Gaussian Approximation and Multiplier Bootstrap for Stochastic Gradient Descent

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper, we establish the non-asymptotic validity of the multiplier bootstrap procedure for constructing the confidence sets using the Stochastic Gradient Descent (SGD) algorithm. Under appropriate regularity conditions, our approach avoids the need to approximate the limiting covariance of Polyak-Ruppert SGD iterates, which allows us to derive approximation rates in convex distance of order up to $1/\sqrt{n}$. Notably, this rate can be faster than the one that can be proven in the Polyak-Juditsky central limit theorem. To our knowledge, this provides the first fully non-asymptotic bound on the accuracy of bootstrap approximations in SGD algorithms. Our analysis builds on the Gaussian approximation results for nonlinear statistics of independent random variables.

1 Introduction

Stochastic Gradient Descent (SGD) is a widely used first-order optimization method that is well suited for large data sets and online learning. The algorithm has attracted significant attention; see [34, 31, 27, 26, 8]. SGD aims to solve the optimization problem:

$$f(\theta) \to \min_{\theta \in \mathbb{R}^d} , \qquad \nabla f(\theta) = \mathbb{E}_{\xi \sim \mathbb{P}_\xi}[F(\theta, \xi)] , \tag{1}$$

where ξ is a random variable defined on a measurable space $(\mathsf{Z}, \mathcal{Z})$. Instead of the exact gradient $\nabla f(\theta)$, the algorithm accesses only unbiased stochastic estimates $F(\theta, \xi)$.

Throughout this work, we focus on the case of strongly convex objective functions and denote by $\theta^\star$ the unique minimizer of (1). The iterates θ_k, $k \in \mathbb{N}$, generated by SGD follow the recursive update:

$$\theta_{k+1} = \theta_k - \alpha_{k+1} F(\theta_k, \xi_{k+1}) , \quad \theta_0 \in \mathbb{R}^d , \tag{2}$$

where $\{\alpha_k\}_{k \in \mathbb{N}}$ is a sequence of step sizes (or learning rates), which may be either diminishing or constant, and $\{\xi_k\}_{k \in \mathbb{N}}$ is an i.i.d. sequence sampled from $\mathbb{P}_\xi$. Theoretical properties of SGD, particularly in the convex and strongly convex settings, have been extensively studied; see, e.g., [28, 26, 8, 23]. Many optimization algorithms build upon the recurrence (2) to accelerate the convergence of the sequence θ_k to $\theta^\star$. Notable examples include momentum acceleration [32], variance reduction techniques [12, 39], and averaging methods. In this work, we focus on Polyak-Ruppert averaging, originally proposed in [36] and [31], which improves convergence by averaging the SGD iterates (2). Specifically, the estimator is defined as

$$\bar{\theta}_n = \frac{1}{n} \sum_{i=0}^{n-1} \theta_i , \quad n \in \mathbb{N} . \tag{3}$$

It has been established (see [31, Theorem 3]) that under appropriate conditions on the objective function f, the noisy gradient estimates F, and the step sizes α_k, the sequence of averaged iterates

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

$\{\bar{\theta}_n\}_{n\in\mathbb{N}}$ is asymptotically normal:

$$\sqrt{n}(\bar{\theta}_n - \theta^\star) \xrightarrow{d} \mathcal{N}(0, \Sigma_\infty) \;, \tag{4}$$

where $\xrightarrow{d}$ denotes convergence in distribution, and $\mathcal{N}(0, \Sigma_\infty)$ is a zero-mean Gaussian distribution with covariance matrix Σ_∞, defined later in Section 2.2. This result raises two key questions:

(i) what is the rate of convergence in (4)?

(ii) how can (4) be leveraged to construct confidence sets for $\theta^\star$, given that Σ_∞ is unknown in practice?

In our paper we aim to answer both questions. To quantify convergence rates in (4), we employ convex distance, which is defined for random vectors $X, Y \in \mathbb{R}^d$ as

$$\mathsf{d}_{\mathsf{C}}(X, Y) = \sup_{B \in \mathsf{C}(\mathbb{R}^d)} \left|\mathbb{P}\big(X \in B\big) - \mathbb{P}(Y \in B\big)\right| \;,$$

where $\mathsf{C}(\mathbb{R}^d)$ denotes the collection of convex subsets of $\mathbb{R}^d$. The authors of [42] derive Berry-Esseen-type bounds for $\mathsf{d}_{\mathsf{C}}(\sqrt{n}\Sigma_n^{-1/2}(\bar{\theta}_n - \theta^\star), \mathcal{N}(0, \mathrm{I}_d))$, where Σ_n is the covariance matrix of the linearized counterpart of (2), see precise definitions (13) . We complement this result with the rates of convergence in (4). Interestingly, we also provide the lower bounds on the convex distance $\mathsf{d}_{\mathsf{C}}(\sqrt{n}(\bar{\theta}_n - \theta^\star), \mathcal{N}(0, \Sigma_\infty))$, which indicates, that for some choice of step sizes α_k in (2), the normal approximation by $\mathcal{N}(0, \Sigma_\infty)$ is less accurate, compared to normal approximation with other covariance matrix, in particular, with Σ_n. This effect has been previously observed in the bootstrap method for i.i.d. observations without the context of gradients methods, see [41, Theorem 3.11].

One of the popular approaches for solving (ii) is based on the *plug-in* methods [11, 9], which aim to construct an estimator $\hat{\Sigma}_n$ of Σ_∞ directly. These methods often provide a non-asymptotic bounds on the closeness $\hat{\Sigma}_n$ is to Σ_∞, often in terms of $\mathbb{E}[\|\hat{\Sigma}_n - \Sigma_\infty\|]$. At the same time, the analysis of this methods typically bypass the item (i) and the issues related with the rate of convergence in (4). In our paper we suggest, to the best of our knowledge, the first fully non-asymptotic analysis of procedure for constructing the confidence intervals, based on the bootstrap approach [15, 16], which avoids the direct approximation of Σ_∞, moreover, theoretical analysis of the underlying procedure together with the results on normal approximation with $\mathcal{N}(0, \Sigma_\infty)$ from (i) shows that the same approximation rate can not be achieved by the plug-in methods, at least for some range of step sizes α_k in (2). Our key contributions are as follows:

- We establish the non-asymptotic validity of the multiplier bootstrap procedure introduced in [16]. Under appropriate regularity conditions, our bounds imply that the quantiles of the exact distribution of $\sqrt{n}(\bar{\theta}_n - \theta^\star)$ can be approximated, up to logarithmic factors, at a rate of $n^{-\gamma/2}$ for step sizes of the form $\alpha_k = c_0/(k+k_0)^\gamma, \gamma \in (1/2, 1)$. To our knowledge, this provides the first fully non-asymptotic bound on the accuracy of bootstrap approximations in SGD algorithms. Notably, this rate can be faster than the one that we can prove in (4). Our rates improve upon recent works [38, 46], which addressed the convergence rate in similar procedures for the LSA algorithm.
- Our analysis of the multiplier bootstrap procedure reveals an interesting property: unlike plug-in estimators, the validity of the bootstrap method does not directly depend on approximating $\sqrt{n}(\bar{\theta}_n - \theta^\star)$ by $\mathcal{N}(0, \Sigma_\infty)$. Instead, it requires approximating $\mathcal{N}(0, \Sigma_n)$ for some matrix Σ_n. The structure of Σ_n and its associated convergence rates play a central role in our present analysis, both for convergence rate in (4) and non-asymptotic bootstrap validity. Precise definitions are provided in Section 2.2.
- We analyze the Polyak-Ruppert averaged SGD iterates (3) for strongly convex minimization problems and establish Gaussian approximation rates in (4) in terms of the convex distance. Specifically, we show that the approximation rate $\mathsf{d}_{\mathsf{C}}(\sqrt{n}(\bar{\theta}_n - \theta^\star), \mathcal{N}(0, \Sigma_\infty))$ is of order $n^{-1/4}$ when using the step size $\alpha_k = c_0/(k+k_0)^{3/4}$ with a suitably chosen α_0. Our result is based on the techniques of [42] and [46]. We also provide the lower bound, which indicate that our rate of normal approximation with $\mathcal{N}(0, \Sigma_\infty)$ is tight in the regime $\alpha_k = c_0/(k+k_0)^\gamma$ with $\gamma \geq 3/4$.

Notations. Throughout this paper, we use the following notations. For a matrix $A \in \mathbb{R}^{d\times d}$ and a vector $x \in \mathbb{R}^d$, we denote by $\|A\|$ and $\|x\|$ their spectral norm and Euclidean norm, respectively. We also write $\|A\|_{\mathrm{F}}$ for the Frobenius norm of matrix A. Given a function $f : \mathbb{R}^d \to \mathbb{R}$, we write $\nabla f(\theta)$ and $\nabla^2 f(\theta)$ for its gradient and Hessian at a point θ. Additionally, we use the standard abbreviations "i.i.d." for "independent and identically distributed" and "w.r.t." for "with respect to".

Literature review. Asymptotic properties of the SGD algorithm, including the asymptotic normality of the estimator $\bar{\theta}_n$ and its almost sure convergence, have been extensively studied for smooth and strongly convex minimization problems [31, 22, 7]. Optimal mean-squared error (MSE) bounds for $\theta_n - \theta^\star$ and $\bar{\theta}_n - \theta^\star$ were established in [27] for smooth and strongly convex objectives, and later refined in [26]. The case of constant-step size SGD for strongly convex problems has been analyzed in depth in [14]. High-probability bounds for SGD iterates were obtained in [33] and later extended in [20]. Both works address non-smooth and strongly convex minimization problems.

It is important to note that the results discussed above do not directly imply convergence rates for $\sqrt{n}(\bar{\theta}_n - \theta^\star)$ to $\mathcal{N}(0, \Sigma_\infty)$ in terms of $\mathsf{d}_{\mathsf{C}}(\cdot, \cdot)$ or the Kantorovich–Wasserstein distance. Among the relevant contributions in this direction, we highlight recent works [44, 38, 46], which provide quantitative bounds on the convergence rate in (4) for iterates of the temporal difference learning algorithm and general linear stochastic approximation (LSA) schemes. However, these algorithms do not necessarily correspond to SGD with a quadratic objective f, as the system matrix in LSA is not necessarily symmetric. Non-asymptotic convergence rates of order $1/\sqrt{n}$ in a smooth Wasserstein distance were established in [2]. Recent paper [1] provide Berry-Essen bounds for last iterate of SGD for high-dimensional linear regression of order up to $n^{-1/4}$.

Bootstrap methods for i.i.d. observations were first introduced in [15]. In the context of SGD methods, [16] proposed the multiplier bootstrap approach for constructing confidence intervals for $\theta^\star$ and established its asymptotic validity. The same algorithm, with non-asymptotic guarantees, was analyzed in [38] for the LSA algorithm, obtaining rate $n^{-1/4}$ when approximating quantiles of the exact distribution of $\sqrt{n}(\bar{\theta}_n - \theta^\star)$.

Popular group of methods for constructing confidence sets for $\theta^\star$ is based on estimating the asymptotic covariance matrix Σ_∞. Plug-in and batch-mean estimators for Σ_∞ attracted lot of attention, see [11, 9, 10], especially in the setting when the stochastic estimates of Hessian are available. The latter two papers focused on learning with contextual bandits. Estimates for Σ_∞ based on batch-mean method and its online modification were considered in [11] and [49]. The authors in [25] considered the asymptotic validity of the plug-in estimator for Σ_∞ in the local SGD setting. [47] refined the validity guarantees for both the multiplier bootstrap and batch-mean estimates of Σ_∞ for nonconvex problems. However, these papers typically provide recovery rates Σ_∞, but only show asymptotic validity of the proposed confidence intervals. A notable exception is the recent paper [46], where the temporal difference (TD) learning algorithm was studied. The authors of [46] provided purely non-asymptotic analysis of their procedure, obtaining the approximation rate $n^{-1/3}$ for quantiles of $\sqrt{n}(\bar{\theta}_n - \theta^\star)$.

2 Main results

This section establishes the nonasymptotic validity of the multiplier bootstrap method proposed in [16]. We focus on smooth and strongly convex minimization problems, following the framework established in [26], [2] and [42]. The underlying procedure is based on perturbing the trajectory (2). We restate the procedure for the sake of clarity. Let $\mathcal{W}^{n-1} = \{w_\ell\}_{1 \le \ell \le n-1}$ be i.i.d. random variables with distribution $\mathbb{P}_w$, each with mean $\mathbb{E}[w_1] = 1$ and variance $\mathrm{Var}[w_1] = 1$. Assume $\mathcal{W}^{n-1}$ is independent of $\Xi^{n-1} = \{\xi_\ell\}_{1 \le \ell \le n-1}$. We then use $\mathcal{W}^{n-1}$ to construct randomly perturbed SGD trajectories, following the same recursive structure as the primary sequence

$$
\begin{aligned}
\theta_k^{\mathsf{b}} &= \theta_{k-1}^{\mathsf{b}} - \alpha_k w_k \{\nabla f(\theta_{k-1}^{\mathsf{b}}) + g(\theta_{k-1}^{\mathsf{b}}, \xi_k) + \eta(\xi_k)\} \;, \quad k \ge 1 \;, \quad \theta_0^{\mathsf{b}} = \theta_0 \;, \\
\bar{\theta}_n^{\mathsf{b}} &= n^{-1} \sum_{k=0}^{n-1} \theta_k^{\mathsf{b}} \;, \quad n \ge 1 \;.
\end{aligned}
\tag{5}
$$

Note that, when generating different weights w_k, we can draw samples from the conditional distribution of $\bar{\theta}_n^{\mathsf{b}}$ given the data Ξ^{n-1}. We further denote $\mathbb{P}^{\mathsf{b}} = \mathbb{P}(\cdot \mid \Xi^{n-1})$ and $\mathbb{E}^{\mathsf{b}} = \mathbb{E}(\cdot \mid \Xi^{n-1})$.

The core principle behind the bootstrap procedure (5) is that the "bootstrap world" probabilities $\mathbb{P}^{\mathsf{b}}\big(\sqrt{n}(\bar{\theta}_n^{\mathsf{b}} - \bar{\theta}_n) \in B\big)$ are close to $\mathbb{P}\big(\sqrt{n}(\bar{\theta}_n - \theta^\star) \in B\big)$ for $B \in \mathsf{C}(\mathbb{R}^d)$. More formally, we say that the procedure (5) is asymptotically valid if

$$
\sup_{B \in \mathsf{C}(\mathbb{R}^d)} \left| \mathbb{P}^{\mathsf{b}}\big(\sqrt{n}(\bar{\theta}_n^{\mathsf{b}} - \bar{\theta}_n) \in B\big) - \mathbb{P}\big(\sqrt{n}(\bar{\theta}_n - \theta^\star) \in B\big) \right| \tag{6}
$$

converges to 0 in $\mathbb{P}$-probability as $n \to \infty$. This result was studied in [16] under assumptions close to the original paper [31]. While an analytical expression for $\mathbb{P}^{\mathsf{b}}(\sqrt{n}(\bar{\theta}_n^{\mathsf{b}} - \bar{\theta}_n) \in B)$ is unavailable, it can be approximated via Monte Carlo simulations by generating M perturbed trajectories according to (5). Standard arguments (see, e.g., [40, Section 5.1]) suggest that the accuracy of this Monte Carlo approximation scales as $\mathcal{O}(M^{-1/2})$ when generating M parallel perturbed trajectories in (5).

Assumptions. We impose the following regularity conditions on the objective function f:

A1. *The function f is two times continuously differentiable and L_1-smooth on $\mathbb{R}^d$, i.e., there is a constant $L_1 > 0$, such that for any $\theta, \theta' \in \mathbb{R}^d$,*

$$\|\nabla f(\theta) - \nabla f(\theta')\| \le L_1 \|\theta - \theta'\| \; .$$

Moreover, we assume that f is μ-strongly convex on $\mathbb{R}^d$, that is, there exists a constant $\mu > 0$, such that for any $\theta, \theta' \in \mathbb{R}^d$, it holds that

$$(\mu/2)\|\theta - \theta'\|^2 \le f(\theta) - f(\theta') - \langle \nabla f(\theta'), \theta - \theta' \rangle \; .$$

A1 implies the following two-sided bound on the Hessian $\nabla^2 f(\theta)$, $\mu \mathrm{I}_d \preceq \nabla^2 f(\theta) \preceq L_1 \mathrm{I}_d$ for all $\theta \in \mathbb{R}^d$. We now formalize the assumptions on $F(\theta, \xi)$. Namely, we rewrite $F(\theta, \xi)$ as

$$F(\theta_{k-1}, \xi_k) = \nabla f(\theta_{k-1}) + \zeta_k ,$$

where $\{\zeta_k\}_{k \in \mathbb{N}}$ is a sequence of d-dimensional random vectors. Then the SGD recursion takes form

$$\theta_k = \theta_{k-1} - \alpha_k (\nabla f(\theta_{k-1}) + \zeta_k) \; , \quad \theta_0 \in \mathbb{R}^d \; . \tag{7}$$

We impose the following assumption on the noise sequence ζ_k:

A2. *For each $k \ge 1$, ζ_k admits the decomposition $\zeta_k = \eta(\xi_k) + g(\theta_{k-1}, \xi_k)$, where*

(i) $\{\xi_k\}_{k=1}^{n-1}$ is a sequence of i.i.d. random variables on $(\mathsf{Z}, \mathcal{Z})$ with distribution $\mathbb{P}_\xi$, $\eta : \mathsf{Z} \to \mathbb{R}^d$ is a function such that $\mathbb{E}[\eta(\xi_1)] = 0$ and $\mathbb{E}[\eta(\xi_1)\eta(\xi_1)^\top] = \Sigma_\xi$. Moreover, $\lambda_{\min}(\Sigma_\xi) > 0$.

(ii) The function $g : \mathbb{R}^d \times \mathsf{Z} \to \mathbb{R}^d$ satisfies $\mathbb{E}[g(\theta, \xi_1)] = 0$ for any $\theta \in \mathbb{R}^d$. Moreover, there exists $L_2 > 0$ such that for any $\theta, \theta' \in \mathbb{R}^d$, it holds that

$$\|g(\theta, \xi) - g(\theta', \xi)\| \le L_2 \|\theta - \theta'\| \quad \textit{and} \quad g(\theta^\star, z) = 0 \;\; \textit{for all } z \in \mathsf{Z} \; . \tag{8}$$

(iii) There exist $C_{1,\xi}, C_{2,\xi} > 0$ such that $\mathbb{P}_\xi$-almost surely that $\|\eta(\xi)\| \le C_{1,\xi}$ and $\sup_\theta \|g(\theta, \xi)\| \le C_{2,\xi}$.

As an example of a sequence ζ_k satisfying conditions (i) and (ii) from A2, consider the case when the oracle function $F(\theta, \xi)$ satisfies:

1. $\mathbb{E}[F(\theta, \xi)] = \nabla f(\theta)$ for all $\theta \in \mathbb{R}^d$;
2. $\|F(\theta, \xi) - F(\theta', \xi)\| \le L\|\theta - \theta'\|$ for all $\xi \in \mathsf{Z}$, and $\sup_\theta |F(\theta, \xi) - F(\theta^\star, \xi)| \le c_\xi$ for some $c_\xi > 0$.

In this case, (i) and (ii) from A2 holds with $\eta(\xi) = F(\theta^\star, \xi)$ and $g(\theta, \xi) = F(\theta, \xi) - F(\theta^\star, \xi)$. Additionally, note that the identity (8) can be relaxed when one considers only last iterate bounds, such as $\mathbb{E}[\|\theta_k - \theta^\star\|^2]$, see [26]. Item (ii) from A2 is often imposed when considering averaged iterates, see [26, Assumption H2'], and [14, 42].

The assumption (iii) from A2 is crucial to prove high-order moment bounds (20), see Lemma 15. In our proof, we closely follow the argument presented in [20, Theorem 4.1], which requires that the noise variables ζ_k be almost sure to be bounded. This setting can be generalized to the case where ζ_k is sub-Gaussian conditioned on $\mathcal{F}_{k-1}$ with variance proxy which is uniformly bounded by a constant factor, that is, there is a constant M, such that $\mathbb{E}[\exp\{\|F(\theta, \xi_1)\|^2/M^2\}] \le 2$ for any $\theta \in \mathbb{R}^d$. This assumption is widely considered in the literature; see [27, 21], and the remarks in [20]. However, when $\zeta_k = g(\theta_{k-1}, \xi_k) + \eta(\xi_k)$ and g is only Lipschitz w.r.t. θ, its moments will naturally scale with $\|\theta_{k-1} - \theta^\star\|$, thus the sub-Gaussian bound with M not depending upon θ is unlikely to hold. Other authors who considered bounds of type (20), e.g. [33], made stronger assumption that $\sup_{\theta \in \mathbb{R}^d} \|F(\theta, \xi)\|$ is a.s. bounded. Another popular direction is to consider schemes for gradient clipping; see e.g. [37]. Unfortunately, employing such schemes change the key representation (12) that we use later in the proof of the main result (see Theorem 1). We leave further studies of clipped gradient schemes for future work. We further impose condition on the Hessian matrix $\nabla^2 f(\theta)$ at $\theta^\star$:

A3. *There exist $L_3, \beta > 0$ such that for all θ with $\|\theta - \theta^\star\| \le \beta$, it holds*

$$\|\nabla^2 f(\theta) - \nabla^2 f(\theta^\star)\| \le L_3 \|\theta - \theta^\star\| \; .$$

A3 ensures that the Hessian of f is Lipschitz continuous in a neighborhood of θ^*. Similar assumptions have been previously considered in [42] and [2], as well as in other works on first-order optimization methods, see, e.g. [24]. Several studies on the non-asymptotic analysis of SGD impose stronger smoothness assumptions, such as bounded derivatives of f up to order four, see [14]. We additionally assume an almost sure co-coercivity of the stochastic gradient:

A 4. *The stochastic gradient $F(\theta, \xi) := \nabla f(\theta) + g(\theta, \xi) + \eta(\xi)$ is almost surely L_4-co-coercive, that is, for any $\theta, \theta' \in \mathbb{R}^d$, it holds $\mathbb{P}_\xi$-almost surely that*

$$L_4 \langle F(\theta, \xi) - F(\theta', \xi), \theta - \theta' \rangle \geq \|F(\theta, \xi) - F(\theta', \xi)\|^2 .$$

In particular, A4 holds (see e.g. [48]), when there is a function $v(\theta, \xi)$, such that $F(\theta, \xi) = \nabla_\theta v(\theta, \xi)$, where $v(\theta, \xi)$ is convex $\mathbb{P}_\xi$-a.s. and L_4-smooth. Co-coercivity is stronger than just requiring $F(\theta, \xi)$ to be monotone. We also impose an assumption on the bootstrap weights W_i used in the algorithm:

A5. *There exist constants $0 < W_{\min} < W_{\max} < +\infty$, such that $W_{\min} \leq w_1 \leq W_{\max}$ a.s.*

The original paper [16] also considered positive bootstrap weights w_i. We have to impose boundedness of w_i due to our high-probability bound on Lemma 15. A particular example of a distribution satisfying A5 is provided in Appendix E.1. We also consider the following bound for step sizes α_k and sample size n:

A 6. *Let $\alpha_k = c_0\{k_0 + k\}^{-\gamma}$, where $\gamma \in (1/2, 1)$, an c_0 satisfies $c_0 W_{\max} \max(2L_4, \mu) \leq 1$ and $k_0 \geq (\frac{2\gamma}{\mu c_0 W_{\min}})^{1/(1-\gamma)}$.*

A 7. *Number of observations n satisfies $n \geq \mathrm{e}^3$ and $\frac{n}{\log(2dn)} \geq \max(1, \frac{(20 C_{Q,\xi} C_\Sigma^2)^2}{9})$, where the constants $C_{Q,\xi}$ and C_Σ are defined in (60) and (27), respectively.*

The particular bound on k_0 in A6 appears due to the high-order moment bounds (see Lemma 15 in appendix). We note that it is possible to remove the co-coercivity assumption A4, but at the price of slightly stronger constraints on c_0 above. We discuss the bound on the number of observations imposed in A7 later in the proof of Theorem 1.

2.1 Non-asymptotic multiplier bootstrap validity

Theorem 1. *Assume A1 - A7. Then with $\mathbb{P}$ - probability at least $1 - 2/n$, it holds*

$$\sup_{B \in \mathsf{C}(\mathbb{R}^d)} |\mathbb{P}^{\mathsf{b}}(\sqrt{n}(\bar{\theta}_n^{\mathsf{b}} - \bar{\theta}_n) \in B) - \mathbb{P}(\sqrt{n}(\bar{\theta}_n - \theta^\star) \in B)| \leq \frac{\mathrm{C}_1 \sqrt{\log n}}{n^{1/2}} + \frac{\mathrm{C}_2 \log n}{n^{\gamma - 1/2}} + \frac{\mathrm{C}_3 (\log n)^{3/2}}{n^{\gamma/2}} ,$$

where $\mathrm{C}_1, \mathrm{C}_2$ and C_3 are given in Appendix E.7, equation (63).

Remark 1. *It is possible to prove the result of Theorem 1 for the step size $\alpha_k = c_0/(k + k_0)$. The required Gaussian approximation result with the covariance matrix Σ_n is proved in [42], and we expect that the only difference with Theorem 1 will occur in extra $\log n$ factors in the corresponding bound and slightly different conditions on c_0 and k_0 in A6.*

Proof sketch of Theorem 1. The proof of non-asymptotic bootstrap validity is based on the Gaussian approximation performed both in the "real" world and bootstrap world together with an appropriate Gaussian comparison inequality:

$$\begin{array}{lcl} \text{Real world:} \quad \sqrt{n}(\bar{\theta}_n - \theta^\star) & \xleftrightarrow{\text{Gaussian approximation, Th. 2}} & \Sigma^{1/2} Y \sim \mathcal{N}(0, \Sigma) \\ & & \Big\updownarrow \text{Gaussian comparison, Lem. 19} \\ \text{Bootstrap world:} \ \sqrt{n}(\bar{\theta}_n^{\mathsf{b}} - \bar{\theta}_n) & \xleftrightarrow{\text{Gaussian approximation, Th. 3}} & \{\Sigma^{\mathsf{b}}\}^{1/2} Y^{\mathsf{b}} \sim \mathcal{N}(0, \Sigma^{\mathsf{b}}) \, . \end{array}$$

Here Σ and Σ^{b} are some covariance matrices to be chosen later. In order to understand where the Gaussian approximation comes from, we consider the process of linearization of statistics $\sqrt{n}(\bar{\theta}_n - \theta^\star)$ and $\sqrt{n}(\bar{\theta}_n^{\mathsf{b}} - \bar{\theta}_n)$. We provide details for $\sqrt{n}(\bar{\theta}_n - \theta^\star)$, and give similar derivations for $\sqrt{n}(\bar{\theta}_n^{\mathsf{b}} - \bar{\theta}_n)$ in Section 2.3. Denote $G = \nabla^2 f(\theta^\star)$. We expand $\sqrt{n}(\bar{\theta}_n - \theta^\star)$ into a weighted sum of independent random vectors, along with the remaining terms of smaller order. By the Newton-Leibniz formula, we obtain

$$\nabla f(\theta) = G(\theta - \theta^\star) + H(\theta), \tag{9}$$

where $H(\theta) = \int_0^1 (\nabla^2 f(\theta^\star + t(\theta - \theta^\star)) - G)(\theta - \theta^\star)\,\mathrm{d}t$. Note that $H(\theta)$ is of the order $\|\theta - \theta^\star\|^2$ (see Lemma 5). The recursion for the SGD algorithm error (7) can be expressed as

$$\theta_k - \theta^\star = (\mathrm{I}_d - \alpha_k G)(\theta_{k-1} - \theta^\star) - \alpha_k(\eta(\xi_k) + g(\theta_{k-1}, \xi_k) + H(\theta_{k-1})) \ . \tag{10}$$

For $i \in \{0, \dots, n-1\}$, we define the matrices

$$Q_i = \alpha_i \sum_{j=i}^{n-1} \prod_{k=i+1}^{j} (\mathrm{I}_d - \alpha_k G) \ , \tag{11}$$

where empty products are defined to be equal to I_d by convention. Then taking average of (10) and rearranging the terms, we obtain the following expansion:

$$\sqrt{n}(\bar{\theta}_n - \theta^\star) = W + D \ , \quad W = -\tfrac{1}{\sqrt{n}} \sum_{i=1}^{n-1} Q_i \eta(\xi_i), \quad D = \sqrt{n}(\bar{\theta}_n - \theta^\star) - W \ . \tag{12}$$

Note that W is a weighted sum of i.i.d. random vectors with mean zero and covariance matrix

$$\Sigma_n = n^{-1} \sum_{k=1}^{n-1} Q_k \Sigma_\xi Q_k^\top \ , \tag{13}$$

and D is the remainder term which is defined in Appendix C, equation (39). Furthermore, in Appendix D.1 we show that Q_i may be approximated by $G^{-\top}$ and Σ_n approximates

$$\Sigma_\infty = G^{-1} \Sigma_\xi G^{-\top} \ .$$

We expect that the summand D does not significantly distort the asymptotic distribution for the linear statistic W, which should be Gaussian by virtue of the central limit theorem. An important question is the choice of the approximating Gaussian distribution $\mathcal{N}(0, \Sigma)$ with $\Sigma = \Sigma_n$ or Σ_∞ as well their bootstrap counterpart Σ^{b}. This choice is instrumental in the sense that it does not change the procedure (5), but only affects the rates in (6). The authors of [16] choose the approximation with $\mathcal{N}(0, \Sigma_\infty)$ for their asymptotic analysis. A similar approach was considered in [38, Theorem 3] for the LSA algorithm setting. However, as it will be shown later in Theorem 4, this choice implies that the rate of normal approximation in (6) is not faster than $n^{-1/4}$. At the same time, Theorem 2 and Theorem 3 below demonstrate that we can achieve approximation rates of up to $n^{-1/2}$ by selecting $\Sigma = \Sigma_n$ in the diagram 2.1, and its bootstrap-world counterpart in the Gaussian approximation. To finish the proof, it remains to apply the Gaussian comparison inequality; see Lemma 19. Detailed proof of Theorem 1 is provided in Appendix E. □

Discussion. In [38] a counterpart of Theorem 1 was established with an approximation rate of the order $n^{-1/4}$ up to logarithmic factors for the setting of the LSA algorithm. The obtained rate is suboptimal, since the authors have chosen $\mathcal{N}(0, \Sigma_\infty)$ for Gaussian approximation when showing bootstrap validity. A recent paper [46] improved this rate to $n^{-1/3}$ for the temporal learning (TD) procedure with linear function approximation. The algorithm they considered is based on the direct estimate of Σ_∞, yielding a rate of order $n^{-1/3}$ when approximating the quantiles of $\sqrt{n}(\bar{\theta}_n - \theta^\star)$, see [46, Theorem 3.4 and 3.5]. The authors in [11] constructed a plug-in estimator $\hat{\Sigma}_n$ of Σ_∞ and showed guarantees of the form $\mathbb{E}[\|\hat{\Sigma}_n - \Sigma_\infty\|] \lesssim C n^{-\gamma/2}$, $\gamma \in (1/2, 1)$ under weaker assumptions than those considered in the current section. At the same time, approximating quantiles of $\sqrt{n}(\bar{\theta}_n - \theta^\star)$ with the method of [11] would require one more step - a Berry-Esseen type bound on the rate of approximation of $\sqrt{n}(\bar{\theta}_n - \theta^\star)$ with $\mathcal{N}(0, \Sigma_\infty)$. As we show in Theorem 4, this rate vanishes as $\gamma \to 1$, which introduces an additional trade-off to the potential analysis of the plug-in procedures based on estimating Σ_∞. This effect highlights the fundamental difference between the multiplier bootstrap approach and the plug-in approach of [11].

Moreover, we highlight that in-expectation bound $\mathbb{E}[\|\hat{\Sigma}_n - \Sigma_\infty\|]$, which are typically studied in literature for plug-in estimates [11, 35], are not sufficient to prove the analogue of the Gaussian comparison result Lemma 1 for $\mathcal{N}(0, \hat{\Sigma}_n)$ and $\mathcal{N}(0, \Sigma_\infty)$ on the set with large $\mathbb{P}$-probability. Thus, the complete non-asymptotic analysis of the confidence sets constructed with the plug-in procedure, remains an open problem.

2.2 Gaussian approximation in the real world

For results of this section, assumptions A2 and A6 can be relaxed. We impose a family of assumptions, denoted as A8(p) with $p \geq 2$, on the noise sequence ζ_k, and A9 on the step sizes α_k:

A 8 (p)**.** *Conditions (i) and (ii) from A 2 holds. Moreover, there exists* $\sigma_p > 0$ *such that* $\mathbb{E}^{1/p}[\|\eta(\xi_1)\|^p] \leq \sigma_p$.

A9. *Suppose that $\alpha_k = c_0/(k_0+k)^\gamma$, where $\gamma \in (1/2, 1)$, $k_0 \geq 1$, and c_0 satisfies $2c_0 L_1 \leq 1$.*

Note that *A*6 implies A9, as well as A2 implies A8(p) for any $p \geq 2$. In the main result of this section we provide the Gaussian approximation result for $\sqrt{n}(\bar{\theta}_n - \theta^\star)$ with $\mathcal{N}(0, \Sigma_n)$, which refines the bounds obtained in [42, Theorem 3.4] and is instrumental for further studies of normal approximation with $\mathcal{N}(0, \Sigma_\infty)$ in Section 2.4.

Theorem 2. *Assume A1, A3, A8(4), A9. Then, with $Y \sim \mathcal{N}(0, \mathrm{I}_d)$, it holds that*

$$\mathsf{d}_{\mathsf{C}}(\sqrt{n}\Sigma_n^{-1/2}(\bar{\theta}_n - \theta^\star), Y) \leq \frac{\mathrm{C}_4}{\sqrt{n}} + \frac{\mathrm{C}_5}{n^{\gamma-1/2}} + \frac{\mathrm{C}_6}{n^{\gamma/2}} \,, \tag{14}$$

where $\mathrm{C}_4, \mathrm{C}_5, \mathrm{C}_6$ are given in Appendix C, equation (40). Moreover, since Σ_n is non-degenerate, and an image of a convex set under non-degenerate linear mapping is a convex set, we have

$$\mathsf{d}_{\mathsf{C}}(\sqrt{n}\Sigma_n^{-1/2}(\bar{\theta}_n - \theta^\star), Y) = \mathsf{d}_{\mathsf{C}}(\sqrt{n}(\bar{\theta}_n - \theta^\star), \Sigma_n^{1/2} Y) \,.$$

Remark 2. *When $\gamma \to 1$, the correction terms above scale as $\mathcal{O}(1/\sqrt{n})$, yielding the overall approximation rate that approaches $1/\sqrt{n}$. Expressions for $\mathrm{C}_4, \mathrm{C}_5, \mathrm{C}_6$ from Theorem 2 depend upon the problem dimension d, parameters specified in A1 - A8(4)- A3-A9. Moreover, C_5 depends upon $\|\theta_0 - \theta^\star\|$. When $\gamma \in (0,1)$, we have that $1/n^{\gamma/2} < 1/n^{\gamma-1/2}$, thus, the term $\mathrm{C}_5\,/n^{\gamma-1/2}$ dominates. We prefer to keep both terms in (14), since they are responsible for the moments of statistics $\frac{1}{\sqrt{n}}\sum_{i=1}^{n-1} Q_i H(\theta_{i-1})$ and $\frac{1}{\sqrt{n}}\sum_{i=1}^{n-1} Q_i g(\theta_{i-1}, \xi_i)$, respectively. The first of them has non-zero mean, since $H(\theta_{i-1})$ is quadratic in $\|\theta_i - \theta^\star\|^2$. When using constant step size SGD, one can correct this term using the Richardson-Romberg technique [14, 43], however, it is unclear if this type of ideas can be generalized for diminishing step size.*

Proof sketch of Theorem 2. The decomposition (12) represents a particular instance of the general problem of Gaussian approximation for nonlinear statistics of the form $\sqrt{n}(\bar{\theta}_n - \theta^\star)$, where the estimator is expressed as the sum of linear and nonlinear components. To establish the Gaussian approximation result, we adapt the arguments from [42], which can be stated as follows. Let $X_1, \ldots, X_n$ be independent random variables taking values in some space $\mathcal{X}$, and let $T = T(X_1, \ldots, X_n)$ be a general d-dimensional statistic that can be decomposed as

$$W := W(X_1, \ldots, X_n) = \textstyle\sum_{\ell=1}^n Z_\ell, \quad D := D(X_1, \ldots, X_n) = T - W \,.$$

Here, we define $Z_\ell = r_\ell(X_\ell)$, where $r_\ell : \mathcal{X} \to \mathbb{R}^d$ is a Borel-measurable function. The term D represents the nonlinear component and is treated as an error term, assumed to be "small" relative to W in an appropriate sense. Suppose that $\mathbb{E}[Z_\ell] = 0$ and that the Z_ℓ is normalized in such a way that $\sum_{\ell=1}^n \mathbb{E}[Z_\ell Z_\ell^\top] = \mathrm{I}_d$ holds. Let $\Upsilon_n = \sum_{\ell=1}^n \mathbb{E}[\|Z_\ell\|^3]$. Then, for $Y \sim \mathcal{N}(0, \mathrm{I}_d)$, the following bound holds:

$$\mathsf{d}_{\mathsf{C}}(T, Y) \leq 259 d^{1/2}\Upsilon_n + 2\mathbb{E}[\|W\|\|D\|] + 2\sum_{\ell=1}^n \mathbb{E}[\|Z_\ell\|\|D - D^{(\ell)}\|], \tag{15}$$

where $D^{(\ell)} = D(X_1, \ldots, X_{\ell-1}, X'_\ell, X_{\ell+1}, \ldots, X_n)$ and X'_ℓ is an independent copy of X_ℓ. This result follows from [42, Theorem 2.1]. Furthermore, this bound can be extended to the case where $\sum_{\ell=1}^n \mathbb{E}[Z_\ell Z_\ell^\top] = \Sigma \succ 0$, as detailed in [42, Corollary 2.3]. In order to apply (15), we let $X_i = \xi_i$, $Z_\ell = h(X_\ell)$, ξ'_i be an i.i.d. copy of ξ_i. Then we need to upper bound $\mathbb{E}^{1/2}[\|D(\xi_1, \ldots, \xi_{n-1})\|^2]$ and $\mathbb{E}^{1/2}[\|D - D'_i\|^2]$, respectively. Detailed proof is given in Appendix C. □

2.3 Gaussian approximation in the bootstrap world

In the main result of this section, we study the Gaussian approximation result for $\sqrt{n}(\bar{\theta}_n^{\mathsf{b}} - \bar{\theta}_n)$ with appropriate normal distribution with respect to $\mathbb{P}^{\mathsf{b}}$. Despite this result is similar in its nature with the one of Theorem 2, it requires to handle some significant challenges that arises when working in the "bootstrap world". Our first steps are the same as in (10) and (5):

$$\begin{aligned}
\theta_k^{\mathsf{b}} - \theta_k = {} & (I - \alpha_k G)(\theta_{k-1}^{\mathsf{b}} - \theta_{k-1}) \\
& - \alpha_k\big(H(\theta_{k-1}^{\mathsf{b}}) + g(\theta_{k-1}^{\mathsf{b}}, \xi_k) - H(\theta_{k-1}) - g(\theta_{k-1}, \xi_k)\big) \\
& - \alpha_k(w_k - 1)\big(G(\theta_{k-1}^{\mathsf{b}} - \theta^\star) + \eta(\xi_k) + g(\theta_{k-1}^{\mathsf{b}}, \xi_k) + H(\theta_{k-1}^{\mathsf{b}})\big) \,.
\end{aligned} \tag{16}$$

Taking an average of (16) and rearranging the terms, we obtain a counterpart of (12):

$$\sqrt{n}(\bar{\theta}_n^{\mathsf{b}} - \bar{\theta}_n) = W^{\mathsf{b}} + D^{\mathsf{b}} \,,$$
$$W^{\mathsf{b}} = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} (w_i - 1) Q_i \eta(\xi_i) \,, \quad D^{\mathsf{b}} = \sqrt{n}(\bar{\theta}_n^{\mathsf{b}} - \bar{\theta}_n) - W^{\mathsf{b}} \,. \tag{17}$$

Here W^{b} is a weighted sum of i.i.d. random variables Ξ^{n-1}, such that $\mathbb{E}^{\mathsf{b}}[W^{\mathsf{b}}] = 0$ and

$$\mathbb{E}^{\mathsf{b}}[W^{\mathsf{b}}\{W^{\mathsf{b}}\}^\top] := \Sigma_n^{\mathsf{b}} = n^{-1} \sum_{i=1}^{n-1} Q_i \eta(\xi_i) \eta(\xi_i)^\top Q_i^\top \,,$$

and D^{b} is a non-linear statistic of Ξ^{n-1}. The principal difficulty arises when considering the conditional distribution of $\sqrt{n}(\bar{\theta}_n^{\mathsf{b}} - \bar{\theta}_n)$ given the data Ξ^{n-1}. In fact, the approach of [42] would require to control the second moments of D^{b} and $D^{\mathsf{b}} - \{D^{\mathsf{b}}\}^{(i)}$ with respect to a bootstrap measure $\mathbb{P}^{\mathsf{b}}$, on the high-probability event with respect to a measure $\mathbb{P}$. At the same time, we loose a martingale structure of the summands in D^{b}, unless we condition on the extended filtration

$$\widetilde{\mathcal{F}}_i = \sigma(w_1, \ldots w_i, \xi_1, \ldots \xi_i) \,, 1 \leq i \leq n-1 \,. \tag{18}$$

Therefore, it is not clear if we can directly apply the approach of [42] discussed in Section 2.2. Instead, we have to use the linearization approach based on the high-order moment bounds for the remainder term D^{b} (see Proposition 3 in Appendix E). This justifies the strong bounded noise assumption A2, that we had to impose. We state the main result of this section below:

Theorem 3. *Assume A1 - A7. Then with $\mathbb{P}$ - probability at least $1 - 2/n$, it holds*

$$\sup_{B \in \mathsf{C}(\mathbb{R}^d)} |\mathbb{P}^{\mathsf{b}}(\sqrt{n}\{\Sigma_n^{\mathsf{b}}\}^{-1/2}(\bar{\theta}_n^{\mathsf{b}} - \bar{\theta}_n) \in B) - \mathbb{P}^{\mathsf{b}}(Y^{\mathsf{b}} \in B)| \leq \frac{M_{3,1}^{\mathsf{b}}}{n^{1/2}} + \frac{M_{3,2}^{\mathsf{b}} \log n}{n^{\gamma - 1/2}} + \frac{M_{3,3}^{\mathsf{b}} \log^{3/2} n}{n^{\gamma/2}} \,,$$

where $\{M_{3,i}^{\mathsf{b}}\}_{i=1}^3$ are defined in Appendix E.6, equation (61).

Proof sketch of Theorem 3. We apply the bound

$$\begin{aligned} &\sup_{B \in \mathsf{C}(\mathbb{R}^d)} |\mathbb{P}^{\mathsf{b}}(\{\Sigma_n^{\mathsf{b}}\}^{-\frac{1}{2}}(W^{\mathsf{b}} + D^{\mathsf{b}}) \in B) - \mathbb{P}^{\mathsf{b}}(Y^{\mathsf{b}} \in B)| \\ &\leq \sup_{B \in \mathsf{C}(\mathbb{R}^d)} |\mathbb{P}^{\mathsf{b}}(\{\Sigma_n^{\mathsf{b}}\}^{-\frac{1}{2}} W^{\mathsf{b}} \in B) - \mathbb{P}^{\mathsf{b}}(Y^{\mathsf{b}} \in B)| + 2c_d (\mathbb{E}^{\mathsf{b}}[\|\{\Sigma_n^{\mathsf{b}}\}^{-\frac{1}{2}} D^{\mathsf{b}}\|^p])^{\frac{1}{1+p}} \,, \end{aligned} \tag{19}$$

where $c_d \leq 4d^{1/4}$ is the isoperimetric constant of the class of convex sets. The proof of (19) is provided in Proposition 3 in Appendix E. We can control $\bar{\mathbb{E}}[\|D^{\mathsf{b}}\|^p]$ by Burkholder's inequality, where $\bar{\mathbb{E}}$ denotes the expectation w.r.t. the product measure $\mathbb{P}_\xi^{\otimes n} \otimes \mathbb{P}_w^{\otimes n}$. Then we proceed with Markov's inequality to obtain $\mathbb{P}$ – high-probability bounds on the behavior of $\mathbb{E}^{\mathsf{b}}[\|D^{\mathsf{b}}\|^p]$. This result requires us to provide bounds for

$$\bar{\mathbb{E}}^{1/p}[\|\theta_k - \theta^\star\|^p] \quad \text{and} \quad \bar{\mathbb{E}}^{1/p}[\|\theta_k^{\mathsf{b}} - \theta^\star\|^p] \,, \quad k \in \{1, \ldots, n-1\} \,, \tag{20}$$

with $p \simeq \log n$ and polynomial dependence on p. To control the second term in the r.h.s. of (19) we note that the matrix Σ_n^{b} concentrates around Σ_n due to the matrix Bernstein inequality (see Lemma 18 for details). Hence, there is a set Ω_1 such that $\mathbb{P}(\Omega_1) \geq 1 - 1/n$ and $\lambda_{\min}(\Sigma_n^{\mathsf{b}}) > 0$ on Ω_1. Moreover, on this set we may use Berry-Essen-type bound for non-i.i.d. sums of random vectors. Detailed proof is given in Appendix E. □

2.4 Rate of convergence in the Polyak–Juditsky central limit theorem

In the final part of this section, we discuss the issue of transition from Σ_n to Σ_∞ and estimation of convergence rates in the Polyak–Juditsky result (4). We utilize the result of Theorem 2 together with the following lemma.

Lemma 1. *Assume that A1 and A9 hold. Let $Y, Y' \sim \mathcal{N}(0, \mathrm{I}_d)$. Then, the Kolmogorov distance between the distributions of $\Sigma_n^{1/2} Y$ and $\Sigma_\infty^{1/2} Y'$ is bounded by*

$$\mathsf{d}_{\mathsf{C}}(\Sigma_n^{1/2} Y, \Sigma_\infty^{1/2} Y') \leq C_\infty n^{\gamma - 1} \,,$$

where the constant C_∞ is defined in (50).

Theorem 2, Lemma 1, and triangle inequality imply the following result on closeness to $\mathcal{N}(0, \Sigma_\infty)$.

Theorem 4. *Assume A1, A3, A8(4), A9. Then, with $Y \sim \mathcal{N}(0, \mathrm{I}_d)$ it holds that*

$$\mathsf{d}_{\mathsf{C}}(\sqrt{n}(\bar{\theta}_n - \theta^\star), \Sigma_\infty^{1/2} Y) \le \frac{\mathrm{C}_4}{\sqrt{n}} + \frac{\mathrm{C}_5}{n^{\gamma - 1/2}} + \frac{\mathrm{C}_6}{n^{\gamma/2}} + \frac{C_\infty}{n^{1-\gamma}} , \tag{21}$$

where $\mathrm{C}_4, \mathrm{C}_5$ and C_6 are given in Theorem 2.

Discussion. Theorem 2 reveals that the normal approximation through $\mathcal{N}(0, \Sigma_n)$ improves when the step sizes α_k are less aggressive, that is, as $\gamma \to 1$. However, Theorem 4 shows that there is a trade-off, since the rate at which Σ_n converges to Σ_∞ also affects the overall quality of the approximation. Optimizing the bound in (21) for γ yields an optimal value of $\gamma = 3/4$, leading to the following approximation rate:

$$\mathsf{d}_{\mathsf{C}}(\sqrt{n}(\bar{\theta}_n - \theta^\star), \Sigma_\infty^{1/2} Y) \le \frac{\mathrm{C}_1'}{n^{1/4}} + \frac{\mathrm{C}_2'}{\sqrt{n}}(\|\theta_0 - \theta^\star\| + \|\theta_0 - \theta^\star\|^2) ,$$

where C_1' and C_2' are instance-dependent quantities (but not depending on $\|\theta_0 - \theta^\star\|$), that can be inferred from Theorem 4. Given the result of Theorem 4 one can proceed with a non-asymptotic evaluation of the methods for constructing confidence intervals based on direct estimation of Σ_∞, such as [11, 49].

Lower bounds We provide a lower bound indicating that the bound Theorem 4 is tight at least in some regimes of step size decay power $\gamma \in (1/2, 1)$. For this aim we consider minimization problem (1) of the following form:

$$f(\theta) = \theta^2/2 \to \min_{\theta \in \mathbb{R}} , \quad \theta_0 = 0 .$$

In this case $\theta^\star = 0$. We consider an additive noise model, that is, the stochastic gradient oracles $F(\theta, \xi)$ have a form $F(\theta, \xi) = \theta + \xi$, where $\xi \sim \mathcal{N}(0, 1)$. Enrolling (2), we get

$$\theta_k = -\sum_{j=1}^{k} \alpha_j \prod_{\ell=j+1}^{k} (1 - a\alpha_\ell)\xi_j \text{ and } \sqrt{n}\bar{\theta}_n = -\frac{1}{\sqrt{n}} \sum_{j=1}^{n-1} Q_j \xi_j , \tag{22}$$

where $Q_j = \alpha_j \sum_{k=j}^{n-1} \prod_{\ell=j+1}^{k} (1 - \alpha_\ell)$. Note that $\sqrt{n}(\bar{\theta}_n - \theta^\star)$ follows normal distribution $\mathcal{N}(0, \sigma_{n,\gamma}^2)$ with $\sigma_{n,\gamma}^2 = \frac{1}{n} \sum_{j=1}^{n-1} Q_j^2$. Due to Lemma 1 (see also equation (49) in the Appendix), we have $G = 1, \Sigma_\infty = 1$, and $\sigma_{n,\gamma}^2 \to 1$ as $n \to \infty$. Moreover, the following lower bound holds:

Proposition 1. *Consider the sequence $\{\theta_k\}_{k \in \mathbb{N}}$ defined by the recurrence (22) with $\alpha_j = c_0/(1+j)^\gamma$. Then it holds, for the number of observations n sufficiently large, that*

$$|\sigma_{n,\gamma}^2 - 1| > \tfrac{C_1(\gamma, c_0)}{n^{1-\gamma}} , \tag{23}$$

where the constant $C_1(\gamma, c_0)$ depends only upon c_0 and γ. Moreover, for n large enough

$$\mathsf{d}_{\mathsf{C}}(\sqrt{n}(\bar{\theta}_n - \theta^\star), \mathcal{N}(0,1)) > \tfrac{C_2(\gamma, c_0)}{n^{1-\gamma}} . \tag{24}$$

Discussion. Proof of Proposition 1 is provided in Appendix F, together with some simple numerical simulations which indicate the tightness of the lower bound (23). Note that the bound (24) reveals that the distribution of $\sqrt{n}(\bar{\theta}_n - \theta^\star)$ can not be approximated by $\mathcal{N}(0, \Sigma_\infty)$ with the rate faster than $1/n^{1-\gamma}$. Moreover, it shows that the rate of normal approximation in Theorem 4 can not be improved when $\gamma \in [3/4; 1)$. This fact is extremely important when taking into account the bootstrap validity result of Theorem 1 and normal approximation in Theorem 2. Indeed, both results suggests that the rates of normal approximation of order up to $1/\sqrt{n}$ can be achieved when $\gamma \to 1$, but they require to consider another covariance matrix Σ_n, corresponding to the linearized recurrence in (13). At the same time, in the regime $\gamma \to 1$, the approximation by $\mathcal{N}(0, \Sigma_\infty)$ can be too slow. It is an interesting and, to the best of our knowledge, open question to provide lower bounds analogous to Proposition 1 which show the tightness of other summands in Theorem 4 in the regime $1/2 < \gamma < 3/4$.

3 Conclusion

In our paper, we performed the fully non-asymptotic analysis of the multiplier bootstrap procedure for SGD applied to strongly convex minimization problems. We showed that the algorithm can achieve approximation rates in convex distances of order up to $1/\sqrt{n}$. We highlight the fact that the validity of the multiplier bootstrap procedure does not require one to consider Berry-Esseen bounds with the asymptotic covariance matrix Σ_∞, which is in sharp contrast to the methods that require direct estimation of Σ_∞.

References

[1] Bhavya Agrawalla, Krishnakumar Balasubramanian, and Promit Ghosal. High-dimensional central limit theorems for linear functionals of online least-squares sgd. *arXiv preprint arXiv:2302.09727*, 2023.

[2] Andreas Anastasiou, Krishnakumar Balasubramanian, and Murat A. Erdogdu. Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale CLT. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 115–137. PMLR, 25–28 Jun 2019.

[3] G. D. Anderson and S.-L. Qiu. A monotoneity property of the gamma function. *Proceedings of the American Mathematical Society*, 125(11):3355–3362, 1997.

[4] Keith Ball. The reverse isoperimetric problem for Gaussian measure. *Discrete & Computational Geometry*, 10(4):411–420, October 1993.

[5] S. Barsov and V. Ulyanov. Estimates for the closeness of Gaussian measures. *Dokl. Akad. Nauk SSSR*, 291(2):273–277, 1986.

[6] V. Bentkus. On the dependence of the berry–esseen bound on dimension. *Journal of Statistical Planning and Inference*, 113(2):385–402, 2003.

[7] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012.

[8] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

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

[10] Xi Chen, Zehua Lai, He Li, and Yichen Zhang. Online statistical inference for contextual bandits via stochastic gradient descent. *arXiv preprint arXiv:2212.14883*, 2022.

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

[12] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

[13] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional Gaussians with the same mean. *arXiv e-prints*, page arXiv:1810.08693, October 2018.

[14] Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *The Annals of Statistics*, 48(3):1348 – 1382, 2020.

[15] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer, 1992.

[16] Yixin Fang, Jinfeng Xu, and Lei Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *Journal of Machine Learning Research*, 19(78):1–21, 2018.

[17] M. V. Fedoryuk. *Metod perevala*. Izdat. "Nauka", Moscow, 1977.

[18] Wolfgang Gabcke. Neue herleitung und explizite restabschätzung der riemann-siegel-formel. 2015.

[19] Friedrich Götze, Alexey Naumov, Vladimir Spokoiny, and Vladimir Ulyanov. Large ball probabilities, Gaussian comparison and anti-concentration. *Bernoulli*, 25(4A):2538–2563, 2019.

[20] Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613. PMLR, 2019.

[21] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

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

[23] Guanghui Lan. *First-order and Stochastic Optimization Methods for Machine Learning*. 01 2020.

[24] Chris Junchi Li, Wenlong Mou, Martin Wainwright, and Michael Jordan. Root-sgd: Sharp nonasymptotics and asymptotic efficiency in a single algorithm. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 909–981. PMLR, 02–05 Jul 2022.

[25] Xiang Li, Jiadong Liang, Xiangyu Chang, and Zhihua Zhang. Statistical estimation and online inference via local sgd. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1613–1661. PMLR, 02–05 Jul 2022.

[26] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24:451–459, 2011.

[27] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[28] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer, 2004.

[29] Frank W. J. Olver. *Asymptotics and special functions*. AKP Classics. A K Peters, Ltd., Wellesley, MA, 1997. Reprint of the 1974 original [Academic Press, New York; MR0435697 (55 #8655)].

[30] Adam Osekowski. *Sharp martingale and semimartingale inequalities*, volume 72. Springer Science & Business Media, 2012.

[31] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

[32] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

[33] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1571–1578, 2012.

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

[35] Abhishek Roy and Krishnakumar Balasubramanian. Online covariance estimation for stochastic gradient descent under markovian sampling. *arXiv preprint arXiv:2308.01481*, 2023.

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

[37] Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *International Conference on Machine Learning*, pages 29563–29648. PMLR, 2023.

[38] Sergey Samsonov, Eric Moulines, Qi-Man Shao, Zhuo-Song Zhang, and Alexey Naumov. Gaussian Approximation and Multiplier Bootstrap for Polyak-Ruppert Averaged Linear Stochastic Approximation with Applications to TD Learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 12408–12460. Curran Associates, Inc., 2024.

[39] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162:83–112, 2017.

[40] Jun Shao. *Mathematical statistics*. Springer Science & Business Media, 2003.

[41] Jun Shao and Dongsheng Tu. *The Jackknife and Bootstrap*. Springer Series in Statistics. Springer New York, NY, 1 edition, 1995. Springer Book Archive, © Springer Science+Business Media New York 1995.

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

[43] Marina Sheshukova, Denis Belomestny, Alain Durmus, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Nonasymptotic analysis of stochastic gradient descent with the richardson-romberg extrapolation. *arXiv preprint arXiv:2410.05106*, 2024.

[44] R Srikant. Rates of Convergence in the Central Limit Theorem for Markov Chains, with an Application to TD learning. *arXiv preprint arXiv:2401.15719*, 2024.

[45] Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

[46] Weichen Wu, Gen Li, Yuting Wei, and Alessandro Rinaldo. Statistical Inference for Temporal Difference Learning with Linear Function Approximation. *arXiv preprint arXiv:2410.16106*, 2024.

[47] Yanjie Zhong, Todd Kuffner, and Soumendra Lahiri. Online Bootstrap Inference with Nonconvex Stochastic Gradient Descent Estimator. *arXiv preprint arXiv:2306.02205*, 2023.

[48] D. L. Zhu and P. Marcotte. Co-coercivity and its role in the convergence of iterative schemes for solving variational inequalities. *SIAM Journal on Optimization*, 6(3):714–726, 1996.

[49] Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online Covariance Matrix Estimation in Stochastic Gradient Descent. *Journal of the American Statistical Association*, 118(541):393–404, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract clearly lays out the paper's main contributions—namely, the non-asymptotic validity of the multiplier bootstrap for constructing confidence sets in SGD, with convex-distance approximation rates up to $O(n^{-\gamma/2})$ (and up to $O(n^{-1/2})$ as $\gamma \to 1$), constituting the first fully non-asymptotic bound of this kind in SGD algorithms . The introduction then explicitly enumerates these same contributions (non-asymptotic bootstrap validity, the role of the linearized covariance Σ_n, and Gaussian approximation rates with matching lower bounds) in its "Our key contributions" section, directly reflecting the theorems and scope developed later in the paper.

   Main results are, respectively, the ones of Theorem 1 and Theorem 3, their statements are complete and supported by the proofs in the Appendix section.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section 2, we explicitly acknowledge key limitations of our analysis, including the almost-sure boundedness requirements on the noise and bootstrap weights (Assumptions A2(ii)–(iii) and A5) as well as Lipschitz and co-coercivity conditions on the gradient. We furthermore note that relaxing these via gradient clipping schemes is left as future work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results are stated with explicit pointers to the underlying assumptions (see Sections 2 and 2.2), and every theorem and corollary is accompanied by a full, detailed proof in the supplementary paper, with clear references to the corresponding sections. To help the reader, we also include the proofs of some non-standard results (on whihc we do not claim originality); we always carefully state sources.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Numerical results for Proposition 1 are given in Appendix F with a full description of the sets of hyperparameters we used.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Even though the experiment is toy-like - most of the paper is theoretical-, we include a link to an anonymous GitHub repository.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental section in Appendix F explicitly provides the settings (step size, specific equations, number of observations, etc.) necessary for understanding and reproducing the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We omit error bars since the experiment contains no stochastic components.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: All necessary information to reproduce experiments is provided in Appendix F.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: This paper is of purely theoretical nature, and the proposed methods do not deal with sensitive attributes that could induce unfairness or privacy issues.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This paper is of purely theoretical nature. We do not foresee any societal harm from the proof of non-asymptotic bootstrap validity.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper contains no models or datasets with potential for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

Justification: No external datasets, software, or assets were used in this paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new assets were introduced in this research.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Paper does not involve crowdsourcing nor research on human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Paper does not involve crowdsourcing nor research on human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The paper does not involve the use of LLMs or related methods.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.