell(\theta, \omega)$ in (1) is the true loss corrupted by an i.i.d. $\mathcal{N}(0,1)$ random noise. We use $d = 20$ and initialize $\hat{\theta}_0$ within $[-20, 20]^d$. We use $a_k = {}^{a}\!/_{(k+1+A)^\alpha}$ with $\alpha = 0.602$ and A equals 10% of the iteration number, $c_k = {}^{c}\!/_{(k+1)^\gamma}$ with $\gamma = 0.101$. Number of replicates is 25 (i.e., all the plots below are averaged performance over 25 replications). The corresponding implementation details ca be found at GitHub. The algorithm we compare against is SPSA [29], which has comparable/better performance than other algorithms reviewed in Section 1.1. During the implementation, both SPSA and HARP use exactly four ZO queries each iteration, so the query complexity *aligns* with the iteration complexity. We see from Figure 1 that that HARP with $\boldsymbol{\Sigma}_k = \hat{\boldsymbol{H}}_k$ outperforms SPSA with $\boldsymbol{\Sigma}_k = \boldsymbol{I}$ for the ill-conditioned problem of minimizing a skew-quartic function.

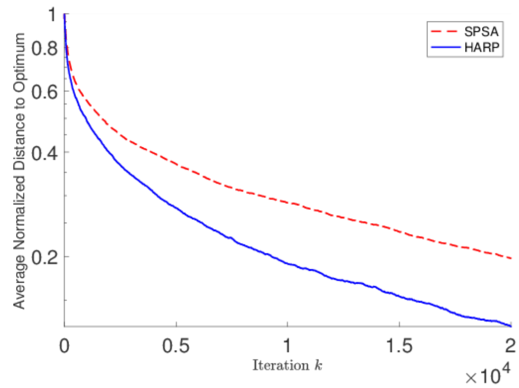

Figure 1: Performance of SPSA and HARP in terms of normalized distance $\|\hat{\theta}_k - \theta^*\| / \|\hat{\theta}_0 - \theta^*\|$ averaged across 25 independent replicates, and both algorithms use four ZO queries per iteration. The underlying loss function is the skew-quartic function with $d = 20$, and the noisy observation is corrupted by a $\mathcal{N}(0, 1)$ noise.

5.2. Universal Image Attack As A Finite-Sum Problem

We consider the problem of generating black-box adversarial examples universally for $I > 1$ images [6, 8] using zeroth-order optimization methods. We consider the constrained problem

$$\begin{cases} \min_{\theta} L(\theta) \equiv \underbrace{\kappa\|\theta\|_2^2}_{\equiv L_1(\theta)} + \underbrace{\frac{1}{I}\sum_{i=1}^{I} \mathrm{loss}(\zeta_i + \theta)}_{\equiv L_2(\theta)}\,, \\ \text{s.t. } (\zeta_i + \theta) \in [-0.5, 0.5]^d\,, \forall i, \end{cases} \tag{31}$$

where the constraint is to normalize the resulting pixels within the range $[-0.5, 0.5]^d$. The $\mathrm{loss}(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$ imposed on each image takes the form

$$\mathrm{loss}(\zeta) = \max_{i: 1\le i \le C} \left\{ \mathrm{ps}(\zeta, i) - \max_{j \neq i: 1 \le j \le C} [\mathrm{ps}(\zeta, j)] \right\}\,, \tag{32}$$

where $\mathrm{ps}(\zeta, i)$ denotes the prediction score of the i-th class given the input ζ. The model $\mathrm{ps}(\cdot, \cdot)$here is trained using the structure specified in [5]. Note that $\sum_{i=1}^{I} \mathrm{loss}(\zeta_i + \theta) = 0$ when the chosen images $\{\zeta_i\}_{i=1}^{I}$ are successfully attacked by the universal perturbation θ. The noisy loss observation $\ell(\theta, \omega)$ is

$$\ell(\theta, \omega) = \kappa\|\theta\|_2^2 + \frac{1}{J}\sum_{j=1}^{J} \mathrm{loss}(\zeta_{i_j(\omega)} + \theta)\,, \tag{33}$$

for $J \le I$, and the J indexes $\{i_1(\omega), \cdots, i_J(\omega)\}$ are i.i.d. uniformly drawn from $\{1, \cdots, I\}$ (without replacement).

Consider (31) with $\kappa = 1/10$. The I images arising in (31) are those *correctly* classified by the trained model. $d = 784$ for MNIST dataset. The algorithm we compare against is ZO-ADAMM [7]. Both algorithms are initialized at $\hat{\theta}_0 = \mathbf{0}$. The ZO-query per iteration for both algorithms is 60, so the query complexity *aligns* with the iteration complexity. We perform 25 independent replicates, each with $K = 1000$ iterations. The stepsizes are $a_k = {}^{a}/_{(k+1+A)^{0.602}}$ and $c_k = {}^{c}/_{(k+1)^{0.101}}$. The details of the hyper-parameters are in GitHub.

Algo	$\mathbb{E}[L(\hat{\boldsymbol{\theta}}_K)]$	$\{\mathrm{Var}[L(\hat{\boldsymbol{\theta}}_K)]\}^{\frac{1}{2}}$	$\mathbb{E}[L_2(\hat{\boldsymbol{\theta}}_K)]$
ADAMM	185.96	16.88	40.95
HARP	138.22	18	12.50

Table 1: Performance of ZO-ADAMM and HARP in terms of loss after $K = 1000$ iterations averaged across 25 independent replicates. The loss function $L(\cdot)$ is the sum of the magnitude cost $L_1(\cdot)$ and the attack loss $L_2(\cdot)$. Here $L_2(\cdot)$ measures the attack loss on $I = 100$ images of the letter one, and its *noisy* query is evaluated using a batch-size of one. A close-to-zero $L_2(\cdot)$ loss is equivalent to a close-to-one attack success rate.

Algo	$\mathbb{E}[L(\hat{\boldsymbol{\theta}}_K)]$	$\{\mathrm{Var}[L(\hat{\boldsymbol{\theta}}_K)]\}^{\frac{1}{2}}$	$\mathbb{E}[L_2(\hat{\boldsymbol{\theta}}_K)]$
ADAMM	56.95	6.89	11.75
HARP	18.46	1.37	0.13

Table 2: Here $L_2(\cdot)$ measures the attack loss on $I = 10$ images of the letter three, and its ZO query is noise-free.

Tables 1–2 summarize the terminal expected loss function evaluation $\mathbb{E}[L(\hat{\boldsymbol{\theta}}_{1000})]$, the terminal standard deviation of loss function evaluation $\{\mathrm{Var}[L(\hat{\boldsymbol{\theta}}_{1000})]\}^{\frac{1}{2}}$, and the terminal attack loss $\mathbb{E}[L_2(\hat{\boldsymbol{\theta}}_{1000})]$, all three of which are averaged across 25 independent replicates. The terminal scaled magnitude of the perturbation $\mathbb{E}[L_1(\hat{\boldsymbol{\theta}}_{1000})]$ can be computed as (31). Moreover, the noisier the loss function observation is (i.e., the larger the discrepancy between collected sample size J and the entire sample size I), the more difficult it is to reduce the $\mathbb{E}[L(\hat{\boldsymbol{\theta}}_K)]$ given a fixed iteration/query budget. In both noisy and noise-free ZO queries, HARP shows faster convergence rate than ZO-ADAMM given a fixed query-budget.

6. Concluding Remarks

This work proposes HARP to use the second-order approximation from ZO queries in both the random perturbation and the parameter update, and demonstrates its superiority in ill-conditioned problems theoretically in Section 4.3 and numerically in Section 5. Note that all the prior work use an identity/scalar matrix as the covariance matrix for the perturbation $\boldsymbol{\Delta}_k$ and use a *deterministic* mapping $\boldsymbol{m}_k(\cdot)$. This work shows the benefits of using non-identity matrix as the covariance matrix for $\boldsymbol{\Delta}_k$ and a *stochastic* mapping $\boldsymbol{m}_k(\cdot)$ which is $\mathcal{F}_k$-measurable. This generalization allows experimenters to incorporate various self-learning structure on the random directions $\boldsymbol{\Delta}_k$.

Some potential future work includes (1) the generalization to root-finding problem where the Jacobian matrix is possibly asymmetric[6]; (2) the generalization to the one-measurement counterpart to (3) as [30] to further reduce query complexity; (3) the extended discussion on global convergence

6. Note that in our discussion, the Hessian matrix for minimization problem is symmetric.

in line of [23]; (4) the extension to constrained minimization problems, and the follow-up discussion when sparsity-promoted constraints are imposed; (5) the potential exploration on (early) stopping SA iterations based on the root-mean-squared error; (6) other forms of $\mathbf{\Sigma}_k$, including diagonal forms to reduce floating point operations per iteration.

Acknowledgment

The author would like to thank Dr. Zhenliang Zhang, Dr. Jian Tan, and Dr. Wotao Yin for inspirational discussion.

References

[1] Richard H. Bartels and George W Stewart. Solution of the matrix equation ax+ xb= c [f4]. *Communications of the ACM*, 15(9):820–826, 1972.

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

[3] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[4] Julius R Blum. Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics*, pages 737–744, 1954.

[5] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Symposium on Security and Privacy*, pages 39–57. IEEE, 2017.

[6] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017.

[7] Xiangyi Chen, Sijia Liu, Kaidi Xu, Xingguo Li, Xue Lin, Mingyi Hong, and David Cox. Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization. In *Advances in Neural Information Processing Systems*, pages 7204–7215, 2019.

[8] Minhao Cheng, Thong Le, Pin-Yu Chen, Huan Zhang, JinFeng Yi, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. In *International Conference on Learning Representations*, 2018.

[9] Kai Lai Chung. *A course in probability theory*. Academic press, 2001.

[10] Yu M Ermol'ev. On the method of generalized stochastic gradients and quasi-fejér sequences. *Cybernetics*, 5(2):208–220, 1969.

[11] Yuri Ermoliev. Stochastic quasigradient methods and their application to system optimization. *Stochastics: An International Journal of Probability and Stochastic Processes*, 9(1-2):1–36, 1983.

[12] V Fabian. Stochastic approximation, optimization methods in statistics, 1971.

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

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

[15] V Ya Katkovnik and KULCHITS. OY. Convergence of a class of random search algorithms. *Automation and Remote Control*, 33(8):1321–1326, 1972.

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

[17] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[18] Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[19] Harold Joseph Kushner and Dean S Clark. *Stochastic approximation methods for constrained and unconstrained systems*, volume 26. Springer Science & Business Media, 1978.

[20] Tze Leung Lai. Extended stochastic lyapunov functions and recursive algorithms in linear stochastic systems. In *Stochastic Differential Systems*, pages 206–220. Springer, 1989.

[21] P Lancaster and HK Farahat. Norms on direct sums and tensor products. *mathematics of computation*, 26(118):401–414, 1972.

[22] Horia Mania, Aurelia Guy, and Benjamin Recht. Simple random search provides a competitive approach to reinforcement learning. *arXiv preprint arXiv:1803.07055*, 2018.

[23] John L Maryak and Daniel C Chin. Global random optimization by simultaneous perturbation stochastic approximation. In *Proceedings of the 2001 American Control Conference.(Cat. No. 01CH37148)*, volume 2, pages 756–762. IEEE, 2001.

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

[25] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[26] Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.

[27] Jerome Sacks. Asymptotic distribution of stochastic approximation procedures. *The Annals of Mathematical Statistics*, 29(2):373–405, 1958.

[28] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

[29] James C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control*, 37(3):332–341, 1992.

[30] James C Spall. A one-measurement form of simultaneous perturbation stochastic approximation. *Automatica*, 33(1):109–112, 1997.

[31] James C Spall. Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE transactions on automatic control*, 45(10):1839–1853, 2000.

[32] Haishan Ye, Zhichao Huang, Cong Fang, Chris Junchi Li, and Tong Zhang. Hessian-aware zeroth-order optimization for black-box adversarial attack. *arXiv preprint arXiv:1812.11377*, 2018.

[33] Jingyi Zhu. *Error Bounds and Applications for Stochastic Approximation with Non-Decaying Gain*. PhD thesis, Johns Hopkins University, 2020. http://jhir.library.jhu.edu/handle/1774.2/62504.

[34] Jingyi Zhu and James C Spall. Tracking capability of stochastic gradient algorithm with constant gain. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 4522–4527. IEEE, 2016.

[35] Jingyi Zhu and James C Spall. Probabilistic bounds in tracking a discrete-time varying process. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 4849–4854. IEEE, 2018.

[36] Jingyi Zhu and James C Spall. Stochastic approximation with nondecaying gain: Error bound and data-driven gain-tuning. *International Journal of Robust and Nonlinear Control*, 30(15): 5820–5870, 2020.

[37] Jingyi Zhu, Long Wang, and James C Spall. Efficient implementation of second-order stochastic approximation algorithms in high-dimensional problems. *Transactions on Neural Networks and Learning Systems*, 31(8):3087–3099, 2020.

Appendix A. Supplementary Proofs

Proof [Proof for Lemma 1] First consider the bias term $\boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k)$ of $\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)$ as an estimator for $\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)$.

$$\begin{aligned}
&\mathbb{E}_k[\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)] \\
&\overset{\text{a.s.}}{=} \mathbb{E}_k\left[\frac{L(\hat{\boldsymbol{\theta}}_k + c_k\boldsymbol{\Delta}_k) - L(\hat{\boldsymbol{\theta}}_k - c_k\boldsymbol{\Delta}_k)}{2c_k}\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\right] + \mathbb{E}_k\left[\frac{\boldsymbol{m}_k(\boldsymbol{\Delta}_k)}{2c_k}\mathbb{E}[(\varepsilon_k^+ - \varepsilon_k^-)\,|\, \hat{\boldsymbol{\theta}}_k, \boldsymbol{\Delta}_k]\right] && (34)\\
&\overset{\text{a.s.}}{=} \mathbb{E}_k[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\boldsymbol{\Delta}_k^T]\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k) + \frac{c_k^2}{12}\mathbb{E}_k\left\{[L^{(3)}(\overline{\boldsymbol{\theta}}_k^+) + L^{(3)}(\overline{\boldsymbol{\theta}}_k^-)](\boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k)\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\right\} && (35)\\
&\overset{\text{a.s.}}{=} \boldsymbol{g}(\hat{\boldsymbol{\theta}}_k) + \boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k)\,, && (36)
\end{aligned}$$

where equation (34) uses [9, Thm. 9.1.3 on p. 315], equation (35) uses the third-order Taylor expansion with mean-value forms of the remainder and $\mathbb{E}[\varepsilon_k^+ - \varepsilon_k^- \,|\, \hat{\boldsymbol{\theta}}_k, \boldsymbol{\Delta}_k] \overset{\text{a.s.}}{=} 0$ in A.3, equation (36) uses the expression (17) and $\mathbb{E}_k[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\boldsymbol{\Delta}_k^T] \overset{\text{a.s.}}{=} \boldsymbol{I}$ assumed in A.2. Then

$$\begin{aligned}
&\mathbb{E}_k[\|\boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k)\|] \\
&\overset{\text{a.s.}}{\leq} \frac{c_k^2}{6}\|L^{(3)}(\boldsymbol{\theta})\|_\infty \mathbb{E}_k[\|\boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k\|\|\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\|] && (37)\\
&\overset{\text{a.s.}}{=} \frac{c_k^2}{6} D_1 \mathbb{E}_k[\|\boldsymbol{\Delta}_k\|^3 \boldsymbol{m}_k(\boldsymbol{\Delta}_k)] && (38)\\
&\overset{\text{a.s.}}{\leq} \frac{c_k^2}{6} D_1 D_2\,, && (39)
\end{aligned}$$

where inequality (37) uses the mean-value theorem ($\int_D |f_1(x)f_2(x)|\,\mathrm{d}x \leq \sup_{x\in D}|f_1(x)|\int_D|f_2(x)|\,\mathrm{d}x$ for two functions f_1 and f_2 and some domain of integration D), equality (38) uses the independence between $\hat{\boldsymbol{\theta}}_k$ and $\boldsymbol{\Delta}_k$ assumed in A.2 and [21], and inequality (39) uses A.2. The representation of $\boldsymbol{\xi}_k(\hat{\boldsymbol{\theta}}_k)$ in (18) follows directly from (4) and (17).

We then consider the second-moment of $\boldsymbol{\xi}_k(\hat{\boldsymbol{\theta}}_k)$ through the following computation:

$$\begin{aligned}
&\mathbb{E}_k\left\{\|\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)\|^2\right\} \\
&\overset{\text{a.s.}}{=} \mathbb{E}_k\left\{\left\|\frac{L(\hat{\boldsymbol{\theta}}_k + c_k\boldsymbol{\Delta}_k) - L(\hat{\boldsymbol{\theta}}_k - c_k\boldsymbol{\Delta}_k)}{2c_k}\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\right\|^2\right\} && (40)\\
&\quad + \frac{1}{4c_k^2}\mathbb{E}_k[(\varepsilon_k^+ - \varepsilon_k^-)^2\|\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\|^2] && (41)\\
&\quad + \frac{1}{2c_k^2}\mathbb{E}_k\Big\{[L(\hat{\boldsymbol{\theta}}_k + c_k\boldsymbol{\Delta}_k) - L(\hat{\boldsymbol{\theta}}_k - c_k\boldsymbol{\Delta}_k)](\varepsilon_k^+ - \varepsilon_k^-)\|\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\|^2\Big\}. && (42)
\end{aligned}$$

The term on (41) becomes $O(c_k^{-2})$ because

$$\begin{aligned}
&\mathbb{E}_k[(\varepsilon_k^+ - \varepsilon_k^-)^2 \|\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\|^2] \\
&\overset{\text{a.s.}}{=} \mathbb{E}_k\left[\|\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\|^2 \mathbb{E}[(\varepsilon_k^+ - \varepsilon_k^-)^2 \,|\, \hat{\boldsymbol{\theta}}_k, \boldsymbol{\Delta}_k]\right] && (43)\\
&\overset{\text{a.s.}}{=} D_3 \cdot \mathbb{E}_k[\|\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\|^2] && (44)\\
&\overset{\text{a.s.}}{\le} D_3 D_2\,, && (45)
\end{aligned}$$

where inequality (43) uses [9, Thm. 9.1.3], inequality (44) uses A.3 and the independence between $\hat{\boldsymbol{\theta}}_k$ and $\boldsymbol{\Delta}_k$, and inequality (45) uses A.2. The term on (42) becomes zero thanks to [9, Thm. 9.1.3] and $\mathbb{E}[\,\varepsilon_k^+ - \varepsilon_k^- \,|\, \hat{\boldsymbol{\theta}}_k, \boldsymbol{\Delta}_k]$ assumed in A.3. The term on (40) can be bounded from above by $D_2\|\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)\|^2 + O(c_k^2)$, as

$$\begin{aligned}
&\mathbb{E}_k\left\{\left\|\frac{L(\hat{\boldsymbol{\theta}}_k + c_k\boldsymbol{\Delta}_k) - L(\hat{\boldsymbol{\theta}}_k - c_k\boldsymbol{\Delta}_k)}{2c_k}\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\right\|^2\right\} \\
&\overset{\text{a.s.}}{=} [\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)]^T \mathbb{E}_k\{\boldsymbol{\Delta}_k[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)]^T \boldsymbol{m}_k(\boldsymbol{\Delta}_k)\boldsymbol{\Delta}_k^T\}\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k) \\
&\quad + \frac{c_k^4}{144}\mathbb{E}_k\left\|[L^{(3)}(\overline{\boldsymbol{\theta}}_k^+) + L^{(3)}(\overline{\boldsymbol{\theta}}_k^-)](\boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k)\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\right\|^2 \\
&\quad + \frac{c_k^2}{6}[\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)]^T \mathbb{E}_k\left\{\boldsymbol{\Delta}_k[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)]^T [L^{(3)}(\overline{\boldsymbol{\theta}}_k^+) + L^{(3)}(\overline{\boldsymbol{\theta}}_k^-)] \times (\boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k)\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\right\} \\
&\overset{\text{a.s.}}{=} O\left(\|\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)\|^2\right) + O(c_k^2)\,, && (46)
\end{aligned}$$

thanks to A.2 and third-order Taylor expansion. ■

Proof [Illustration for Paragraph 4.1.1] The proof directly follows from the second-order Taylor expansion and the Lipschitz Hessian condition on the remainder terms.

$$\begin{aligned}
&\mathbb{E}_k[\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)] \\
&\overset{\text{a.s.}}{=} \mathbb{E}_k\left[\frac{L(\hat{\boldsymbol{\theta}}_k + c_k\boldsymbol{\Delta}_k) - L(\hat{\boldsymbol{\theta}}_k - c_k\boldsymbol{\Delta}_k)}{2c_k}\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\right] + \mathbb{E}_k\left[\frac{\boldsymbol{m}_k(\boldsymbol{\Delta}_k)}{2c_k}\mathbb{E}\left[(\varepsilon_k^+ - \varepsilon_k^-) \,|\, \hat{\boldsymbol{\theta}}_k, \boldsymbol{\Delta}_k\right]\right] \\
&\overset{\text{a.s.}}{=} \mathbb{E}_k[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\boldsymbol{\Delta}_k^T]\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k) + \frac{c_k}{4}\mathbb{E}_k\left\{\boldsymbol{\Delta}_k^T[\boldsymbol{H}(\overline{\boldsymbol{\theta}}_k^+) - \boldsymbol{H}(\overline{\boldsymbol{\theta}}_k^-)]\boldsymbol{\Delta}_k\right\} && (47)\\
&\overset{\text{a.s.}}{=} \boldsymbol{g}(\hat{\boldsymbol{\theta}}_k) + \boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k)\,,
\end{aligned}$$

where (47) follows from the second-order Taylor expansion. Then $\boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k)$ satisfies

$$\begin{aligned}
\mathbb{E}_k\|\boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k)\| &\overset{\text{a.s.}}{\le} \frac{c_k}{4}\mathbb{E}_k\left\{\boldsymbol{\Delta}_k^T\left[O(1)\|2c_k\boldsymbol{\Delta}_k\|\right]\boldsymbol{\Delta}_k\right\} && (48)\\
&\overset{\text{a.s.}}{=} O(c_k^2) && (49)
\end{aligned}$$

where the $O(1)$ in (48) represents the Lipschitz parameter of $\boldsymbol{H}(\cdot)$. Note that the explicit scaling constant in (49) is no longer available as (17). ■

Proof [Proof for Theorem 1]

Under assumptions A.4, and A.5, we known from [19, Thm. 2.3.1 on p. 39] that Thm. 1 holds when the following two conditions hold:

i) $\|\boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k)\| < \infty$ for all k and $\boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k) \to \mathbf{0}$ a.s.

ii) $\lim_{k\to\infty} \mathbb{P}\left\{\sup_{j\geq k} \|\sum_{i=k}^{j} a_i \boldsymbol{\xi}_i(\hat{\boldsymbol{\theta}}_k)\| \geq \eta\right\} = 0$ for any $\eta > 0$.

Obviously, i) holds thanks to Lemma 1. Under assumption A.3, $\boldsymbol{\xi}_k(\hat{\boldsymbol{\theta}}_k)$ defined in (18) is an $\mathcal{F}_k$-martingale. Using [18, Eq. (4.1.4)], we have

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{j\geq k} \|\sum_{i=k}^{j} a_i \boldsymbol{\xi}_i(\hat{\boldsymbol{\theta}}_i)\| \geq \eta\right\} \\
&\leq \eta^{-2}\mathbb{E}\|\sum_{i=k}^{\infty} a_i \boldsymbol{\xi}_i(\hat{\boldsymbol{\theta}}_i)\|^2 \quad (50)\\
&= \eta^{-2}\sum_{i=k}^{\infty} a_i^2 \mathbb{E}\|\boldsymbol{\xi}_i(\hat{\boldsymbol{\theta}}_i)\|^2, \quad (51)
\end{aligned}$$

where inequality (50) uses Markov's inequality, equality (51) uses $\mathbb{E}[]\boldsymbol{\xi}_i(\hat{\boldsymbol{\theta}}_i)^T\boldsymbol{\xi}_j(\hat{\boldsymbol{\theta}}_j)] = \mathbb{E}\{\boldsymbol{\xi}_i(\hat{\boldsymbol{\theta}}_i)^T\mathbb{E}[\boldsymbol{\xi}_j(\hat{\boldsymbol{\theta}}_j)\big|\, \hat{\boldsymbol{\theta}}_j]\} = 0$ for all $i < j$. Given A.5, ii) is also satisfied. The a.s. convergence from $\hat{\boldsymbol{\theta}}_k$ to $\boldsymbol{\theta}^*$ is arrived. ■

Proof [Proof for Theorem 1'] Let us first show part i). Under A.4', we have

$$\begin{aligned}
&\mathbb{E}_k[L(\hat{\boldsymbol{\theta}}_k)] \\
&\overset{\text{a.s.}}{\leq} \mathbb{E}_k\left\{L(\hat{\boldsymbol{\theta}}_k) - a_k[\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)]^T\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k) + \frac{D_4 a_k^2}{2}\|\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)\|^2\right\} \quad (52)\\
&\overset{\text{a.s.}}{\leq} L(\hat{\boldsymbol{\theta}}_k) - a_k\|\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)\|^2 + a_k O(c_k^2)\|\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)\| + \frac{D_4 a_k^2}{2}\left[O(c_k^2) + O(c_k^{-2}) + O(\|\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)\|)^2\right] \quad (53)\\
&\overset{\text{a.s.}}{=} L(\hat{\boldsymbol{\theta}}_k) - a_k\|\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)\|^2 + O(a_k c_k^2)\|\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)\| + O(a_k^2 c_k^2)) + O\left(\frac{a_k^2}{c_k^2}\right) + O(a_k^2)\|\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)\|^2\\
&\overset{\text{a.s.}}{\leq} L(\hat{\boldsymbol{\theta}}_k) - \frac{a_k}{2}\left(\|\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)\| - O(c_k^2)\right)^2 + O\left(a_k^2 c_k^2\right) + O(a_k^2 c_k^{-2}), \text{for large } k \text{ s.t. } O(a_k) < 1/2, \quad (54)
\end{aligned}$$

where (52) uses A.4' and mean-value theorem, (53) uses Cauchy-Schwartz inequality and (40)–(42), and (54) uses A.5.

Therefore, for sufficiently large k, we have

$$\mathbb{E}_k[L(\hat{\boldsymbol{\theta}}_k) - L(\boldsymbol{\theta}^*)] \overset{\text{a.s.}}{\leq} L(\hat{\boldsymbol{\theta}}_k) - L(\boldsymbol{\theta}^*) + O(a_k^2 c_k^2) + O(a_k^2 c_k^{-2}) \quad - \frac{a_k}{2}\left(\|\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)\| - O(c_k^2)\right)^2, \quad (55)$$

Under A.4' and A.5, [20, Thm. 1] ensures that the nonnegative random variable $[L(\hat{\boldsymbol{\theta}}_k) - L(\boldsymbol{\theta}^*)]$ converges to a *finite* random variable on a.s. Now that A.4' assumes $\sup\{\|\boldsymbol{\theta}\| : L(\boldsymbol{\theta}) \leq L(\boldsymbol{\theta}^*) + C_1\}$, the boundedness of $L(\hat{\boldsymbol{\theta}}_k)$ a.s. implies the iterate boundedness $\sup_k \|\hat{\boldsymbol{\theta}}_k\| < \infty$ a.s.

Next we show part ii). When (55) hold, [26] ensures that $\lim_{k\to\infty}\sum_{i=1}^{k} a_i[\|\boldsymbol{g}(\hat{\boldsymbol{\theta}}_i)\| - O(c_i^2)]^2 < \infty$ a.s. Together with A.5, we have $\|\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)\| \to 0$ as $k \to \infty$ a.s.

For any *fixed* sample point within a subset of $\Omega \times \Omega_{\boldsymbol{\Delta}}$ with a measure of 1, the sequence $\{\hat{\theta}_0, \cdots, \hat{\theta}_k, \cdots\}$ is a bounded sequence per i). By Bolzano-Weierstrass theorem, we can pick a sub-sequence $\{\hat{\theta}_{k_0}, \cdots, \hat{\theta}_{k_i}, \cdots\}$ such that $\|\boldsymbol{g}(\hat{\theta}_{k_i})\| \to \mathbf{0}^+$ as $i \to \infty$ a.s. Moreover, the fact that $\|\boldsymbol{g}(\hat{\theta}_k)\| \to 0$ a.s. and the smoothness of $\boldsymbol{g}(\cdot)$ ensure that the limit point of the sub-sequence $\{\hat{\theta}_{k_0}, \cdots, \hat{\theta}_{k_i}, \cdots\}$ as $i \to \infty$ coincides with the limit point of the entire sequence $\{\hat{\theta}_0, \cdots, \hat{\theta}_k, \cdots\}$ as $k \to \infty$. Finally, A.4' asserts that θ^* is the unique minimizer such that all neighboring points around it have nonzero gradient evaluation, so the claim in ii) is shown. ■

Proof [Proof for Theorem 2] First consider the term $\widetilde{c}_k^{-1}\overline{\ell}_k \boldsymbol{m}_k(\widetilde{\boldsymbol{\Delta}}_k)$.

$$\mathbb{E}(\widetilde{c}_k^{-1}\overline{\ell}_k \boldsymbol{m}_k(\widetilde{\boldsymbol{\Delta}}_k) \Big| \hat{\theta}_k, \boldsymbol{\Delta}_k) \overset{\text{a.s.}}{=} \boldsymbol{g}(\hat{\theta}_k + c_k\boldsymbol{\Delta}_k) - \boldsymbol{g}(\hat{\theta}_k - c_k\boldsymbol{\Delta}_k) + O(c_k^3)\,, \tag{56}$$

where the $O(c_k^3)$ term in (56) is the difference of the two $O(c_k^2)$ bias terms in the one-sided gradient approximations for $\boldsymbol{g}(\hat{\theta}_k \pm c_k\boldsymbol{\Delta}_k)$ in $\widetilde{c}_k^{-1}\overline{\ell}_k \boldsymbol{m}_k(\widetilde{\boldsymbol{\Delta}}_k)$ and $\widetilde{c}_k = O(c_k)$. Hence, by an expansion of each of $\boldsymbol{g}(\hat{\theta}_k \pm c_k\boldsymbol{\Delta}_k)$, we have for any i, j

$$\mathbb{E}\left(\frac{\overline{\ell}_k}{2c_k\widetilde{c}_k} \boldsymbol{m}_k(\widetilde{\boldsymbol{\Delta}}_k)[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)]^T \Big| \mathcal{F}_k, \boldsymbol{\Delta}_k \right) \overset{\text{a.s.}}{=} \boldsymbol{H}(\hat{\theta}_k) + O(c_k^2)\,, \tag{57}$$

where (57) uses (56) and $\mathbb{E}_k(\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\boldsymbol{\Delta}_k^T) = \boldsymbol{I}$ in A.2. Note that the $O(c_k^2)$ term in (57) absorbs higher-order terms in the Taylor expansion of $\boldsymbol{g}(\hat{\theta}_k + c_k\boldsymbol{\Delta}_k) - \boldsymbol{g}(\hat{\theta}_k - c_k\boldsymbol{\Delta}_k)$ in (56).

Another symmetrization operation of $(2c_k\widetilde{c}_k)^{-1}\overline{\ell}_k \boldsymbol{m}_k(\widetilde{\boldsymbol{\Delta}}_k)[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)]^T$ gives the latter part of (5), in order to ensure a symmetric Hessian estimate.

Given (57), the statement that $\overline{\boldsymbol{H}}_k \xrightarrow{\text{a.s.}} \boldsymbol{H}(\theta^*)$ follows from the Theorem 1 or Theorem 1', the updating recursion (5), the algorithmic form in Algorithm 1 and the corresponding analysis in [37].
■

Proof [Proof for Theorem 3] The asymptotic normality result will be shown once the conditions (2.2.1), (2.2.2), and (2.2.3) of [13] hold.

We first show that [13, Eq. (2.2.1)] hold. We see that $\boldsymbol{\Gamma}_k \to a\boldsymbol{H}(\theta^*)$ a.s. by the result in Thm. 1 and the continuity of $\boldsymbol{H}(\cdot)$ as assumed in A.1. When $\alpha < 6\gamma$, we have $\boldsymbol{t}_k \to \mathbf{0}$ a.s., as Lemma 1 shows that $\|\boldsymbol{\beta}_k(\hat{\theta}_k)\| = O(c_k^2) = O(k^{-2\gamma})$ a.s. When $\alpha = 6\gamma$, using A.2 and Thm. 1, we know that $\boldsymbol{t}_k = -a(k+1)^{2\gamma} \cdot O(c_k^2) = O(1)$. Using (17), A.1, and Thm. 1, we have

$$\boldsymbol{\beta}_k \overset{k\to\infty}{\longrightarrow} \frac{1}{6} c_k^2 \mathbb{E}[L^{(3)}(\theta^*) \cdot (\boldsymbol{\Delta} \otimes \boldsymbol{\Delta} \otimes \boldsymbol{\Delta}) \cdot \boldsymbol{m}(\boldsymbol{\Delta})] \,\text{a.s.}\,, \tag{58}$$

thanks to the dominated convergence theorem. Multiplying $-a(k+1)^{\tau/2} = -a(k+1)^{2\gamma}$ on both sides of (58) gives (23). Combined the cases for $\alpha < 6\gamma$ and $\alpha = 6\gamma$, we know that $\boldsymbol{t}_k$ converges to a finite vector for $\alpha \leq 6\gamma$.

We then show that [13, Eq. (2.2.2)] hold. By definition (4), $\boldsymbol{\xi}_k(\hat{\boldsymbol{\theta}}_k)$ is a $\mathcal{F}_k$-measurable martingale sequence, and so is $\boldsymbol{v}_k$.

$$
\begin{aligned}
&\mathbb{E}_k(\boldsymbol{v}_k\boldsymbol{v}_k^T) \\
&\overset{\text{a.s.}}{=} \frac{a^2}{(k+1)^{2\gamma}}\left(\mathbb{E}_k\{\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)[\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)]^T\} - \mathbb{E}_k[\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)]\{\mathbb{E}_k[\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)]\}^T\right) && (59)\\
&\overset{\text{a.s.}}{=} \frac{a^2}{c^2}c_k^2\mathbb{E}_k\{\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)[\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)]^T\} + \frac{a^2}{c^2}c_k^2[\boldsymbol{g}_k(\hat{\boldsymbol{\theta}}_k)+\boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k)][\boldsymbol{g}_k(\hat{\boldsymbol{\theta}}_k)+\boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k)]^T && (60)\\
&\overset{\text{a.s.}}{=} \frac{a^2}{c^2}\cdot\mathbb{E}_k\left[\left(\frac{\varepsilon_k^+ - \varepsilon_k^-}{2}\right)^2\boldsymbol{m}_k(\boldsymbol{\Delta}_k)[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)]^T\right] + o(1)\\
&\overset{\text{a.s.}}{=} \frac{a^2}{4c^2}\mathbb{E}_k\left\{\boldsymbol{m}_k(\boldsymbol{\Delta}_k)[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)]^T\mathbb{E}[(\varepsilon_k^+ - \varepsilon_k^-)^2\,|\,\hat{\boldsymbol{\theta}}_k,\boldsymbol{\Delta}_k]\right\} + o(1)\\
&\overset{\text{a.s.}}{=} \frac{a^2}{c^2}\frac{2\mathrm{Var}[\ell(\boldsymbol{\theta}^*,\omega)]}{4}\mathbb{E}\{\boldsymbol{m}_k(\boldsymbol{\Delta}_k)[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)]^T\} + o(1) && (61)\\
&\xrightarrow{\text{a.s.}} \frac{a^2\mathrm{Var}[\ell(\boldsymbol{\theta}^*,\omega)]}{2c^2}\boldsymbol{\Sigma},\ \text{as } k\to\infty\,, && (62)
\end{aligned}
$$

where (59) follows from (4), the $o(1)$ term on (60) is due to A.2, (17), Lemma 1, and Theorem 1, both (61) and (62) are due to A.6 and Remark 2.

We finally show that either (2.2.3) or (2.2.4) in [13] hold. That is, for every $\eta > 0$, $\lim_{k\to\infty}\mathbb{E}(\|\boldsymbol{v}_k\|^2\mathbb{I}_{\{\|\boldsymbol{v}_k\|^2\geq\eta k^\alpha\}}) = 0$. For any $C_5 \in (0, C_4/2)$, we have

$$
\begin{aligned}
&\lim_{k\to\infty}\mathbb{E}\left(\|\boldsymbol{v}_k\|^2\mathbb{I}_{\{\|\boldsymbol{v}_k\|^2\geq\eta k^\alpha\}}\right)\\
&\leq \limsup_{k\to\infty}[\mathbb{P}(\|\boldsymbol{v}_k\|^2\geq\eta k^\alpha)]^{\frac{C_5}{1+C_4}}\cdot[\mathbb{E}(\|\boldsymbol{v}_k\|^{2(1+C_5)})]^{\frac{1}{1+C_5}}\\
&\leq \limsup_{k\to\infty}\left(\frac{\mathbb{E}(\|\boldsymbol{v}_k\|^2)}{\eta k^\alpha}\right)^{\frac{C_5}{1+C_4}}\cdot[\mathbb{E}(\|\boldsymbol{v}_k\|^{2(1+C_5)})]^{\frac{1}{1+C_5}}\,, && (63)
\end{aligned}
$$

where the first inequality is due to Holder's inequality and the second inequality is due to Markov's inequality.

Using Minkowski inequality, we have $\|\boldsymbol{v}_k\|^{2(1+C_5)} \leq 2(1+C_5)k^{-2(1+C_5)\gamma}\left[\|\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)^{2(1+C_5)} + \|\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)\|^{2(1+C_5)} + \|\boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k)\|^{2(1+C_5)}\|\right]$. From Lemma 1 and A.4, we know that there exists some K such that both $\boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k)$ and $\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)$ are uniformly bounded a.s. for all $k \geq K$. Lemma 1 also implies that $\|\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)\| = O(c_k^{-2})$. Combined, we have $\mathbb{E}\|\boldsymbol{v}_k\|^{2(1+C_5)} = O(1)$.

Now that all relevant conditions in [13] are met to ensure the asymptotic normality. ■

Proof [Proof of Lemma 2]

Under A.3',

$$
\begin{aligned}
&\mathbb{E}_k[\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)[\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)]^T]\\
&\overset{\text{a.s.}}{=} \frac{1}{4c_k^2}\mathbb{E}_k\left\{\boldsymbol{m}_k(\boldsymbol{\Delta}_k)[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)]^T\times[\ell(\hat{\boldsymbol{\theta}}_k + c_k\boldsymbol{\Delta}_k,\omega_k) - \ell(\hat{\boldsymbol{\theta}}_k - c_k\boldsymbol{\Delta}_k,\omega_k)]^2\right\}\\
&\overset{\text{a.s.}}{=} \frac{1}{4c_k^2}\mathbb{E}_k\left\{\boldsymbol{m}_k(\boldsymbol{\Delta}_k)[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)]^T\times\mathbb{E}[[\ell(\hat{\boldsymbol{\theta}}_k + c_k\boldsymbol{\Delta}_k,\omega_k) - \ell(\hat{\boldsymbol{\theta}}_k - c_k\boldsymbol{\Delta}_k,\omega_k)]^2\,\middle|\,\hat{\boldsymbol{\theta}}_k,\boldsymbol{\Delta}_k]\right\}. && (64)
\end{aligned}
$$

Similar to the third-order Taylor expansion in Lemma 1, we have

$$\begin{aligned}
&\frac{1}{4c_k^2}\mathbb{E}[[\ell(\hat{\theta}_k + c_k\boldsymbol{\Delta}_k, \omega_k) - \ell(\hat{\theta}_k - c_k\boldsymbol{\Delta}_k, \omega_k)]^2 \Big| \hat{\theta}_k, \boldsymbol{\Delta}_k] \\
&\stackrel{\text{a.s.}}{=} \mathbb{E}\left\{ [\boldsymbol{\Delta}_k^T \mathrm{g}(\hat{\theta}_k, \omega_k)]^2 \Big| \hat{\theta}_k, \boldsymbol{\Delta}_k \right\} + O(c_k^4) \\
&\stackrel{\text{a.s.}}{=} \left[\boldsymbol{\Delta}_k^T \mathrm{g}(\hat{\theta}_k, \omega_k)\right]^2 + O(c_k^4)\,.
\end{aligned} \tag{65}$$

Whence, (64) becomes

$$\begin{aligned}
&\mathbb{E}_k[\hat{\boldsymbol{g}}_k(\hat{\theta}_k)[\hat{\boldsymbol{g}}_k(\hat{\theta}_k)]^T] \\
&\stackrel{\text{a.s.}}{=} \mathbb{E}_k\left\{ \boldsymbol{m}_k(\boldsymbol{\Delta}_k)\boldsymbol{\Delta}_k^T \mathrm{g}(\hat{\theta}_k, \omega_k)[\mathrm{g}(\hat{\theta}_k, \omega_k)]^T \boldsymbol{\Delta}_k [\boldsymbol{m}_k(\boldsymbol{\Delta}_k)]^T \right\} + o(1)\,.
\end{aligned} \tag{66}$$

Now that A.2 assumes independence between $\hat{\theta}_k$ and $\boldsymbol{\Delta}_k$, then the $(i, j)-$th component of (66) equals the following a.s.:

$$\mathbb{E}\left[\sum_{p=1}^{d}\sum_{q=1}^{d} m_{k,i}\Delta_{k,p}\Delta_{k,q}m_{k,j}\right] \cdot \mathbb{E}_k\left(\mathrm{g}_{k,p}\mathrm{g}_{k,q}\right) + o(1)$$

$$\stackrel{\text{a.s.}}{=} \left[\mathbb{I}_{\{i=j\}}\mathbb{I}_{\{p=q\}} + \mathbb{I}_{\{i\neq j\}}(\mathbb{I}_{\{p=i,q=j\}} + \mathbb{I}_{\{p=j,q=i\}})\right] \times \mathbb{E}_k\left(\mathrm{g}_{k,p}\mathrm{g}_{k,q}\right) + o(1) \tag{67}$$

$$\stackrel{\text{a.s.}}{=} \begin{cases} \sum_{p=1}^{d} \mathbb{E}_k(\mathrm{g}_{k,p})^2 + o(1), & \text{if } i = j, \\ 2\mathbb{E}_k(\mathrm{g}_{k,i}\mathrm{g}_{k,j}) + o(1), & \text{if } i \neq j. \end{cases} \tag{68}$$

where $m_{k,i}$ is the ith component of $\boldsymbol{m}_k(\boldsymbol{\Delta}_k)$, $\Delta_{k,p}$ is the pth component of $\boldsymbol{\Delta}_k$, $\mathrm{g}_{k,p}$ is the pth component of $\mathrm{g}(\hat{\theta}_k, \omega_k)$, equality (67) uses $\mathbb{E}_k[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\boldsymbol{\Delta}_k^T] = \boldsymbol{I}$ in A.2. Taking the diagonal terms of (68) gives (25). ■

Proof [Proof for Theorem 4]

We first show that [13, Eq. (2.2.1)] hold. As in the proof for Thm. 3, $\boldsymbol{\Gamma}_k \to a\boldsymbol{H}(\theta^*)$ a.s. When $\alpha < 4\gamma$, $\boldsymbol{t}_k = O(k^{\alpha/2-2\gamma}) \to \boldsymbol{0}$. Hence, [13, Eq. (2.2.1)] is met.

We then show [13, Eq. (2.2.2)] hold. Following the same reasoning as (60), we have

$$\mathbb{E}_k(\boldsymbol{v}_k\boldsymbol{v}_k^T) \stackrel{\text{a.s.}}{=} a^2\mathbb{E}_k\{\hat{\boldsymbol{g}}_k(\hat{\theta}_k)[\hat{\boldsymbol{g}}_k(\hat{\theta}_k)]^T\} + o(1)\,, \tag{69}$$

which is exactly (68). Under A.3', ω_k is independent from $\mathcal{F}_k$, we have $\mathbb{E}_k\mathrm{g}_{k,p}^2 \stackrel{\text{a.s.}}{=} \mathbb{E}\mathrm{g}_{k,p}^2 \stackrel{\text{a.s.}}{=} \int_{\omega\in\Omega}[\mathrm{g}(\hat{\theta}_k, \omega)]_p^2 \mathrm{d}\mathbb{P}(\omega) \xrightarrow{\text{a.s.}} \int_{\omega\in\Omega}[\mathrm{g}(\theta^*, \omega)]_p^2 \mathrm{d}\mathbb{P}(\omega)$ as $k \to \infty$, where the asymptotic relationship is due to dominated convergence theorem and A.3'. Following the same line of reasoning, $\mathbb{E}_k(\mathrm{g}_{k,i}\mathrm{g}_{k,j}) \stackrel{\text{a.s.}}{=} \int_{\omega\in\Omega}[\mathrm{g}(\hat{\theta}_k, \omega)]_i[\mathrm{g}(\hat{\theta}_k, \omega)]_j \mathrm{d}\mathbb{P}(\omega) \xrightarrow{\text{a.s.}} \int_{\omega\in\Omega}[\mathrm{g}(\theta^*, \omega)]_i[\mathrm{g}(\theta^*, \omega)]_j \mathrm{d}\mathbb{P}(\omega)$. Combined, we have (28).

The proof of showing [13, Eq. (2.2.3)] is exactly the same as that in proof for Theorem 3. ■