Proceedings of Machine Learning Research vol 145:1137–1160, 2021 2nd Annual Conference on Mathematical and Scientific Machine Learning

Hessian-Aided Random Perturbation (HARP) Using Noisy Zeroth-Order Queries

Jingyi Zhu JINGYI.ZHU@JHU.EDU
DAMO Academy, Alibaba Inc., Bellevue, WA

Editors: Joan Bruna, Jan S Hesthaven, Lenka Zdeborova

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 unit-covariance distribution. In contrast, this work considers the generalization where the perturbations may have non-isotropic covariance matrix constructed from the ZO queries. We propose to feed the Hessian-inverse approximation into the covariance of the random perturbation, so it is dubbed as Hessian-Aided Random Perturbation (HARP). HARP collects two or more (depending on the specific estimator form) zeroth-order queries per iteration to form approximations for both the gradient and the Hessian. We show the almost surely convergence 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 with isotropic-covariance random perturbation. [1]

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, see Spall (2005, Chaps. 3–8) for details. When only noisy zeroth-order (ZO) oracle is available (see Larson et al. (2019) for a comprehensive review), it is common practice to generate deterministic perturbation (Kiefer and Wolfowitz, 1952; Blum, 1954) or random perturbation (Ermol'ev, 1969; Katkovnik and OY, 1972; Spall, 1992) in finding extrema. SA methods that utilize ZO oracle only have regained their popularity in evolutionary strategy (as an alternative to reinforcement learning) (Salimans et al., 2017; Mania et al., 2018) and adversarial image attack (Kurakin et al., 2016; Carlini and Wagner, 2017). To the best of our knowledge, majorities of the existing random-perturbation-based methods generate the perturbation from a distribution with a mean of zero and a covariance of identity of scalar matrix, which enforce that every component of the perturbation vector is of the same magnitude on average and is independent with all other components. The resulting gradient estimate may not be robust to scaling and correlation of different parameters, and the non-robust estimation may further slow down the optimization process. Therefore, this work 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

[1] Part of the work was presented at the 12th International Workshop—a venue that does *not* have publication proceedings—on "Optimization for Machine Learning" as a part of the NeurIPS 2020 conference.

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 system stochasticity be represented by the random variable $\omega \in \Omega$, whose underlying distribution $\mathbb{P}$ is generally unknown. Consider finding the minimizer for a twice-differentiable bounded-from-below loss function $L(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$:

$$\theta^* \equiv \arg\min_{\theta \in \mathbb{R}^d} L(\theta)\,, \text{ where } L(\theta) \equiv \mathbb{E}_{\omega \sim \mathbb{P}}\left[\ell(\theta, \omega)\right]\,. \tag{1}$$

In (1), 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 $L(\theta)$ when one realization of $\omega \sim \mathbb{P}$ is drawn from Ω. Besides, the evaluation of the noisy ZO queries $\ell(\theta, \omega)$ is generally *expensive*. Under this setting, we implement 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. Let us focus 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 $\boldsymbol{0}$-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), and it typically arises in *simulation-based* optimization.

1.1. Prior Work on Gradient Estimation Using ZO Queries

The generic form for gradient estimate in (3) subsumes random direction stochastic approximation (RDSA) (Ermol'ev, 1969; Ermoliev, 1983) 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) (Katkovnik and OY, 1972) with $\boldsymbol{\Delta}_k$ being standard multivariate normally distributed and $\boldsymbol{m}_k(\boldsymbol{\Delta}_k) = \boldsymbol{\Delta}_k$, simultaneous perturbation stochastic approximation (SPSA) (Spall, 1992) with each component of $\boldsymbol{\Delta}_k$ being Rademacher distributed and $\boldsymbol{m}_k(\boldsymbol{\Delta}_k) = \boldsymbol{\Delta}_k$, and many other variants. Overall, majorities of SA schemes enforce the covariance matrix $\boldsymbol{\Sigma}_k^{-1} \succ \boldsymbol{0}$ for the random perturbation $\boldsymbol{\Delta}_k$ to be the *identity* matrix or a *scalar* matrix. Although the randomized scheme (3) exists for a long time and demonstrates numerical advantages over FDSA (Kiefer and Wolfowitz, 1952), theoretical foundation regarding the *optimal* choices of $\boldsymbol{\Delta}_k$ is lacking and extra caution is required in its implementation.

We propose a novel algorithm dubbed as Hessian-aided random perturbation (HARP). The choice of feeding Hessian approximation into $\boldsymbol{\Sigma}_k$ is motivated by mitigating the shortcomings of

$\boldsymbol{\Sigma}_k = \boldsymbol{I}$ in Section 2.1, analyzed theoretically through proving the almost surely convergence and convergence rate in Section 4.2, and demonstrated through two numerical experiments in Section 5. Previously, in both stochastic optimization (Spall, 2000) and deterministic optimization, the Hessian approximation 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.

Maheswaranathan et al. (2018) considers a non-isotropic $\boldsymbol{\Sigma}_k$, yet it is built upon the assumption that both ZO and first-order queries are accessible—which is no longer derivative-free/black-box. Ye et al. (2018) shares some similarities with us in terms of leveraging the Hessian estimate to achieve faster convergence. The results therein have to be interpreted carefully: the random perturbation $\boldsymbol{\Delta}_k$ affects both the gradient and the Hessian estimates at every iteration, yet the proofs ignore the randomness lying in the Hessian estimate.

1.2. Overview and Contribution

The remainder of this paper is organized as follows. Sect. 2 conveys the motivation behind adding structures to the covariance matrix for the random perturbation and presents implementation details of HARP. Sect. 4 provides theoretical justification (including the a.s. convergence and the rate of convergence) for (2) under IID noise. Sect. 5 illustrates the numerical performance of HARP and other ZO algorithms. Sect. 6 includes some concluding remarks and envisions some future directions. Before proceeding, let us outline the key contributions.

- We present a general framework of gradient estimation techniques and prove its theoretical properties (including almost surely convergence and rate of convergence). Its ZO queries per-iteration may range from one to four based on Zhu (2021), depending on the structure of the perturbation sequence. This framework unifies existing methods in Sect. 1.1 where the mapping $\boldsymbol{m}_k(\cdot)$ in (3) is *deterministic*. This framework allows us to explore new possible ways of constructing perturbation and design new algorithms beyond the existing ones by allowing $\boldsymbol{m}_k(\cdot)$ to be *random* and *measurable* with respect to the information available up till time k.

- We propose a new algorithm called "Hessian-Aided Random Perturbation" (HARP). The choice of feeding Hessian approximation into $\boldsymbol{\Sigma}_k$ is driven by our theoretical analysis in Section 4 and demonstrated in our numerical experiments in Section 5. Previously, in both stochastic optimization (Spall, 2000) and deterministic optimization, the Hessian is applied in parameter update *only*. HARP adaptively changes the covariance $\boldsymbol{\Sigma}_k$ 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. Theory and numerical experiments show that HARP outperforms other gradient approximation schemes whose random perturbation has an identity/scalar covariance matrix, especially for ill-conditioned problems.

2. Motivation and Pseudo-Code

Let us provide the motivation for HARP here and then present the pseudo code.

2.1. Why Use Non-Isotropic Covariance Σ^{-1} for Random Perturbation Δ?

Prior work summarized Section 1.1 enforce $\boldsymbol{\Sigma}_k = \boldsymbol{I}$. Let us illustrate the potential setback while estimating the gradient in (3) via SPSA/SFSA scheme with $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma} = \boldsymbol{I}$ and $\boldsymbol{m}(\boldsymbol{\Delta}) = \boldsymbol{\Delta}$. The RDSA algorithm where $\boldsymbol{\Sigma}$ is a scalar matrix can be similarly discussed.

2.1.1. Different Scalings on Individual Directions

One salient feature of $\boldsymbol{\Sigma} = \boldsymbol{I}$ is *equal* diagonal elements. Note that each diagonal element of $\boldsymbol{\Sigma}^{-1/2}$ *linearly* impacts the absolute value of the corresponding component of $\boldsymbol{\Delta}$. Consequently, the resulting SPSA/SFSA estimate $\hat{\boldsymbol{g}}(\hat{\theta})$ will subsequently perturb every component of $\hat{\theta}$ in (2) by the *same* magnitude *on average*. Naturally, feeding Hessian estimate into $\boldsymbol{\Sigma}$ arises from generating gradient estimate that is robust to different scalings on each direction of the underlying loss function.

Suppose we try to minimize $L(\theta) = (100\theta_1^2 + \theta_2^2)$ whose optimum is the origin. For illustration, suppose the initial estimate $\hat{\theta} = [1, 1]^T$, the differencing magnitude $c = 0.1$, and we get to observe noise-free loss function for the moment.

SPSA uses independent Rademacher-distributed random perturbation $\boldsymbol{\Delta}^{\text{SPSA}}$, whose four equally-likely possible values are $[1, 1]^T, [1, -1]^T, [-1, -1]^T, [-1, 1]^T$. The expectation of the gradient estimate $\mathbb{E}_{\boldsymbol{\Delta}^{\text{SPSA}}}[\hat{\boldsymbol{g}}^{\text{SPSA}}(\hat{\theta})]$ equals the true gradient $\boldsymbol{g}(\hat{\theta}) = [100, 1]^T$ now that noise-free ZO queries are accessible. However, the Euclidean norm of its covariance matrix $\text{Var}_{\boldsymbol{\Delta}^{\text{SPSA}}}[\hat{\boldsymbol{g}}^{\text{SPSA}}(\hat{\theta})]$ is in the order of 10^4. SFSA using $\boldsymbol{\Delta}^{\text{SFSA}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ also gives an unbiased gradient estimate, but the corresponding covariance matrix's magnitude is twice as large as that for SPSA for this example.

HARP (algorithm details to appear) draws $\boldsymbol{\Delta}^{\text{HARP}}$ from a distribution with $\boldsymbol{0}$-mean and a covariance $\boldsymbol{\Sigma} = \hat{\boldsymbol{H}}(\hat{\theta})^{-1}$, where $\hat{\boldsymbol{H}}(\hat{\theta})$ represents the estimate for the Hessian function $\boldsymbol{H}(\theta) \equiv \nabla^2 L(\theta)$ evaluated at $\hat{\theta}$. Let us defer the discussion on estimating $\hat{\boldsymbol{H}}(\cdot)$ and suppose $\boldsymbol{H}(\cdot)$ is perfectly recovered for the moment. Letting $\boldsymbol{\Delta}^{\text{HARP}} = [\boldsymbol{H}(\hat{\theta})]^{-1/2}\boldsymbol{\Delta}^{\text{SPSA}}$ and $\boldsymbol{m}(\boldsymbol{\Delta}^{\text{HARP}}) = \boldsymbol{H}(\hat{\theta})\boldsymbol{\Delta}^{\text{HARP}}$ is a valid choice. Note that any distributions satisfying C.2 to appear are welcomed, as they will *on average* impose 10% of the change magnitude in θ_2 onto that of θ_1. Predictably, the resulting estimator $\hat{\boldsymbol{g}}^{\text{HARP}}(\hat{\theta})$ is unbiased too. Nonetheless, the covariance matrix of $\hat{\boldsymbol{g}}^{\text{HARP}}(\hat{\theta})$ has a Euclidean norm of 2×10^2, which is smaller than 10^4 for SPSA/SFSA.

2.1.2. Correlation Across Directions

Another salient feature of $\boldsymbol{\Sigma} = \boldsymbol{I}$ is *zero* off-diagonal elements. It immediately follows that the perturbations along all the components of $\boldsymbol{\Delta}$ are *independent* with each other. Fortunately, feeding Hessian estimate into $\boldsymbol{\Sigma}$ innately helps in generating gradient estimate that is robust to various correlations between different components of the parameter.

Say, the loss function of interest becomes $L(\theta) = (100\theta_1^2 + \theta_2^2 + \theta_1\theta_2)$ with an extra cross-term, and $\hat{\theta}$ and c stay the same. Similarly, we have unbiased gradient estimator $\hat{\boldsymbol{g}}^{\text{SPSA}}(\hat{\theta})$ and $\hat{\boldsymbol{g}}^{\text{HARP}}(\hat{\theta})$. However, the covariance magnitude of the former is around 4×10^4 while the latter is around 8×10^2.

2.1.3. HARP

The aforementioned contrived example explains that we can reduce the variance of the gradient estimator $\hat{\boldsymbol{g}}(\hat{\theta})$ in (3) by also estimating Hessian matrix on the fly. The gain from the gradient-estimation side can be further propagated to a gain from the optimization perspective. With a more reliable (in terms of smaller variance) gradient estimator, we can expect and prove that HARP can reduce the variance of the parameter estimate $\hat{\theta}$ generated from (2).

2.2. Algorithm Development

Though 2.1.3 lays out the main motivation of HARP—reducing the variance of $\hat{\boldsymbol{g}}_k(\hat{\theta}_k)$ in *gradient estimation* and subsequently reducing the variance of $\hat{\theta}_k$ in *stochastic optimization*—there are a few gaps need to be handled before we carry out the intuitive idea of injecting Hessian estimate into perturbation covariance.

i) **The Hessian information has to be estimated in black-box problem**. Fortunately, there are prior work on estimating Hessian using ZO oracles. The deterministic-perturbation strategy in Fabian (1971) uses $O(d^2)$ ZO queries per iteration. The randomized-perturbation scheme in Spall (2000) uses *four* ZO queries per iteration, and Bhatnagar and Prashanth (2015) uses *three* ZO queries (but the Hessian estimate involves multiple contrived constants). Zhu (2021) extends these estimators with the aid of Stein's Identity. All the Hessian estimation $\hat{\boldsymbol{H}}(\hat{\theta})$ is computed *recursively* based on the gradient estimate $\hat{\boldsymbol{g}}(\hat{\theta})$ after collecting ZO queries.

To form a $\mathcal{F}_k$-measurable second-order approximation, HARP is comprised of two recursions, a natural form inspired by 2SPSA Spall (2000) is

$$\hat{\boldsymbol{H}}_k = \left\{\boldsymbol{m}_k(\widetilde{\boldsymbol{\Delta}}_k)[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)]^T + \boldsymbol{m}_k(\boldsymbol{\Delta}_k)[\boldsymbol{m}_k(\widetilde{\boldsymbol{\Delta}}_k)]^T\right\} \bar{\ell}_k/(4c_k\widetilde{c}_k)\,. \tag{4}$$

In (4), 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 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 $\bar{\ell}_k = \ell(\hat{\theta}_k + c_k\boldsymbol{\Delta}_k + \widetilde{c}_k\widetilde{\boldsymbol{\Delta}}_k, \omega_k^{+,+}) - \ell(\hat{\theta}_k + c_k\boldsymbol{\Delta}_k, \omega_k^{+}) - \ell(\hat{\theta}_k - c_k\boldsymbol{\Delta}_k + \widetilde{c}_k\widetilde{\boldsymbol{\Delta}}_k, \omega_k^{-,+}) + \ell(\hat{\theta}_k - c_k\boldsymbol{\Delta}_k, \omega_k^{+})$.

Inspired by Zhu (2021) (5) provides another estimator when $\boldsymbol{\Delta}_k$ is drawn from multivariate standard normal distribution and $\boldsymbol{m}_k(\cdot)$ is an identity mapping. For succinctness, write the noisy ZO queries $\ell_k^{\pm} \equiv \ell(\hat{\theta}_k \pm c\boldsymbol{\Delta}_k, \omega_k^{\pm})$ and $\ell_k \equiv \ell(\hat{\theta}_k, \omega_k)$, and write the observation noise $\varepsilon_k^{\pm} \equiv \ell_k^{\pm} - L(\hat{\theta}_k \pm c\boldsymbol{\Delta}_k)$ and $\varepsilon_k \equiv \ell_k - L(\hat{\theta}_k)$. Depending on the number of noisy ZO queries per iteration, we have several possible Hessian estimates Zhu (2021):

$$\hat{\boldsymbol{H}}_k = \begin{cases} c_k^{-2}\ell_k^{+}[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\boldsymbol{\Delta}_k^T - \boldsymbol{I}]\,, & \text{(5a)} \\ c_k^{-2}(\ell_k^{+} - \ell_k)[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\boldsymbol{\Delta}_k^T - \boldsymbol{I}]\,, & \text{(5b)} \\ (2c_k^2)^{-1}(\ell_k^{+} + \ell_k^{-})[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\boldsymbol{\Delta}_k^T - \boldsymbol{I}]\,, & \text{(5c)} \\ (2c_k^2)^{-1}(\ell_k^{+} + \ell_k^{-} - 2\ell_k)[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\boldsymbol{\Delta}_k^T - \boldsymbol{I}]\,. & \text{(5d)} \end{cases}$$

ii) **A valid covariance matrix has to be positive-definite, i.e., $\boldsymbol{\Sigma} \succ \boldsymbol{0}$.** However, due to the randomness ω in (1), there is no guarantee that the random matrix $\hat{\boldsymbol{H}}(\hat{\theta})$ can be a valid covariance matrix a.s. To ensure positive-definiteness of the covariance matrix, we impose a mapping $\boldsymbol{f}(\cdot)$ from $\mathbb{R}^{d\times d}$ to the set of symmetric and positive definite matrices at every iteration. The straightforward form for $\boldsymbol{f}(\cdot)$ is the Levenberg-Marquardt method: $\boldsymbol{f}(\boldsymbol{H}) = \boldsymbol{H} + \eta\boldsymbol{I}$ for η larger than the smallest eigenvalue of the input matrix $\boldsymbol{H}$. Another possible mapping is $\boldsymbol{f}(\boldsymbol{H}) = (\boldsymbol{H}^T\boldsymbol{H} + \eta\boldsymbol{I})^{1/2}$ for small $\eta > 0$ (say, 10^{-6}), which can be implemented in $O(d^2)$ FLOPs (Zhu et al., 2020), where the matrix square root is the unique positive-definite square root (implementable via `sqrtm` in MATLAB).

Overall, we propose to solve the stochastic optimization problem (1) using the recursion (2) to mitigate i)–ii), where the gradient estimate $\hat{\boldsymbol{g}}_k(\hat{\theta}_k)$ takes the form in (3) where the perturbation covariance $\boldsymbol{\Sigma}_k$ is computed as below:

$$\begin{cases} \boldsymbol{\Sigma}_k = \boldsymbol{f}_k(\overline{\boldsymbol{H}}_k)\,, & \text{(6a)} \\ \overline{\boldsymbol{H}}_{k+1} = \overline{\boldsymbol{H}}_k - w_k(\overline{\boldsymbol{H}}_k - \hat{\boldsymbol{H}}_k)\,, & \text{(6b)} \end{cases}$$

for $k \geq 0$ with weights $w_k > 0$. To handle i), (6b) produces estimate for $\boldsymbol{H}(\cdot)$ through a weighted average of the seen Hessian estimates $\hat{\boldsymbol{H}}_k$'s. The weights w_k governs the smoothing rate for Hessian estimate and $\sum_k w_k^2 c_k^{-4} < \infty$ is crucial to ensure the convergence of (1). For $\overline{\boldsymbol{H}}_0 = \boldsymbol{I}$, the early iteration of HARP resembles the randomized-perturbation strategy reviewed in Sect. 1.1. Responding to ii), (6a) imposes a positive-definite mapping on the recursive Hessian estimate $\overline{\boldsymbol{H}}_k$. Zhu et al. (2020, Algorithms 1–2) provides a way to achieve $O(d^2)$ FLOPs. Other forms of $\boldsymbol{f}_k(\cdot)$ satisfying conditions in Spall (2000) also work.

The implementation procedure for HARP using estimator (5) is summarized in Algorithm 1. This special case requires $\boldsymbol{\Delta}_k$ is drawn from multivariate standard normal distribution and $\boldsymbol{m}_k(\cdot)$ is an identity mapping. For other general distribution of $\boldsymbol{\Delta}_k$, additional hyper-parameter $\widetilde{c}_k$ and random perturbation $\widetilde{\boldsymbol{\Delta}}_k$ are required per (4).

Algorithm 1 A Special Case of Hessian-Amended Random Perturbation

```
Input: initialization $\hat{\theta}_0$, $\hat{\boldsymbol{H}}_0 = \boldsymbol{I}$, $\boldsymbol{\Sigma}_0 = \boldsymbol{I}$, and coefficients $a_k, c_k, w_k$ for $0 \leq k \leq K$.
set iteration index $k = 0$.
for $k = 0, 1, \cdots K$ do
    generate $\boldsymbol{\Delta}_k$ from a distribution with mean-$\boldsymbol{0}$ and a covariance of $\boldsymbol{\Sigma}_k^{-1}$ per (6a) and compute $\boldsymbol{m}_k(\boldsymbol{\Delta}_k) = \boldsymbol{\Sigma}_k \boldsymbol{\Delta}_k$.
    collect two ZO queries $\ell_k^{\pm}$ and estimate $\hat{\boldsymbol{g}}_k(\hat{\theta}_k)$ via (3).
    update parameter $\hat{\theta}_k$ via (2).
    estimate $\hat{\boldsymbol{H}}_k$ via (5) and update $\overline{\boldsymbol{H}}_k$ via (6b).        ▷ We may collect $\ell_k$ is if (5d) is used.
end
Output: terminal estimate $\hat{\theta}_K$.
```

2.3. 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 $\boldsymbol{x}$, $\|\boldsymbol{x}\|_\infty$ returns the infinity 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 Let $\mathcal{F}_k$ represent the history of the process (2) until the kth iteration:

$$\mathcal{F}_k = \{\hat{\theta}_0, \cdots, \hat{\theta}_k; \boldsymbol{\Delta}_0, \cdots, \boldsymbol{\Delta}_{k-1}; \omega_0^{\pm}, \cdots, \omega_{k-1}^{\pm}\}\,. \tag{7}$$

Note that the precise definition of $\mathcal{F}_k$ may vary, depending on the estimator forms (5) and the corresponding ZO queries. Furthermore, let $\mathbb{E}_k(\cdot)$ denote the conditional expectation $\mathbb{E}[\cdot \,|\, \mathcal{F}_k]$.

Miscellaneous notation $\mathbb{I}_E$ represents the indicator function of a logical expression E. In addition to $\boldsymbol{g}(\theta) \equiv \nabla L(\theta) \in \mathbb{R}^{d\times 1}$ and $\boldsymbol{H}(\theta) \equiv \nabla^2 L(\theta) \in \mathbb{R}^{d\times d}$, we also let $\nabla^3 L(\theta) \in \mathbb{R}^{1\times d^3}$ (as a row vector) represent the third-order derivative of $L(\cdot)$ evaluated at θ.

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

$$\hat{\theta}_k \text{ converges almost surely (strongly) to } \theta^*\,,\ (\hat{\theta}_k \xrightarrow{\text{a.s.}} \theta^*)\,, \tag{8}$$

or

$$\hat{\theta}_k \text{ converges to } \theta^* \text{ in mean-squared sense}\,,\ (\hat{\theta}_k \xrightarrow{\text{m.s.}} \theta^*)\,. \tag{9}$$

Robbins and Monro (1951) gave conditions for (8) whereas Blum (1954) for (9)[2]. We will prove (8) in Section 4.

3.2. Rate of Convergence

When either (8) or (9) 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:

$$\begin{cases} \max_{\mathcal{S}}\ \tau\,, \\ \text{s.t. random sequence } (\hat{\theta}_k - \theta^*) \text{ is } O_P(k^{-\tau/2})\,, \end{cases} \tag{10}$$

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 (18), and Billingsley (2008, Sect. 27), the constraint in (10) always takes the following form:

$$k^{\tau/2}(\hat{\theta}_k - \theta^*) \xrightarrow{\text{dist.}} \mathcal{N}(\mu, \boldsymbol{B}) \text{ for finite } \mu, \boldsymbol{B} \succ \boldsymbol{0}\,, \tag{11}$$

where $\xrightarrow{\text{dist.}}$ represents "convergence in distribution," and $(\tau, \mu, \boldsymbol{B})$ are functions of $\mathcal{S}$. When (11) holds and $[k^{\tau/2}(\hat{\theta}_k - \theta^*)]$ is uniformly integrable for any $\tau \le \tau^*$, (9) holds. The RMS error is asymptotic to $\lim_{k\to\infty}[\mathbb{E}(\|\hat{\theta}_k - \theta^*\|^2)]^{1/2} = k^{-\tau/2}[\|\mu\|^2 + \operatorname{tr}(\boldsymbol{B})]$.

[2] Neither (8) nor (9) implies the other (Billingsley, 2013, Chap. 5). Both (8) and (9) 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

$$\min_{\mathcal{S}} \left\{ k^{-\tau/2} \left[\|\mu\|^2 + \operatorname{tr}(\boldsymbol{B}) \right] \right\} , \tag{12}$$

as opposed to (10). 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 (12) reduces to (10). Sections 4.1–4.2 show that the solution to (10) is

$$\tau^* = \begin{cases} 2/3\,, & \text{for IID noise}\,, \quad (13) \\ 1\,, & \text{for CRN noise}\,, \quad (14) \end{cases}$$

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 (12) is in use. To achieve

$$\epsilon\text{-accurate estimate } \hat{\theta}_k \text{ s.t. } [\mathbb{E}(\|\hat{\theta}_k - \theta^*\|^2)]^{1/2} \leq \epsilon\,, \tag{15}$$

the the *average* desired number of iteration is

$$\left\{{}^{[\|\mu\|^2 + \operatorname{tr}(\boldsymbol{B})]}/_{\epsilon}\right\}^{2/\tau^*} = \begin{cases} O(\epsilon^{-3})\,, & \text{IID noise}\,, \\ O(\epsilon^{-2})\,, & \text{CRN noise}\,. \end{cases} \tag{16}$$

3.3. Other Forms of "Convergence" Rate

Nesterov and Spokoiny (2017, Sect. 4) uses the following notion

$$\epsilon\text{-accurate estimate } \hat{\theta}_k \text{ s.t. } \mathbb{E}[L(\hat{\theta}_k) - L(\theta^*)] \leq \epsilon\,, \tag{17}$$

as opposed to (15), and (17) is popular for analyzing ZO algorithms (Ghadimi and Lan, 2013). Let us offer a few remarks on the differences between (15) and (17). First of all, the resultant "convergence" rate under the notion (17) require *non*-decaying rate. Zhu (2020, 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 Zhu and Spall (2020), as $\hat{\theta}_k$ will be "random-walking" within a neighborhood of θ^* even for sufficiently large k Zhu and Spall (2018). Second, Nesterov and Spokoiny (2017); Ghadimi and Lan (2013) and all the subsequent work on ZO algorithms require *additive* CRN noise, and the corresponding analysis can *not* be generalized to the general CRN noise case discussed in Section 4.2, not to mention the IID noise case in Section 4.1. Third, the complexity result (Nesterov and Spokoiny, 2017, Eq. (59)) does not reveal the eigen-structure of $\boldsymbol{H}(\cdot)$ under certain smoothness assumption. On the contrary, $\boldsymbol{B}$ in (11) 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

[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 $p \to \infty$ (Fabian, 1971).

not least, $[\mathbb{E}(\|\hat{\theta}_k - \theta^*\|^2)]^{1/2} \leq \epsilon$ implies $\mathbb{E}[L(\hat{\theta}_k) - L(\theta^*)] \leq \epsilon'$, but generally not the other way around.

Overall, the notion (17) and the analysis in Nesterov and Spokoiny (2017); Ghadimi and Lan (2013) 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 (Zhu and Spall, 2020, 2016). Finally, it is advisable to use "concentration" and "concentration rate" Kushner and Yin (2003, Chaps. 7–8).

4. Convergence Result

4.1. IID Scenario

Although all estimators in (5) has bias term decreasing at the same rate, we use (5d) which has the smallest covariance matrix Zhu (2021). For clarity, this section analyzes $\hat{\boldsymbol{g}}_k$ in (3) (using two ZO queries) and $\hat{\boldsymbol{H}}_k$ in (5d) (using three ZO queries). Besides, we focus on the uncontrolled noise scenario where $\omega_k^{\pm}, \omega_k$ are i.i.d., frequently encountered in datastream for online learning. $O(k^{-1/3})$ in terms of root-mean-square (RMS) error $[\mathbb{E}(\|\hat{\theta}_k - \theta^*\|^2)]^{1/2}$ is the *fastest* rate possible for $a_k = O(k^{-1})$, $c_k = O(k^{-1/6})$, $\sum_k w_k^2 c_k^{-4} < \infty$, when $L(\cdot)$ is thrice continuously differentiable and is not quadratic. The difference between Algorithm 1 and prior work in Sect. 1.1 lies in the covariance of the resulting estimate.

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 $\varepsilon_k^{\pm} \equiv \ell(\hat{\theta}_k \pm c_k\boldsymbol{\Delta}_k, \omega_k^{\pm}) - L(\hat{\theta}_k \pm c\boldsymbol{\Delta}_k)$ and $\varepsilon_k \equiv \ell(\hat{\theta}_k, \omega_k) - L(\hat{\theta}_k)$.

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

Assumption A.2 (Perturbation) *Assume that the perturbation sequence $\{\boldsymbol{\Delta}_k\}$ are independently distributed with a mean of $\boldsymbol{0}$ 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 $\hat{\theta}_k$. 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 (Noise) *Assume $\mathbb{E}[\varepsilon_k^+ - \varepsilon_k^- \,|\, \hat{\theta}_k, \boldsymbol{\Delta}_k] \overset{\text{a.s.}}{=} 0$, and $\mathbb{E}[(\varepsilon_k^+ - \varepsilon_k^-)^2 \,|\, \hat{\theta}_k, \boldsymbol{\Delta}_k] \overset{\text{a.s.}}{\leq} D_3$ uniformly for all k.*

Remark 1 *For example, $\boldsymbol{m}_k(\boldsymbol{\Delta}_k) = \boldsymbol{\Sigma}_k\boldsymbol{\Delta}_k$ is a valid choice to enable $\mathbb{E}_k[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\boldsymbol{\Delta}_k^T] \overset{\text{a.s.}}{=} \boldsymbol{I}$ for all k. Alternatively, we may generate independent and identically distributed (i.i.d.) sequence $\{\boldsymbol{\delta}_k\}$*

from a zero-mean and unit-covariance distribution, and then let $\boldsymbol{\Delta}_k = \boldsymbol{\Sigma}_k^{-1/2}\boldsymbol{\delta}_k$ *and* $\boldsymbol{m}_k(\boldsymbol{\Delta}_k) = \boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\delta}_k$. *Also note that* $\mathbb{E}_k[\|\boldsymbol{\Delta}_k\|^6\|\boldsymbol{m}_k(\hat{\boldsymbol{\theta}}_k)\|^2]$ *being bounded implies all the lower-moments are bounded.*

A.1 and A.3 are assumptions imposed on the underlying loss function and the stochasticity, which we are not aware of in blackbox optimization, yet they are fundamental so that the resulting estimators make sense. On the contrary, C.2 is some verifiable conditions that the experimenter can control. Let us first discuss the asymptotically unbiasedness of both the gradient estimator and the Hessian estimator. Note that $\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)$ in (2) can be thought of as an estimate of $\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)$ and can be rewritten as:

$$\begin{aligned}&\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)\\ &\quad= \boldsymbol{g}(\hat{\boldsymbol{\theta}}_k) + \mathbb{E}_k[\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k) - \boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)] + \{\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k) - \mathbb{E}_k[\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)]\}\\ &\quad\equiv \boldsymbol{g}(\hat{\boldsymbol{\theta}}_k) + \boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k) + \boldsymbol{\xi}_k(\hat{\boldsymbol{\theta}}_k)\,,\end{aligned} \tag{18}$$

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.

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

$$\boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k) \overset{\text{a.s.}}{=} \frac{c_k^2}{12}\mathbb{E}_k\left\{[L^{(3)}(\overline{\boldsymbol{\theta}}_k^+) + L^{(3)}(\overline{\boldsymbol{\theta}}_k^-)](\boldsymbol{\Delta}_k\otimes\boldsymbol{\Delta}_k\otimes\boldsymbol{\Delta}_k)\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\right\}, \tag{19}$$

$$\begin{aligned}\boldsymbol{\xi}_k(\hat{\boldsymbol{\theta}}_k) &\overset{\text{a.s.}}{=} \frac{(\varepsilon_k^+ - \varepsilon_k^-)}{2c_k}\boldsymbol{m}_k(\boldsymbol{\Delta}_k) + [\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\boldsymbol{\Delta}_k^T - \boldsymbol{I}]\,\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)\\ &\qquad + \frac{c_k^2}{12}[L^{(3)}(\overline{\boldsymbol{\theta}}_k^+) + L^{(3)}(\overline{\boldsymbol{\theta}}_k^-)](\boldsymbol{\Delta}_k\otimes\boldsymbol{\Delta}_k\otimes\boldsymbol{\Delta}_k)\boldsymbol{m}_k(\boldsymbol{\Delta}_k) - \boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k),\end{aligned} \tag{20}$$

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* $O(c_k^{-2})$. *Besides, the Hessian estimator (5d) satisfies* $\mathbb{E}_k(\hat{\boldsymbol{H}}_k) \overset{\text{a.s.}}{=} \boldsymbol{H}(\hat{\boldsymbol{\theta}}_k) + O(c_k^2)$ *and* $\mathbb{E}_k(\|\hat{\boldsymbol{H}}_k\|^2) \overset{\text{a.s.}}{=} O(c_k^{-4})$.

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

With the gain sequence properly weighting the bias and variance in gradient and Hessian estimators, we can establish the almost surely convergence $\hat{\boldsymbol{\theta}}_k \xrightarrow{\text{a.s.}} \boldsymbol{\theta}^*$ as $k \to \infty$.

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* θ^* *is the unique minimizer such that* $\sup\{\|\theta\| : L(\theta) \leq L(\theta^*) + C_1\} < \infty$ *for every* $C_1 > 0$, $\inf_{\|\theta-\theta^*\|>C_2}[L(\theta) - L(\theta^*)] > 0$ *for every* $C_2 > 0$, $\inf_{\|\theta-\theta^*\|>C_3} \|\boldsymbol{g}(\theta)\| > 0$ *for every* $C_3 > 0$. *Moreover, there exists some* K*, such that for* $k \geq K$, $\boldsymbol{H}(\cdot)$ *satisfies* $\|\boldsymbol{H}(\theta)\|_\infty < D_4$ *for all* θ *in an open neighborhood of* $\hat{\theta}_k$ *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* $\hat{\theta}_k \overset{k\to\infty}{\longrightarrow} \theta^*$ *a.s. and* $\overline{\boldsymbol{H}}_k \overset{k\to\infty}{\longrightarrow} \boldsymbol{H}(\theta^*)$ *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. and and* $\overline{\boldsymbol{H}}_k \overset{k\to\infty}{\longrightarrow} \boldsymbol{H}(\theta^*)$ *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. Kushner and Clark (1978, 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* $\mathbb{E}[(\varepsilon_k^+ - \varepsilon_k^-)^{2+C_4} | \hat{\theta}_k, \boldsymbol{\Delta}_k] \overset{\text{a.s.}}{<} \infty$ *uniformly for all* k. *Finally,* $\boldsymbol{H}(\theta^*) \succ \boldsymbol{0}$.

Remark 2 *Note that under IID scenario for the observation noise, we have* $\mathbb{E}[(\varepsilon_k^+ - \varepsilon_k^-)^2 | \hat{\theta}_k, \boldsymbol{\Delta}_k] \to 2\mathrm{Var}(\ell(\theta^*, \omega))$ *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.*

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 $\alpha \in ({}^1/_2, 1]$, and $\gamma \in (0, \alpha - {}^1/_2)$. Granted, there are other forms for stepsizes (a_k, c_k). However, they do not necessarily provide improved rates (Sacks, 1958). Before stating Theorem 2, 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 $\boldsymbol{v}_k \equiv -ak^{-\gamma}\xi_k(\hat{\theta}_k)$.

Theorem 2 (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*

$$k^{\tau/2}(\hat{\theta}_k - \theta^*) \overset{\text{dist.}}{\longrightarrow} \mathcal{N}(\mu, \boldsymbol{B}) \ , \tag{21}$$

where $(\mu, \boldsymbol{B})$ *satisfies the linear system* (22) *and the Lyapunov equation* (23) *respectively:*

$$\begin{cases} (\boldsymbol{\Gamma} - {}^{\tau_+\boldsymbol{I}}/_{2})\mu = \boldsymbol{t}\,, & (22)\\ (\boldsymbol{\Gamma} - {}^{\tau_+\boldsymbol{I}}/_{2})\boldsymbol{B} + \boldsymbol{B}(\boldsymbol{\Gamma}^T - {}^{\tau_+\boldsymbol{I}}/_{2}) = \dfrac{a^2\mathrm{Var}[\ell(\theta^*, \omega)]}{2c^2}\boldsymbol{\Sigma}\,. & (23)\end{cases}$$

In (22–23)*,* $\boldsymbol{\Gamma} = \lim_{k\to\infty}\boldsymbol{\Gamma}_k = a\boldsymbol{H}(\theta^*)$*, the* $\mathrm{Var}[\ell(\theta^*, \omega)]$ *and* $\boldsymbol{\Sigma}$ *are defined in Remark* 2 *and A.*6 *respectively, and*

$$\boldsymbol{t} = \lim_{k\to\infty}\boldsymbol{t}_k = -\frac{ac^2}{6}\mathbb{I}_{\{\alpha=6\gamma\}}\mathbb{E}[L^{(3)}(\theta^*)\cdot(\boldsymbol{\Delta}\otimes\boldsymbol{\Delta}\otimes\boldsymbol{\Delta})\cdot\boldsymbol{m}(\boldsymbol{\Delta})]\,, \tag{24}$$

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

Remark 3 *Bartels and Stewart (1972) provides the explicit solution to* (23)*:*

$$\boldsymbol{B} = \frac{a^2\mathrm{Var}[\ell(\theta^*, \omega)]}{2c^2}\int_0^\infty e^{t({}^{\tau_+\boldsymbol{I}}/_{2}-\boldsymbol{\Gamma})}\boldsymbol{\Sigma}e^{t({}^{\tau_+\boldsymbol{I}}/_{2}-\boldsymbol{\Gamma}^T)}\mathrm{d}t\,. \tag{25}$$

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,*

$$\mathbb{E}_k\{\|\hat{\boldsymbol{g}}_k(\hat{\theta}_k)^2\|\} \overset{\text{a.s.}}{=} \mathbb{E}\|\mathrm{g}(\hat{\theta}_k, \omega_k)\|^2 + o(1) \overset{\text{a.s.}}{=} \int_{\omega\in\Omega}\|\mathrm{g}(\hat{\theta}_k, \omega)\|^2\mathrm{d}\mathbb{P}(\omega) + o(1)\,. \tag{26}$$

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 3, 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 $\boldsymbol{v}_k = -a\boldsymbol{\xi}_k(\hat{\theta}_k)$.

Theorem 3 (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*

$$k^{\alpha/2}(\hat{\theta}_k - \theta^*) \xrightarrow{\text{dist.}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{B})\,, \tag{27}$$

where $\boldsymbol{B}$ *satisfies*

$$(\boldsymbol{\Gamma} - {}^{\alpha_+\boldsymbol{I}}/_{2})\boldsymbol{B} + \boldsymbol{B}(\boldsymbol{\Gamma}^T - {}^{\alpha_+\boldsymbol{I}}/_{2}) = a^2\Sigma\,. \tag{28}$$

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

$$\Sigma_{i,j} = \mathbb{I}_{\{i=j\}}\int_{\omega\in\Omega}\|\mathrm{g}(\theta^*, \omega)\|^2\mathrm{d}\mathbb{P}(\omega) + \mathbb{I}_{\{i\neq j\}}\int_{\omega\in\Omega}[\mathrm{g}(\theta^*, \omega)]_i[\mathrm{g}(\theta^*, \omega)]_j\mathrm{d}\mathbb{P}(\omega)\,, \tag{29}$$

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

Recall that in IID scenario, (21) 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 (27) 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 (25) 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 $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$. Then $\boldsymbol{B}$ in (23) equals $\boldsymbol{P}\boldsymbol{M}\boldsymbol{P}^T$, where the (i, j)th elements of $\boldsymbol{M}$ is

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

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

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

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

$$\frac{a^2 \mathrm{Var}[\ell(\theta^*, \omega)]}{2c^2} \sum_{i=1}^{d} \frac{1}{2a - \tau_+/\lambda_i} . \tag{31}$$

Note that both (30) and (31) diverge when *any* one of the eigenvalues of $\boldsymbol{H}(\theta^*)$ is close to zero. Nonetheless, (31) is smaller than (30) when $\lambda_i \ll 1$ for some $1 \leq i \leq d$, under which circumstance the iteration complexity (16) 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 Spall (2000) 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 the attached code. The algorithm we compare against is SPSA (Spall, 1992), 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.

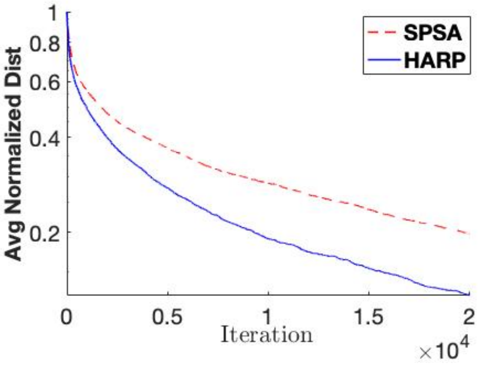

Figure 1: Performance of SPSA and HARP in terms of normalized distance $\|\hat{\theta}_k - \theta^*\| / \|\hat{\theta}_0 - \theta^*\|$ averaged across 25 independent replicates, and both algorithms use four ZO queries per iteration. The underlying loss function is the skew-quartic function with $d = 20$, 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 (Chen et al., 2017; Cheng et al., 2018) using zeroth-order optimization methods. We consider the constrained problem

$$\begin{cases} \min_{\theta} L(\theta) \equiv \underbrace{\kappa\|\theta\|_2^2}_{\equiv L_1(\theta)} + \underbrace{\frac{1}{I}\sum_{i=1}^{I} \operatorname{loss}(\zeta_i + \theta)}_{\equiv L_2(\theta)} , \\ \text{s.t. } (\zeta_i + \theta) \in [-0.5, 0.5]^d , \forall i, \end{cases} \tag{32}$$

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

$$\operatorname{loss}(\zeta) = \max_{i:1\le i\le C} \left\{ \operatorname{ps}(\zeta, i) - \max_{j\ne i:1\le j\le C} [\operatorname{ps}(\zeta, j)] \right\} , \tag{33}$$

where $\operatorname{ps}(\zeta, i)$ denotes the prediction score of the i-th class given the input ζ. The model $\operatorname{ps}(\cdot,\cdot)$here is trained using the structure specified in Carlini and Wagner (2017). Note that $\sum_{i=1}^{I} \operatorname{loss}(\zeta_i + \theta) = 0$ 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

$$\ell(\theta, \omega) = \kappa\|\theta\|_2^2 + \frac{1}{J}\sum_{j=1}^{J} \operatorname{loss}(\zeta_{i_j(\omega)} + \theta) , \tag{34}$$

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 (32) with $\kappa = {}^1/_{10}$. The I images arising in (32) are those *correctly* classified by the trained model. $d = 784$ for MNIST dataset. The algorithm we compare against is ZO-ADAMM (Chen et al., 2019). Both algorithms are initialized at $\hat{\theta}_0 = \mathbf{0}$, i.e., no attack is imposed initially. 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 the code attached.

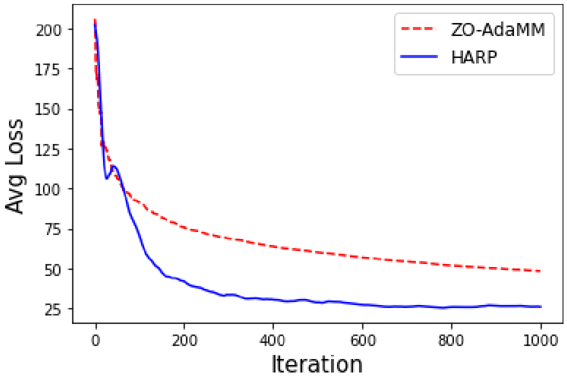

Figure 2: Average (expected) loss function evaluation of ZO-ADAMM and HARP in terms of loss after $K = 1000$ iterations averaged across 25 independent replicates. The query complexity aligns with the iteration complexity.

Figure 2 shows that HARP exhibits faster convergence rate than ZO-ADAMM given a fixed query-budget. Recall that 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 $I = 10$ images of the letter three, and its *noisy* query is evaluated using a batch-size of one. $\mathbb{E}[L_2(\hat{\theta}_K)]$ for ZO-ADAMM and HARP are approximately 9.8 and 0.2. A close-to-zero $L_2(\cdot)$ loss is equivalent to a close-to-one attack success rate.

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 $\mathbf{\Delta}_k$ and use a *deterministic* mapping $\boldsymbol{m}_k(\cdot)$. This work shows the benefits of using non-isotropic matrix as the covariance matrix for $\mathbf{\Delta}_k$ 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 $\mathbf{\Delta}_k$—at the cost of two additional ZO queries per iteration, see Algorithm 1.

Some potential future work includes (1) the generalization to root-finding problem where the Jacobian matrix is possibly asymmetric[5]; (2) the generalization to the one-measurement counterpart

[5] Note that in our discussion, the Hessian matrix for minimization problem is symmetric.

to (3) as Spall (1997) to further reduce query complexity; (3) the extended discussion on global convergence in line of Maryak and Chin (2001); (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

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.

Shalabh Bhatnagar and LA Prashanth. Simultaneous perturbation newton algorithms for simulation optimization. *Journal of Optimization Theory and Applications*, 164(2):621–643, 2015.

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

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

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

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

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.

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.

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.

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

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

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.

V Fabian. Stochastic approximation, optimization methods in statistics, 1971.

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

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

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

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

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

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

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

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

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

Jeffrey Larson, Matt Menickelly, and Stefan M Wild. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019.

Niru Maheswaranathan, Luke Metz, George Tucker, Dami Choi, and Jascha Sohl-Dickstein. Guided evolutionary strategies: escaping the curse of dimensionality in random search. 2018.

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

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.

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

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

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.

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

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.

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

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

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

James C Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. John Wiley & Sons, 2005.

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.

Jingyi Zhu. *Error Bounds and Applications for Stochastic Approximation with Non-Decaying Gain*. PhD thesis, Johns Hopkins University, 2020.

Jingyi Zhu. Hessian estimation via stein's identity in black-box problems. In *2nd Conference on Mathematical and Scientific Machine Learning*, 2021.

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

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.

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.

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 $\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)$.

$$\begin{aligned}
&\mathbb{E}_k[\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)] \\
&\overset{\text{a.s.}}{=} \mathbb{E}_k\left[\frac{L(\hat{\boldsymbol{\theta}}_k + c_k\boldsymbol{\Delta}_k) - L(\hat{\boldsymbol{\theta}}_k - c_k\boldsymbol{\Delta}_k)}{2c_k}\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\right] + \mathbb{E}_k\left[\frac{\boldsymbol{m}_k(\boldsymbol{\Delta}_k)}{2c_k}\mathbb{E}[(\varepsilon_k^+ - \varepsilon_k^-)|\,\hat{\boldsymbol{\theta}}_k, \boldsymbol{\Delta}_k]\right] \qquad (35) \\
&\overset{\text{a.s.}}{=} \mathbb{E}_k[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\boldsymbol{\Delta}_k^T]\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k) + \frac{c_k^2}{12}\mathbb{E}_k\left\{[L^{(3)}(\overline{\boldsymbol{\theta}}_k^+) + L^{(3)}(\overline{\boldsymbol{\theta}}_k^-)](\boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k)\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\right\} \qquad (36) \\
&\overset{\text{a.s.}}{=} \boldsymbol{g}(\hat{\boldsymbol{\theta}}_k) + \boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k)\,, \qquad (37)
\end{aligned}$$

where equation (35) uses Chung (2001, Thm. 9.1.3 on p. 315), equation (36) 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 (37) uses the expression (19) 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

$$\begin{aligned}
\mathbb{E}_k[\|\boldsymbol{\beta}_k(\hat{\boldsymbol{\theta}}_k)\|] &\overset{\text{a.s.}}{\leq} \frac{c_k^2}{6}\|L^{(3)}(\boldsymbol{\theta})\|_\infty \mathbb{E}_k[\|\boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k\|\|\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\|] && (38) \\
&\overset{\text{a.s.}}{=} \frac{c_k^2}{6} D_1 \mathbb{E}_k[\|\boldsymbol{\Delta}_k\|^3 \boldsymbol{m}_k(\boldsymbol{\Delta}_k)] && (39) \\
&\overset{\text{a.s.}}{\leq} \frac{c_k^2}{6} D_1 D_2\,, && (40)
\end{aligned}$$

where inequality (38) 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 (39) uses the independence between $\hat{\boldsymbol{\theta}}_k$ and $\boldsymbol{\Delta}_k$ assumed in A.2 and Lancaster and Farahat (1972), and inequality (40) uses A.2. The representation of $\boldsymbol{\xi}_k(\hat{\boldsymbol{\theta}}_k)$ in (20) follows directly from (18) and (19).

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

$$\begin{aligned}
\mathbb{E}_k\left\{\|\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)\|^2\right\} \overset{\text{a.s.}}{=} \mathbb{E}_k&\left\{\left\|\frac{L(\hat{\boldsymbol{\theta}}_k + c_k\boldsymbol{\Delta}_k) - L(\hat{\boldsymbol{\theta}}_k - c_k\boldsymbol{\Delta}_k)}{2c_k}\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\right\|^2\right\} && (41) \\
&+ \frac{1}{4c_k^2}\mathbb{E}_k[(\varepsilon_k^+ - \varepsilon_k^-)^2\|\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\|^2] && (42) \\
&+ \frac{1}{2c_k^2}\mathbb{E}_k\Big\{[L(\hat{\boldsymbol{\theta}}_k + c_k\boldsymbol{\Delta}_k) - L(\hat{\boldsymbol{\theta}}_k - c_k\boldsymbol{\Delta}_k)](\varepsilon_k^+ - \varepsilon_k^-)\|\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\|^2\Big\}. && (43)
\end{aligned}$$

The term on (42) becomes $O(c_k^{-2})$ because

$$\begin{aligned}
\mathbb{E}_k[(\varepsilon_k^+ - \varepsilon_k^-)^2\|\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\|^2] &\overset{\text{a.s.}}{=} \mathbb{E}_k\left[\|\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\|^2\mathbb{E}[(\varepsilon_k^+ - \varepsilon_k^-)^2|\,\hat{\boldsymbol{\theta}}_k, \boldsymbol{\Delta}_k]\right] && (44) \\
&\overset{\text{a.s.}}{=} D_3 \cdot \mathbb{E}_k[\|\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\|^2] && (45) \\
&\overset{\text{a.s.}}{\leq} D_3 D_2\,, && (46)
\end{aligned}$$

where inequality (44) uses Chung (2001, Thm. 9.1.3), inequality (45) uses A.3 and the independence between $\hat{\theta}_k$ and $\boldsymbol{\Delta}_k$, and inequality (46) uses A.2. The term on (43) becomes zero thanks to Chung (2001, Thm. 9.1.3) and $\mathbb{E}[\,\varepsilon_k^+ - \varepsilon_k^- \,|\, \hat{\theta}_k, \boldsymbol{\Delta}_k]$ assumed in A.3. The term on (41) can be bounded from above by $D_2\|\boldsymbol{g}(\hat{\theta}_k)\|^2 + O(c_k^2)$, as

$$
\begin{aligned}
&\mathbb{E}_k\left\{\left\|\frac{L(\hat{\theta}_k + c_k\boldsymbol{\Delta}_k) - L(\hat{\theta}_k - c_k\boldsymbol{\Delta}_k)}{2c_k}\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\right\|^2\right\} \\
&\overset{\text{a.s.}}{=} [\boldsymbol{g}(\hat{\theta}_k)]^T\mathbb{E}_k\{\boldsymbol{\Delta}_k[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)]^T\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\boldsymbol{\Delta}_k^T\}\boldsymbol{g}(\hat{\theta}_k) \\
&\quad + \frac{c_k^4}{144}\mathbb{E}_k\left\|[L^{(3)}(\overline{\theta}_k^+) + L^{(3)}(\overline{\theta}_k^-)](\boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k)\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\right\|^2 \\
&\quad + \frac{c_k^2}{6}[\boldsymbol{g}(\hat{\theta}_k)]^T\mathbb{E}_k\left\{\boldsymbol{\Delta}_k[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)]^T[L^{(3)}(\overline{\theta}_k^+) + L^{(3)}(\overline{\theta}_k^-)] \times (\boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k)\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\right\} \\
&\overset{\text{a.s.}}{=} O\left(\|\boldsymbol{g}(\hat{\theta}_k)\|^2\right) + O(c_k^2)\,,
\end{aligned}
\tag{47}
$$

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

If we adopt the Hessian estimator form in (4), we shall first consider the term $\widetilde{c}_k^{-1}\overline{\ell}_k\boldsymbol{m}_k(\widetilde{\boldsymbol{\Delta}}_k)$.

$$
\mathbb{E}(\widetilde{c}_k^{-1}\overline{\ell}_k\boldsymbol{m}_k(\widetilde{\boldsymbol{\Delta}}_k)\,\big|\,\hat{\theta}_k, \boldsymbol{\Delta}_k) \overset{\text{a.s.}}{=} \boldsymbol{g}(\hat{\theta}_k + c_k\boldsymbol{\Delta}_k) - \boldsymbol{g}(\hat{\theta}_k - c_k\boldsymbol{\Delta}_k) + O(c_k^3)\,,
\tag{48}
$$

where the $O(c_k^3)$ term in (48) 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

$$
\mathbb{E}\left(\frac{\overline{\ell}_k}{2c_k\widetilde{c}_k}\boldsymbol{m}_k(\widetilde{\boldsymbol{\Delta}}_k)[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)]^T\,\bigg|\,\mathcal{F}_k, \boldsymbol{\Delta}_k\right) \overset{\text{a.s.}}{=} \boldsymbol{H}(\hat{\theta}_k) + O(c_k^2)\,,
\tag{49}
$$

where (49) uses (48) 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 (49) 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 (48). 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 (4), in order to ensure a symmetric Hessian estimate. Given (49), 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 (4), the algorithmic form in Algorithm 1 and the corresponding analysis in Zhu et al. (2020).

If we adopt the Hessian estimator form in (5), we can conclude that $\mathbb{E}_k(\hat{\boldsymbol{H}}_k) \overset{\text{a.s.}}{=} \boldsymbol{H}(\hat{\theta}_k) + O(c_k)$ and $\mathbb{E}_k(\|\hat{\boldsymbol{H}}_k\|^2) \overset{\text{a.s.}}{=} O(c_k^{-4})$ using C.2 and following Zhu (2021, Proof of Lemma. 3).

■

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.

$$
\begin{aligned}
&\mathbb{E}_k[\hat{\boldsymbol{g}}_k(\hat{\theta}_k)] \\
&\overset{\text{a.s.}}{=} \mathbb{E}_k\left[\frac{L(\hat{\theta}_k + c_k\boldsymbol{\Delta}_k) - L(\hat{\theta}_k - c_k\boldsymbol{\Delta}_k)}{2c_k}\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\right] + \mathbb{E}_k\left[\frac{\boldsymbol{m}_k(\boldsymbol{\Delta}_k)}{2c_k}\mathbb{E}\left[(\varepsilon_k^+ - \varepsilon_k^-)\,\big|\,\hat{\theta}_k, \boldsymbol{\Delta}_k\right]\right] \\
&\overset{\text{a.s.}}{=} \mathbb{E}_k[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\boldsymbol{\Delta}_k^T]\boldsymbol{g}(\hat{\theta}_k) + \frac{c_k}{4}\mathbb{E}_k\left\{\boldsymbol{\Delta}_k^T[\boldsymbol{H}(\overline{\theta}_k^+) - \boldsymbol{H}(\overline{\theta}_k^-)]\boldsymbol{\Delta}_k\right\} \qquad (50) \\
&\overset{\text{a.s.}}{=} \boldsymbol{g}(\hat{\theta}_k) + \boldsymbol{\beta}_k(\hat{\theta}_k)\,,
\end{aligned}
$$

where (50) follows from the second-order Taylor expansion. Then $\boldsymbol{\beta}_k(\hat{\theta}_k)$ satisfies

$$\mathbb{E}_k\|\boldsymbol{\beta}_k(\hat{\theta}_k)\| \overset{\text{a.s.}}{\le} \frac{c_k}{4}\mathbb{E}_k\left\{\boldsymbol{\Delta}_k^T\left[O(1)\|2c_k\boldsymbol{\Delta}_k\|\right]\boldsymbol{\Delta}_k\right\} \tag{51}$$

$$\overset{\text{a.s.}}{=} O(c_k^2) \tag{52}$$

where the $O(1)$ in (51) represents the Lipschitz parameter of $\boldsymbol{H}(\cdot)$. Note that the explicit scaling constant in (52) is no longer available as (19).

■

Proof [Proof for Theorem 1]

Under assumptions A.4, and A.5, we known from Kushner and Clark (1978, Thm. 2.3.1 on p. 39) that Thm. 1 holds when the following two conditions hold:

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

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

Obviously, i) holds thanks to Lemma 1. Under assumption A.3, $\boldsymbol{\xi}_k(\hat{\theta}_k)$ defined in (20) is an $\mathcal{F}_k$-martingale. Using Kushner and Yin (2003, Eq. (4.1.4)), we have

$$\mathbb{P}\left\{\sup_{j\ge k}\|\sum_{i=k}^{j} a_i\boldsymbol{\xi}_i(\hat{\theta}_i)\| \ge \eta\right\} \le \eta^{-2}\mathbb{E}\|\sum_{i=k}^{\infty} a_i\boldsymbol{\xi}_i(\hat{\theta}_i)\|^2 \tag{53}$$

$$= \eta^{-2}\sum_{i=k}^{\infty} a_i^2\mathbb{E}\|\boldsymbol{\xi}_i(\hat{\theta}_i)\|^2\,, \tag{54}$$

where inequality (53) uses Markov's inequality, equality (54) uses $\mathbb{E}[]\boldsymbol{\xi}_i(\hat{\theta}_i)^T\boldsymbol{\xi}_j(\hat{\theta}_j)] = \mathbb{E}\{\boldsymbol{\xi}_i(\hat{\theta}_i)^T\mathbb{E}[\boldsymbol{\xi}_j(\hat{\theta}_j)\Big|\,\hat{\theta}_j]\} = 0$ for all $i < j$. Given A.5, ii) is also satisfied. The a.s. convergence from $\hat{\theta}_k$ to θ^* is arrived. ■

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

$$\mathbb{E}_k[L(\hat{\theta}_k)]$$

$$\overset{\text{a.s.}}{\le} \mathbb{E}_k\left\{L(\hat{\theta}_k) - a_k[\boldsymbol{g}(\hat{\theta}_k)]^T\hat{\boldsymbol{g}}_k(\hat{\theta}_k) + \frac{D_4a_k^2}{2}\|\hat{\boldsymbol{g}}_k(\hat{\theta}_k)\|^2\right\} \tag{55}$$

$$\overset{\text{a.s.}}{\le} L(\hat{\theta}_k) - a_k\|\boldsymbol{g}(\hat{\theta}_k)\|^2 + a_kO(c_k^2)\|\boldsymbol{g}(\hat{\theta}_k)\| + \frac{D_4a_k^2}{2}\left[O(c_k^2) + O(c_k^{-2}) + O(\|\boldsymbol{g}(\hat{\theta}_k)\|)^2\right] \tag{56}$$

$$\overset{\text{a.s.}}{=} L(\hat{\theta}_k) - a_k\|\boldsymbol{g}(\hat{\theta}_k)\|^2 + O(a_kc_k^2)\|\boldsymbol{g}(\hat{\theta}_k)\| + O(a_k^2c_k^2)) + O\left(\frac{a_k^2}{c_k^2}\right) + O(a_k^2)\|\boldsymbol{g}(\hat{\theta}_k)\|^2$$

$$\overset{\text{a.s.}}{\le} L(\hat{\theta}_k) - \frac{a_k}{2}\left(\|\boldsymbol{g}(\hat{\theta}_k)\| - O(c_k^2)\right)^2 + O\left(a_k^2c_k^2\right) + O(a_k^2c_k^{-2})\,, \text{for large } k \text{ s.t. } O(a_k) < 1/2\,, \tag{57}$$

where (55) uses A.4' and mean-value theorem, (56) uses Cauchy-Schwartz inequality and (41)–(43), and (57) uses A.5.

Therefore, for sufficiently large k, we have

$$\mathbb{E}_k[L(\hat{\theta}_k) - L(\theta^*)] \overset{\text{a.s.}}{\le} L(\hat{\theta}_k) - L(\theta^*) + O(a_k^2 c_k^2) + O(a_k^2 c_k^{-2}) \quad - \frac{a_k}{2}\left(\|\boldsymbol{g}(\hat{\theta}_k)\| - O(c_k^2)\right)^2 , \tag{58}$$

Under A.4' and A.5, Lai (1989, Thm. 1) ensures that the nonnegative random variable $[L(\hat{\theta}_k) - L(\theta^*)]$ converges to a *finite* random variable on a.s. Now that A.4' assumes $\sup\{\|\theta\| : L(\theta) \le L(\theta^*) + C_1\}$, the boundedness of $L(\hat{\theta}_k)$ a.s. implies the iterate boundedness $\sup_k \|\hat{\theta}_k\| < \infty$ a.s.

Next we show part ii). When (58) hold, Robbins and Siegmund (1971) ensures that $\lim_{k\to\infty}\sum_{i=1}^k a_i[\|\boldsymbol{g}(\hat{\theta}_i)\| - O(c_i^2)]^2 < \infty$ a.s. Together with A.5, we have $\|\boldsymbol{g}(\hat{\theta}_k)\| \to 0$ 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] The asymptotic normality result will be shown once the conditions (2.2.1), (2.2.2), and (2.2.3) of Fabian et al. (1968) hold.

We first show that Fabian et al. (1968, 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 $\|\beta_k(\hat{\theta}_k)\| = O(c_k^2) = O(k^{-2\gamma})$ 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 (19), A.1, and Thm. 1, we have

$$\beta_k \overset{k\to\infty}{\longrightarrow} \frac{1}{6} c_k^2 \mathbb{E}[L^{(3)}(\theta^*) \cdot (\boldsymbol{\Delta} \otimes \boldsymbol{\Delta} \otimes \boldsymbol{\Delta}) \cdot \boldsymbol{m}(\boldsymbol{\Delta})] \text{ a.s.} , \tag{59}$$

thanks to the dominated convergence theorem. Multiplying $-a(k+1)^{\tau/2} = -a(k+1)^{2\gamma}$ on both sides of (59) gives (24). Combined the cases for $\alpha < 6\gamma$ and $\alpha = 6\gamma$, we know that $\boldsymbol{t}_k$ converges to a finite vector for $\alpha \le 6\gamma$.

We then show that Fabian et al. (1968, Eq. (2.2.2)) hold. By definition (18), $\xi_k(\hat{\theta}_k)$ is a $\mathcal{F}_k$-measurable martingale sequence, and so is $\boldsymbol{v}_k$.

$$\begin{aligned}
\mathbb{E}_k(\boldsymbol{v}_k\boldsymbol{v}_k^T) &\overset{\text{a.s.}}{=} \frac{a^2}{(k+1)^{2\gamma}}\left(\mathbb{E}_k\{\hat{\boldsymbol{g}}_k(\hat{\theta}_k)[\hat{\boldsymbol{g}}_k(\hat{\theta}_k)]^T\} - \mathbb{E}_k[\hat{\boldsymbol{g}}_k(\hat{\theta}_k)]\{\mathbb{E}_k[\hat{\boldsymbol{g}}_k(\hat{\theta}_k)]\}^T\right) && (60)\\
&\overset{\text{a.s.}}{=} \frac{a^2}{c^2} c_k^2 \mathbb{E}_k\{\hat{\boldsymbol{g}}_k(\hat{\theta}_k)[\hat{\boldsymbol{g}}_k(\hat{\theta}_k)]^T\} + \frac{a^2}{c^2} c_k^2 [\boldsymbol{g}_k(\hat{\theta}_k) + \beta_k(\hat{\theta}_k)][\boldsymbol{g}_k(\hat{\theta}_k) + \beta_k(\hat{\theta}_k)]^T && (61)\\
&\overset{\text{a.s.}}{=} \frac{a^2}{c^2} \cdot \mathbb{E}_k\left[\left(\frac{\varepsilon_k^+ - \varepsilon_k^-}{2}\right)^2 \boldsymbol{m}_k(\boldsymbol{\Delta}_k)[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)]^T\right] + o(1)\\
&\overset{\text{a.s.}}{=} \frac{a^2}{4c^2}\mathbb{E}_k\left\{\boldsymbol{m}_k(\boldsymbol{\Delta}_k)[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)]^T \mathbb{E}[(\varepsilon_k^+ - \varepsilon_k^-)^2 \,|\, \hat{\theta}_k, \boldsymbol{\Delta}_k]\right\} + o(1)\\
&\overset{\text{a.s.}}{=} \frac{a^2}{c^2}\frac{2\text{Var}[\ell(\theta^*, \omega)]}{4}\mathbb{E}\{\boldsymbol{m}_k(\boldsymbol{\Delta}_k)[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)]^T\} + o(1) && (62)\\
&\overset{\text{a.s.}}{\longrightarrow} \frac{a^2\text{Var}[\ell(\theta^*, \omega)]}{2c^2}\boldsymbol{\Sigma}, \text{ as } k \to \infty , && (63)
\end{aligned}$$

where (60) follows from (18), the $o(1)$ term on (61) is due to A.2, (19), Lemma 1, and Theorem 1, both (62) and (63) are due to A.6 and Remark 2.

We finally show that either (2.2.3) or (2.2.4) in Fabian et al. (1968) 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

$$\begin{aligned}\lim_{k\to\infty} \mathbb{E}\left(\|\boldsymbol{v}_k\|^2 \mathbb{I}_{\{\|\boldsymbol{v}_k\|^2 \geq \eta k^\alpha\}}\right) &\leq \limsup_{k\to\infty} [\mathbb{P}(\|\boldsymbol{v}_k\|^2 \geq \eta k^\alpha)]^{\frac{C_5}{1+C_4}} \cdot [\mathbb{E}(\|\boldsymbol{v}_k\|^{2(1+C_5)})]^{\frac{1}{1+C_5}} \\ &\leq \limsup_{k\to\infty} \left(\frac{\mathbb{E}(\|\boldsymbol{v}_k\|^2)}{\eta k^\alpha}\right)^{\frac{C_5}{1+C_4}} \cdot [\mathbb{E}(\|\boldsymbol{v}_k\|^{2(1+C_5)})]^{\frac{1}{1+C_5}}\,, \end{aligned} \tag{64}$$

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

Now that all relevant conditions in Fabian et al. (1968) are met to ensure the asymptotic normality. ■

Proof [Proof of Lemma 2]

Under A.3',

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

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

$$\begin{aligned}\frac{1}{4c_k^2}\mathbb{E}[[\ell(\hat{\theta}_k + c_k\boldsymbol{\Delta}_k, \omega_k) - \ell(\hat{\theta}_k - c_k\boldsymbol{\Delta}_k, \omega_k)]^2 \Big| \hat{\theta}_k, \boldsymbol{\Delta}_k] &\overset{\text{a.s.}}{=} \mathbb{E}\left\{[\boldsymbol{\Delta}_k^T \mathrm{g}(\hat{\theta}_k, \omega_k)]^2 \Big| \hat{\theta}_k, \boldsymbol{\Delta}_k\right\} + O(c_k^4) \\ &\overset{\text{a.s.}}{=} \left[\boldsymbol{\Delta}_k^T \mathrm{g}(\hat{\theta}_k, \omega_k)\right]^2 + O(c_k^4)\,. \end{aligned} \tag{66}$$

Whence, (65) becomes

$$\mathbb{E}_k[\hat{\boldsymbol{g}}_k(\hat{\theta}_k)[\hat{\boldsymbol{g}}_k(\hat{\theta}_k)]^T] \overset{\text{a.s.}}{=} \mathbb{E}_k\left\{\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\boldsymbol{\Delta}_k^T \mathrm{g}(\hat{\theta}_k, \omega_k)[\mathrm{g}(\hat{\theta}_k, \omega_k)]^T \boldsymbol{\Delta}_k[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)]^T\right\} + o(1)\,. \tag{67}$$

Now that A.2 assumes independence between $\hat{\theta}_k$ and $\boldsymbol{\Delta}_k$, then the $(i,j)-$th component of (67) equals the following a.s.:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{p=1}^{d}\sum_{q=1}^{d} m_{k,i}\Delta_{k,p}\Delta_{k,q}m_{k,j}\right] \cdot \mathbb{E}_k\left(\mathrm{g}_{k,p}\mathrm{g}_{k,q}\right) + o(1) \\
&\overset{\text{a.s.}}{=} \left[\mathbb{I}_{\{i=j\}}\mathbb{I}_{\{p=q\}} + \mathbb{I}_{\{i\neq j\}}(\mathbb{I}_{\{p=i,q=j\}} + \mathbb{I}_{\{p=j,q=i\}})\right] \times \mathbb{E}_k\left(\mathrm{g}_{k,p}\mathrm{g}_{k,q}\right) + o(1) && (68) \\
&\overset{\text{a.s.}}{=} \begin{cases} \sum_{p=1}^{d} \mathbb{E}_k(\mathrm{g}_{k,p})^2 + o(1), & \text{if } i=j, \\ 2\mathbb{E}_k(\mathrm{g}_{k,i}\mathrm{g}_{k,j}) + o(1), & \text{if } i \neq j. \end{cases} && (69)
\end{aligned}$$

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 (68) uses $\mathbb{E}_k[\boldsymbol{m}_k(\boldsymbol{\Delta}_k)\boldsymbol{\Delta}_k^T] = \boldsymbol{I}$ in A.2. Taking the diagonal terms of (69) gives (26). ■