From Cursed to Competitive: Closing the ZO–FO Gap via Input-to-State Stability

Amir Ali Farzin
School of Engineering
Australian National University
Canberra, ACT 2601
amirali.farzin@anu.edu.au

Philipp Braun
School of Engineering
Australian National University
Canberra, ACT 2601
philipp.braun@anu.edu.au

Iman Shames
Department of Electrical and Electronic Engineering
University of Melbourne
Melbourne, VIC 3010
iman.shames@unimelb.edu.au

Abstract

While it is generally understood that zeroth-order (ZO) algorithms incur an extra dimension dependence in their iteration complexity compared to their first-order (FO) counterparts, for any choice of parameters, we show that under several conditions, in expectation, ZO methods do not suffer from extra dimension dependence in their convergence rates. We look at optimisation algorithms from the dynamical systems perspective and identify conditions under which the average of a ZO algorithm can be written as its FO counterpart with bounded perturbations whose size depends on design parameters. Then, using input-to-state stability properties, we show that ZO methods follow the same decay rate as their FO counterparts and converge to a neighbourhood of the fixed point of FO methods, where its radius depends on the bound of the norm of the perturbations, which can be made arbitrarily small through the smoothing parameter, step size and number of sampled directions. The theoretical findings are illustrated via numerical examples.

1 Introduction

In this paper, we study a minimisation problem of the form

$$\min_{x\in\mathbb{R}^n} \quad f(x) \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, and bounded below. Such problems arise routinely in machine learning, signal processing, and control. While previously, in the literature, it was commonly understood that convergence complexity bounds for zeroth-order (ZO) methods have extra dimension dependencies compared to their first-order (FO) counterparts, in this work, we analyse conditions under which it is guaranteed that ZO methods, in expectation, have no extra dimension dependencies in their convergence rate compared to their FO counterparts. Here, we study iterative algorithms and their convergence properties from the dynamical system perspective and analyse them by studying input-to-state stability (ISS) properties.

Motivation: In many modern applications, gradient information is either unavailable, prohibitively expensive to compute, or simply does not exist in closed form. Examples include black-box adversarial attacks [CZS$^+$17, YHF$^+$18, MMS$^+$18], reinforcement learning with opaque simu-

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

lators [SHC+17, CRS+18], hyperparameter tuning [SLA12], simulation-based engineering design [CSV09], and memory-efficient fine-tuning of large language models [MGN+23]. In these settings, one must resort to ZO methods that construct gradient surrogates from function evaluations alone. While attractive for their simplicity and broad applicability, ZO methods have traditionally been seen as having fundamentally weaker convergence guarantees than their FO counterparts, independent of the parameter selection.

The dimension-dependence gap: The seminal work of [NS17] established the modern framework for ZO optimisation based on Gaussian smoothing and showed that, for smooth convex problems, ZO gradient descent requires at most n times more iterations than standard gradient descent, where n is the dimension. This $\mathcal{O}(n)$ factor has become the standard penalty attributed to working without gradient information. Subsequent works have refined and extended these bounds across convex [DJWW15, BG18], nonconvex stochastic [GL13], and structured non-convex settings such as quasar-convex, submodular, and Polyak-Łojasiewicz functions [FPBS25a, FPB+25b, FS24]. In each case, the convergence bounds carry an explicit dependence on n beyond what FO methods require. Efforts to reduce this overhead via sparsity [WDBS18, BG22] or effective dimension [YYFL23] typically require additional structural assumptions (e.g., gradient sparsity or a fast-decaying Hessian spectrum) that may not hold in general.

Related work: On the ZO optimisation side, building on [NS17, GL13], ZO variants of classical FO algorithms have been developed, including accelerated methods [DGG21], variance-reduced schemes [LKC+18], and minimax methods [LLC+20, WBMR20, FPB+25a, FPB+26]. Notably, the MeZO framework [MGN+23] demonstrated that ZO-SGD can fine-tune LLMs with billions of parameters at competitive performance, attributed to the favourable loss landscape induced by pre-training; subsequent work [ZLH+24] benchmarked broader ZO optimiser families in this setting. On the dynamical systems side, interpreting optimisation algorithms as discrete-time dynamical systems is well established [LRP16, WRJ21, FPBS26, DP18, FPBS25b]. Input-to-state stability (ISS), introduced in [Son89] and extended to discrete-time systems in [JW01], provides a natural framework for analysing robustness of stable equilibria to bounded perturbations and has been widely applied in nonlinear control [Son08, KB23]. ISS has recently been used to show that perturbed FO methods converge to a neighbourhood of the fixed point [CJSB25]. These works certify stability or boundedness, whereas we show that the ZO dynamics inherit the *decay rate* of the FO method and use it to bridge ZO and FO convergence theory, which, to the best of our knowledge, is novel. Structured orthogonal directions [KMR+23] also improve dimension dependence, at the cost of up to $O(n)$ queries per iteration. A detailed comparison of our results with those of prior ZO methods is provided in Table 2 (Appendix A).

Contributions: The main contributions of this paper are as follows.

(i) We show that, under suitable conditions, the averaged dynamics of a Gaussian ZO algorithm approximate its FO counterpart, where the approximation error is bounded and can be made arbitrarily small by tuning the smoothing parameter μ, the step size h, and the number of sampled directions t.

(ii) Leveraging the ISS framework, we prove that ZO methods inherit the same transient convergence rate as their FO counterparts, i.e., the same exponential/asymptotic decay, and converge to a neighbourhood of the FO fixed point whose radius depends on the perturbation bound but carries no additional dependence on the number of iterations (Theorems 1 and 2). This provides a fundamentally different explanation of ZO convergence behaviour. The dimension dependence does not appear in the iteration complexity but rather in the size of the convergence neighbourhood. Moreover, the framework serves as a general-purpose tool. Whenever the FO algorithm is known to converge, the ISS machinery automatically yields convergence of its ZO counterpart to a neighbourhood of the FO fixed point, without requiring a separate, algorithm-specific proof. Our results concern *iteration complexity*. Each ZO step uses t function evaluations, and the t required for a target radius may depend on the oracle's variance and hence, in the worst case, on n. These evaluations are, however, parallel within an iteration rather than sequential. (iii) We apply this framework to GD, the heavy ball, and Nesterov's accelerated gradient. Under strong convexity, we derive explicit perturbation bounds (Lemmas 2, 3, 4) and dimension-free convergence rate results (Theorems 3, 4, 5). For non-strongly-convex objectives, we show that L_2 regularisation provides sufficient contraction, enabling the same dimension-free analysis with a controlled trade-off between regularisation strength and neighbourhood size (Lemma 5 and Theorem 6).

(iv) We validate the theoretical findings through numerical experiments, confirming that ZO trajectories closely track their FO counterparts and demonstrating how μ, h, and t control the proximity.

Outline: Preliminaries are introduced in Section 2. In Section 3, the main results on the dimensional independency of ZO methods are given. In Section 4, some illustrative examples are studied. The conclusions and future research directions are discussed in Section 5. Additional details and the proof of the technical results can be found in Appendices A- E.

Notation: For $x, y \in \mathbb{R}^n$, we write $\langle x, y\rangle = x^\top y$, and let $\|\cdot\|$ denote the Euclidean norm of a vector and the corresponding spectral norm of a matrix. For $\delta > 0$ and $z^* \in \mathbb{R}^n$, we define $\mathcal{B}_\delta(z^*) := \{z \in \mathbb{R}^n \mid \|z - z^*\| \le \delta\}$. For $f : \mathbb{R}^n \to \mathbb{R}$, we denote the gradient and Hessian by ∇f and $\nabla^2 f$, respectively. The identity matrix of proper size is denoted by I, and Id denotes the identity function $\mathrm{Id}(s) = s$. A continuous function $\alpha : [0, a) \to [0, \infty)$ (with $a > 0$ or $a = \infty$) is of class $\mathcal{K}$ if it is strictly increasing and $\alpha(0) = 0$. A function α is of class $\mathcal{K}_\infty$ if $\alpha \in \mathcal{K}$ and $\alpha(s) \to \infty$ as $s \to \infty$. A continuous function $\beta : [0, \infty) \times [0, \infty) \to [0, \infty)$ is of class $\mathcal{KL}$ if, for each fixed $t \ge 0$, the map $s \mapsto \beta(s, t)$ belongs to $\mathcal{K}$, and for each fixed $s \ge 0$, the map $t \mapsto \beta(s, t)$ is decreasing with $\beta(s, t) \to 0$ as $t \to \infty$.

2 Preliminaries

In this section, we present preliminaries on dynamical systems, ISS, and ZO Gaussian oracles.

2.1 Dynamical systems and perturbations

We study optimisation problem (1) through the lens of discrete-time dynamical systems, interpreting an algorithm as

$$z_{k+1} = w(z_k) \qquad (\text{or } z^+ = w(z)), \tag{2}$$

with state $z \in \mathbb{R}^n$ and update map $w : \mathbb{R}^n \to \mathbb{R}^n$. A point $z^e \in \mathbb{R}^n$ is an equilibrium (fixed point) of w if $w(z^e) = z^e$. To analyse stability of equilibria, we use standard notions from discrete-time systems theory; see, e.g., [KB23, Ch. 2, Ch. 5].

Definition 1. *An equilibrium $z^e \in \mathbb{R}^n$ of* (2) *is* stable *if there exist $\alpha \in \mathcal{K}_\infty$ and $\delta > 0$ such that $\|z_k - z_e\| \le \alpha(\|z_0 - z_e\|)$ for all $k \in \mathbb{N}$ and $z_0 \in \mathcal{B}_\delta(z^e)$. If z^e is not stable, it is called* unstable. *The equilibrium is* asymptotically stable *if there exist $\beta \in \mathcal{KL}$ and $\delta > 0$ such that*

$$\|z_k - z_e\| \le \beta(\|z_0 - z_e\|, k) \qquad \forall\, k \in \mathbb{N}, \quad \forall\, z_0 \in \mathcal{B}_\delta(z^e). \tag{3}$$

If β in (3) *has the form $\beta(r, k) = Mr\nu^k$ for $M > 0$ and $\nu \in (0, 1)$, then z^e is* exponentially stable.

For an optimisation algorithm, $\mathcal{KL}$-stability or asymptotic stability of an equilibrium of (2) is equivalent to local convergence to the associated critical point. We now consider a perturbed version of (2):

$$z_{k+1} = \bar{w}(z_k, q_k) = w(z_k) + q_k, \tag{4}$$

where $\{q_k\}_{k\in\mathbb{N}}$ models perturbations (e.g., inexact gradients or noise). When these perturbations are bounded, robustness can be studied via ISS [SW96, JW01].

Definition 2 (Input-to-state stability)**.** *The system* (4) *with $z_k, q_k \in \mathbb{R}^n$ is (locally) input-to-state stable (ISS) if there exist $\delta > 0$, $\beta \in \mathcal{KL}$, and $\gamma \in \mathcal{K}$ such that, for all sequences $\{q_k\}_{k\ge 0}$,*

$$\|z_k - z^e\| \le \beta(\|z_0 - z^e\|, k) + \gamma\Big(\sup_{0\le j\le k} \|q_j\|\Big) \qquad \forall\, k \in \mathbb{N}, \quad \forall\, z_0 \in \mathcal{B}_\delta(z^e). \tag{5}$$

The ISS estimate (5) ensures that the state remains bounded for bounded inputs and recovers asymptotic stability when $q_k \equiv 0$ [SW96, KB23]. We adopt the additive perturbation structure in (4), which is sufficient for our analysis and simplifies the ISS arguments. A convenient way to verify ISS is via Lyapunov functions. The following theorems show that (4) inherits the transient decay of (2) and converges to a neighbourhood of its equilibria, whenever the perturbations are uniformly bounded and the unperturbed system is exponentially or asymptotically stable.

Theorem 1. *Consider the system* (4) *with equilibrium $z^e \in \mathbb{R}^n$ for $q_k = 0$ for $k \in \mathbb{N}$, and suppose there exist a locally Lipschitz function $V : \mathbb{R}^n \to \mathbb{R}_{\ge 0}$, constants $c \in (0, 1)$ and $R > 0$, and constants $c_1, c_2 \in \mathbb{R}_{>0}$, such that*

$$c_1\|z - z^e\|^2 \le V(z) \le c_2\|z - z^e\|^2, \qquad \forall\, z \in \mathcal{B}_R(z^e) \tag{6}$$

$$V(\bar{w}(z, 0)) - V(z) \le -cV(z), \qquad \forall\, z \in \mathcal{B}_R(z^e) \tag{7}$$

whenever $q_k = 0$ for all $k \in \mathbb{N}$. Then there exist a class-$\mathcal{K}$ function γ and radii $\bar{r}, \bar{q} > 0$ such that, for the perturbed system (4) *with* $\sup_k \|q_k\| \le \bar{q}$ *and initial condition satisfying $z_0 \in \mathcal{B}_{\bar{r}}(z^e)$, the corresponding solution satisfies*

$$\|z_k - z^e\| \le \sqrt{\frac{c_2}{c_1}}\|z_0 - z^e\|(1-c)^{\frac{k}{2}} + \gamma\Big(\sup_{0\le j\le k}\|q_j\|\Big), \qquad k \ge 0. \tag{8}$$

In particular, z^e is locally exponentially stable for $q_k = 0$, $k \in \mathbb{N}$, and locally ISS with respect to perturbations $q_k \in \mathcal{B}_{\bar{q}}(0)$, $k \in \mathbb{N}$.

The function V in Theorem 1 is called Lyapunov function. A proof of Theorem 1 is given in Appendix B. The above theorem shows that the perturbed system exhibits the same local exponential decay rate as the unperturbed system and converges to a neighbourhood of the equilibrium of the unperturbed system, where the size of the neighbourhood depends on the supremum of the norm of the perturbations. For extension to the setting where the unperturbed system is asymptotically stable but not exponentially, we first present the following lemma.

Lemma 1. *Let $\varphi\colon [0,\infty) \to [0,\infty)$ be continuous, strictly increasing, with $\varphi(0) = 0$, and $\mathrm{Id} - \varphi$ a class-$\mathcal{K}$ function. Let $\{v_k\}_{k\ge 0}$ be a non-negative sequence satisfying*

$$v_{k+1} \le \varphi(v_k) + d, \qquad \forall\, k \ge 0,$$

for some constant $d \in [0, b)$, where $b = \lim_{r\to\infty} r - \varphi(r)$. Then the following properties are satisfied.

(i) There exists a class-$\mathcal{KL}$ function $\hat{\beta}$ such that, if $d = 0$, then $v_k \le \hat{\beta}(v_0, k)$ for all $k \ge 0$.
(ii) For every $\varepsilon > 0$, there exists $\bar{d}(\varepsilon) > 0$ such that, if $d \le \bar{d}(\varepsilon)$, then $\limsup_{k\to\infty} v_k \le \varepsilon$.
(iii) For every $R > 0$, there exists a class-$\mathcal{K}$ function $\hat{\gamma}_R$ such that, with $\hat{\beta}$ as in (i)*,*

$$v_k \le \hat{\beta}(v_0, k) + \hat{\gamma}_R(d), \qquad \forall\, k \ge 0, \quad \forall\, v_0 \in [0, R], \quad \forall\, d \in [0, b).$$

Theorem 2. *Consider the system* (4) *with equilibrium $z^e \in \mathbb{R}^n$ for $q_k = 0$ for $k \in \mathbb{N}$, and suppose there exist a locally Lipschitz function $V : \mathbb{R}^n \to \mathbb{R}_{\ge 0}$, class-$\mathcal{K}_\infty$ functions $\alpha_1, \alpha_2, \alpha_3$, and a constant $R > 0$ such that*

$$\alpha_1(\|z - z^e\|) \le V(z) \le \alpha_2(\|z - z^e\|), \qquad \forall\, z \in \mathcal{B}_R(z^e), \tag{9}$$
$$V(\bar{w}(z,0)) - V(z) \le -\alpha_3(\|z - z^e\|), \qquad \forall\, z \in \mathcal{B}_R(z^e). \tag{10}$$

Then there exist a class-$\mathcal{KL}$ function β, a class-$\mathcal{K}$ function γ, and radii $\bar{r}, \bar{q} > 0$ such that, for the perturbed system (4) *with $\sup_k \|q_k\| \le \bar{q}$ and $z_0 \in \mathcal{B}_{\bar{r}}(z^e)$, the solution satisfies*

$$\|z_k - z^e\| \le \beta(\|z_0 - z^e\|, k) + \gamma\Big(\sup_{0\le j\le k}\|q_j\|\Big), \qquad k \ge 0. \tag{11}$$

Proofs of Lemma 1 and Theorem 2 are given in Appendix B. Together with Theorem 1, they show that the perturbed system converges to a neighbourhood of the equilibrium of the unperturbed system with the same transient decay profile, without any extra dependence on the iteration index k, where the neighbourhood size is determined by the perturbation bound. Unlike classical ISS results, which certify stability of the perturbed system, Theorems 1–2 guarantee that it keeps the decay profile of the unperturbed system, with γ uniform in k. The equilibrium z^e is arbitrary, and its uniqueness is not needed since the results are local, and existence holds in all instantiations below. This provides a general mechanism for transferring convergence guarantees from FO algorithms to their ZO counterparts. Whenever an FO algorithm is known to converge, one only needs to verify that the averaged ZO dynamics can be written in the form (4) with bounded perturbations, and the ISS machinery automatically guarantees convergence to a neighbourhood of the same fixed point, inheriting the same decay profile. In Section 3, we analyse conditions under which such a formulation is possible. Crucially, since the perturbations originate from the ZO approximation rather than a model mismatch, disturbances or sensor noise, their magnitude can be manipulated through appropriate parameter selections.

2.2 Gaussian smoothing

Following [NS17], we define the Gaussian-smoothed version of f with parameter $\mu > 0$ as $f_\mu(x) = E_u[f(x + \mu u)]$, where $u \sim \mathcal{N}(0, I)$. Regardless of whether f is differentiable, f_μ is always

differentiable with $\nabla f_\mu(x) = E_u\left[\frac{f(x+\mu u)}{\mu} u\right]$ [NS17]. We use the two-point random oracle

$$g_\mu(x) = \tfrac{1}{\mu}(f(x+\mu u) - f(x))\, u, \tag{12}$$

which satisfies $E_u[g_\mu(x)] = \nabla f_\mu(x)$. We call an algorithm a Gaussian ZO method if it uses (12) in place of gradients.

Assumption 1. *The variance of the oracle g_μ defined in* (12) *is uniformly upper bounded by $\sigma^2 \geq 0$; i.e., $E_u[\|g_\mu(x) - \nabla f_\mu(x)\|^2] \leq \sigma^2$.*

This is a standard assumption; see, e.g., [MMSN21, LLC+20, FPB+25a]. The variance can be reduced by averaging t independent oracle evaluations per iteration, yielding $E_u\left[\|g_\mu(x_k) - \nabla f_\mu(x_k)\|^2\right] \leq \sigma^2/t$ while preserving unbiasedness [BG22]. We note that since our convergence results are local (Theorems 1 and 2 hold on a ball $B_{\bar{r}}(z^e)$), Assumption 1 needs only to hold on this compact region. A relaxation to state-dependent variance is analysed in Remark 4.

3 ZO methods as perturbed FO methods

Without further assumptions, there is no guarantee that the expectation of a Gaussian ZO method can be formulated as a perturbed FO method with characterisable bounded perturbations. Here, we identify conditions under which such a formulation is possible. Then, Theorems 1 and 2 imply that the convergence of the Gaussian ZO method has no extra dependence on the problem dimension and follows the same transient decay as the FO method. The proofs of all the lemmas and theorems of this section are given in Appendix C.

3.1 Gradient descent and strong convexity

We focus on the GD algorithm and assume that f is strongly convex. The GD update rule is

$$z_{k+1} = z_k - h\nabla f(z_k), \tag{13}$$

where $h > 0$ is the step size. Suppose f has Lipschitz gradients with constant $L_1 > 0$. The ZO counterpart of GD is

$$\tilde{z}_{k+1} = \tilde{z}_k - h g_\mu(\tilde{z}_k). \tag{14}$$

From [NS17], we know that $E[g_\mu] = \nabla f_\mu$. Let $U_k = \{u_1, \ldots, u_k\}$ and define the averaged (deterministic) iterate of ZO-GD and its update rule as

$$\bar{z}_k = E_{U_k}[\tilde{z}_k], \qquad \bar{z}_{k+1} = \bar{z}_k - h E_{U_k}[\nabla f_\mu(\tilde{z}_k)]. \tag{15}$$

This can be reformulated as

$$\bar{z}_{k+1} = \bar{z}_k - h\nabla f(\bar{z}_k) + q_k, \tag{16}$$

where

$$q_k = h\big(\nabla f(\bar{z}_k) - E_{U_k}[\nabla f_\mu(\tilde{z}_k)]\big) = h\big(\nabla f(\bar{z}_k) - \nabla f_\mu(\bar{z}_k)\big) + h\big(\nabla f_\mu(\bar{z}_k) - E_{U_k}[\nabla f_\mu(\tilde{z}_k)]\big). \tag{17}$$

Lemma 2. *Consider* (13), (16)*, and* (17)*. Let f be β-strongly convex with Lipschitz gradients with constant $L_1 > 0$, and let $h \in \big(0, \frac{\beta}{2L_1^2}\big)$. Then*

$$\sup_k \|q_k\| \leq h\frac{\mu}{2} L_1 (n+3)^{3/2} + hL_1\sqrt{\frac{2h\sigma^2}{t(\beta - 2L_1^2 h)}}. \tag{18}$$

Lemma 2 shows that the averaged ZO-GD iterates satisfy a perturbed version of the GD dynamics, where the perturbations are bounded and their magnitude can be manipulated through appropriate parameter selection. It is known that, for a smooth strongly convex objective, GD converges to the minimiser exponentially fast [Nes18]. In [NS17], ZO-GD was analysed under the same setting, and it was shown that the iteration complexity carries an extra dimension-dependent factor for any choice of parameters. The following theorem shows that, instead, the dimension affects only the size of the convergence neighbourhood, and there are parameter choices such that ZO-GD closely follows GD.

Theorem 3. *Consider* (13) *and* (16). *Let f be β-strongly convex with Lipschitz gradients with constant $L_1 > 0$, let $h \in \big(0, \frac{\beta}{2L_1^2}\big)$ and for $r > 0$ let $\mathcal{B}_r(z^e)$ denote a convergence neighbourhood around the fixed point z^e of GD. Then there exist constants $\bar{\mu}, \bar{q} > 0$ and $\bar{t} \in \mathbb{N}$ (depending on h, β, L_1, and the Lyapunov function of the GD dynamics) such that, for any $\mu \in (0, \bar{\mu})$ and $t \geq \bar{t}$ satisfying $\sup_k \|q_k\| \leq \bar{q}$ via the bound in Lemma 2, ZO-GD has the same exponential decay rate as GD converges to $\mathcal{B}_r(z^e)$, and there exists a class-$\mathcal{K}$ function α such that*

$$r = \alpha\left(h\frac{\mu}{2}L_1(n+3)^{3/2} + hL_1\sqrt{\frac{2h\sigma^2}{t(\beta - 2L_1^2 h)}}\right). \tag{19}$$

Remark 1 (Single iterate)**.** *Since $\bar{z}_k = E[\tilde{z}_k]$, we have $E[\|\tilde{z}_k - z^*\|^2] = \|\bar{z}_k - z^*\|^2 + E[\|\tilde{z}_k - \bar{z}_k\|^2]$, where the last term is at most $\frac{2h\sigma^2}{t(\beta - 2L_1^2 h)}$ by* (55)*. Hence the single ZO iterate inherits the GD decay rate and converges to a slightly larger neighbourhood in expectation, still controlled by h, μ, and t.*

3.2 Heavy ball method and Nesterov's accelerated gradient under strong convexity

In this section, we consider both the heavy ball (HB) method and Nesterov's accelerated gradient (NAG) under strong convexity. The two algorithms share the same augmented state-space structure and differ only in the point at which the gradient is evaluated, so we treat them in a unified framework.

The HB and NAG update rules are, respectively,

$$x_{k+1} = x_k + h_2(x_k - x_{k-1}) - h_1\nabla f(x_k), \tag{20}$$

$$x_{k+1} = x_k + h_2(x_k - x_{k-1}) - h_1\nabla f\big(x_k + h_2(x_k - x_{k-1})\big), \tag{21}$$

and their ZO counterparts (ZO-HB, ZO-NAG) replace ∇f with g_μ defined in (12). Define the augmented state vector $z_k = \begin{bmatrix} x_{k-1}^\top & x_k^\top \end{bmatrix}^\top \in \mathbb{R}^{2n}$, the state transition matrix

$$A = \begin{bmatrix} 0 & I_n \\ -h_2 I_n & (1+h_2)I_n \end{bmatrix},$$

and the output map $T = [0 \;\; I_n] \in \mathbb{R}^{n \times 2n}$, so that $x_k = Tz_k$. Note that $TAz_k = x_k + h_2(x_k - x_{k-1})$ is the lookahead point used by NAG. Both algorithms can be written compactly as

$$z_{k+1} = Az_k + \begin{bmatrix} 0 \\ -h_1\nabla f(Pz_k) \end{bmatrix}, \tag{22}$$

where $P = T$ for HB and $P = TA$ for NAG. The ZO counterpart replaces $\nabla f(Pz_k)$ with $g_\mu(Pz_k)$.

Following the same steps as in Section 3.1, conditioning on z_k, taking expectations, and adding and subtracting the gradient evaluated at the averaged state, we obtain the averaged dynamics

$$\bar{z}_{k+1} = A\bar{z}_k + \begin{bmatrix} 0 \\ -h_1\nabla f(P\bar{z}_k) \end{bmatrix} + q_k, \tag{23}$$

where the perturbation term is

$$q_k = \begin{bmatrix} 0 \\ -h_1\Big(E_{U_k}[\nabla f_\mu(Pz_k)] - E_{U_k}[\nabla f(Pz_k)] + E_{U_k}[\nabla f(Pz_k)] - \nabla f(P\bar{z}_k)\Big) \end{bmatrix}. \tag{24}$$

Equation (23) shows that the averaged trajectory follows the FO dynamical system (HB or NAG, depending on the choice of P) perturbed by q_k.

Lemma 3. *Consider* (22)–(24) *with $P = T$ (HB). Let f be β-strongly convex with L_1-Lipschitz gradients, and let $h_1, h_2 > 0$ satisfy*

$$\rho_{\mathrm{HB}} := (1 + 2h_2)^2 - h_1(1+h_2)\beta + 4h_1h_2L_1 + 2h_1^2L_1^2 < 1. \tag{25}$$

Then

$$\sup_k \|q_k\| \leq \frac{h_1\mu L_1(n+3)^{3/2}}{2} + \frac{h_1^2 L_1\sigma}{\sqrt{t}}\sqrt{\frac{2}{1-\rho_{\mathrm{HB}}}}. \tag{26}$$

A sufficient condition for (25) *is $h_2 \in (0, ch_1)$ and $h_1 \in \big(0, \frac{\beta - 4c}{4c^2 + 4cL_1 + 2L_1^2}\big)$ with $c \in \big(0, \frac{\beta}{4}\big)$.*

Lemma 4. *Consider* (22)–(24) *with $P = TA$ (NAG). Let f be β-strongly convex with L_1-Lipschitz gradients, and let $h_1, h_2 > 0$ satisfy*

$$\rho_{\mathrm{NAG}} := (1 - h_1\beta + 2h_1^2L_1^2)(1 + 2h_2)^2 < 1. \tag{27}$$

Then

$$\sup_k \|q_k\| \le \frac{h_1\mu L_1(n+3)^{3/2}}{2} + \frac{(1+2h_2)h_1^2L_1\sigma}{\sqrt{t}}\sqrt{\frac{2}{1-\rho_{\mathrm{NAG}}}}\,. \tag{28}$$

A sufficient condition for (27) *is $h_1 \in \big(0, \frac{\beta}{2L_1^2}\big)$ and $h_2 \in \big(0, \frac{1}{2}\big(\frac{1}{\sqrt{1-h_1\beta+2h_1^2L_1^2}} - 1\big)\big)$.*

Both lemmas show that the averaged ZO iterates satisfy a perturbed version of their FO dynamics with bounded, controllable perturbations. The key structural difference is that, for NAG, the strong convexity contraction acts directly on the lookahead error, eliminating the momentum cross term that appears in the HB analysis. For HB, it is well known that the method locally converges exponentially fast [Pol64]; for NAG, global convergence with an accelerated rate is established in [Nes18]. The following theorems provide a unified perspective that for both methods, the dimension dependence does not appear in the iteration complexity, instead, the convergence is to a neighbourhood of FO dynamics, where the size of the neighbourhood can be made arbitrarily small.

Theorem 4. *Consider* (22) *and* (23) *with $P = T$. Let f be β-strongly convex with L_1-Lipschitz gradients, let the hypotheses of Lemma 3 be satisfied, and for $r > 0$ let $\mathcal{B}_r(z^e)$ denote a convergence neighbourhood around the fixed point z^e of HB. Then there exist positive scalars $\bar{\mu}$ and $\bar{q}$, and $\bar{t} \in \mathbb{N}$ such that, for any $\mu \in (0, \bar{\mu})$ and $t \ge \bar{t}$ satisfying $\sup_k \|q_k\| \le \bar{q}$ via the bound in Lemma 3, ZO-HB has the same exponential decay rate as HB and converges to $\mathcal{B}_r(z^e)$, and there exists a class-$\mathcal{K}$ function α such that*

$$r = \alpha\Big(\frac{h_1\mu L_1(n+3)^{3/2}}{2} + \frac{h_1^2L_1\sigma}{\sqrt{t}}\sqrt{\frac{2}{1-\rho_{\mathrm{HB}}}}\Big). \tag{29}$$

Theorem 5. *Consider* (22) *and* (23) *with $P = TA$. Let f be β-strongly convex with L_1-Lipschitz gradients, let the hypotheses of Lemma 4 be satisfied, and for $r > 0$ let $\mathcal{B}_r(z^e)$ denote a convergence neighbourhood around the fixed point z^e of NAG. Then there exist $\bar{\mu}, \bar{q} > 0$ and $\bar{t} \in \mathbb{N}$ such that, for any $\mu \in (0, \bar{\mu})$ and $t \ge \bar{t}$ satisfying $\sup_k E\|q_k\| \le \bar{q}$ via the bound in Lemma 4, ZO-NAG has the same decay rate as NAG and converges to $\mathcal{B}_r(z^e)$, and there exists a class-$\mathcal{K}$ function α such that*

$$r = \alpha\Big(\frac{h_1\mu L_1(n+3)^{3/2}}{2} + \frac{(1+2h_2)h_1^2L_1\sigma}{\sqrt{t}}\sqrt{\frac{2}{1-\rho_{\mathrm{NAG}}}}\Big). \tag{30}$$

3.3 L_2 regularisation

In the preceding sections, we showed that strong convexity is sufficient to guarantee that the averaged ZO dynamics can be written as a bounded perturbation of the FO dynamics. When the objective f is not strongly convex, one can introduce an L_2 regularisation term to induce the necessary contraction to recover the same perturbation framework, even when f itself is not strongly convex.

We therefore consider, in place of (1), the regularised problem

$$\min_{x\in\mathbb{R}^n} \; f(x) + \lambda\|x\|^2, \tag{31}$$

where $\lambda > 0$ is the regularisation parameter. We focus on the GD case; analogous results for HB and NAG are discussed in Remark 10. Setting $\lambda = \frac{1-c}{2h}$, where $h > 0$ is the step size and $c \in [0, 1)$ (thus λ can be arbitrary small but not larger than $\frac{1}{2h}$), the GD update (13) applied to (31) becomes

$$z_{k+1} = cz_k - h\nabla f(z_k), \tag{32}$$

the ZO-GD update (14) becomes

$$\tilde{z}_{k+1} = c\tilde{z}_k - hg_\mu(\tilde{z}_k), \tag{33}$$

and the averaged ZO update is

$$\bar{z}_{k+1} = c\bar{z}_k - h\nabla f(\bar{z}_k) + q_k, \tag{34}$$

where

$$q_k = h\big(\nabla f(\bar{z}_k) - E_{U_k}[\nabla f_\mu(\tilde{z}_k)]\big) = h\big(\nabla f(\bar{z}_k) - \nabla f_\mu(\bar{z}_k)\big) + h\big(\nabla f_\mu(\bar{z}_k) - E_{U_k}[\nabla f_\mu(\tilde{z}_k)]\big). \tag{35}$$

Lemma 5. *Consider* (32)*,* (34)*, and* (35)*. Let f be a possibly non-strongly-convex function with Lipschitz gradients with constant $L_1 > 0$, and let $h \in \big(0, \frac{\sqrt{2(1+c^2)}-2c}{2L_1}\big)$. Then*

$$\sup_k \|q_k\| \leq h\frac{\mu}{2}L_1(n+3)^{3/2} + hL_1\sqrt{\frac{2h^2\sigma^2}{t(1-c^2-4cL_1h-2L_1^2h^2)}}. \tag{36}$$

Lemma 5 shows that the contraction induced by the regularisation term ensures the averaged ZO-GD iterates satisfy a perturbed version of the GD dynamics, where the perturbations are bounded and controllable through the algorithm parameters. The following theorem establishes that, whenever GD converges to a fixed point, ZO-GD inherits the same convergence rate with no extra dimension dependence in the iteration complexity; the dimension affects only the size of the convergence neighbourhood.

Theorem 6. *Consider* (32) *and* (34)*. Let f be a possibly non-strongly-convex function with Lipschitz gradients with constant $L_1 > 0$, and let $h \in \big(0, \frac{\sqrt{2(1+c^2)}-2c}{2L_1}\big)$, and for $r > 0$ let $\mathcal{B}_r(z^e)$ denote a convergence neighbourhood around the fixed point z^e of GD. Suppose that GD applied to the regularised problem* (31) *converges to a fixed point z^e. Then there exist constants $\bar{\mu}, \bar{q} > 0$ and $\bar{t} \in \mathbb{N}$ (depending on h, c, L_1, and the Lyapunov function of the GD dynamics) such that, for any $\mu \in (0, \bar{\mu})$ and $t \geq \bar{t}$ satisfying $\sup_k \|q_k\| \leq \bar{q}$ via the bound in Lemma 5, ZO-GD has the same decay rate as GD and converges to $\mathcal{B}_r(z^e)$ and there exists a class-$\mathcal{K}$ function α such that*

$$r = \alpha\left(h\frac{\mu}{2}L_1(n+3)^{3/2} + hL_1\sqrt{\frac{2h^2\sigma^2}{t(1-c^2-4cL_1h-2L_1^2h^2)}}\right). \tag{37}$$

The admissible parameter range in Lemma 5 implies that the regularised objective is strongly convex (Remark 9). Nevertheless, the framework provides a systematic way to apply ISS to originally non-convex objectives with explicit control over the trade-off between proximity to the original minimiser and neighbourhood size. Next, we illustrate the theoretical findings via numerical examples.

4 Numerical examples

We illustrate the main results through two settings: a strongly convex quadratic (Section 4.1) and a non-convex neural network with L_2 regularisation (Section 4.2). Extended parameter studies are provided in Appendices E.1 and E.2. Both examples have cheap gradients and are chosen to validate the theory, not to advocate ZO where FO is available.

4.1 Quadratic objective

We consider $\min_{x\in\mathbb{R}^n} x^\top Ax$, where $n = 1000$ and A is positive definite with condition number 100 and maximum eigenvalue 100, comparing GD and ZO-GD (averaged over 5 runs). Figure 1 shows the effect of varying $h \in \{10^{-4}, 10^{-5}, 10^{-7}\}$ with $\mu = 10^{-6}$ and $t = 1$ fixed. When h is too large, the perturbation exceeds the admissible bound of Theorem 1 and ZO-GD diverges. As h decreases, ZO-GD converges to a progressively smaller neighbourhood of the GD trajectory, with no extra dimension dependence. Analogous sweeps over μ and t, as well as parameter compensation experiments, are presented in Appendix E.1. In all cases, the observations are consistent with the bound in Lemma 2: h and μ control the smoothing bias, while h and t control the variance term.

4.2 Binary classification with a neural network

We consider binary classification on MNIST (digits 0 vs 1) using a two-layer fully connected network with ReLU activations and sigmoid output ($d = 784$, $H = 128$ (hidden layer neurons), $n = 100{,}609$ parameters). We minimise the L_2-regularised logistic loss with $\lambda = 10^{-2}$ and $M = 2000$ training samples. Since the logistic loss composed with a neural network is non-convex, this example falls in the framework of Section 3.3.

Convergence comparison. Figure 2 shows the training loss for all six methods (GD, HB, NAG and their ZO counterparts). Shaded regions indicate the standard deviation across 3 independent

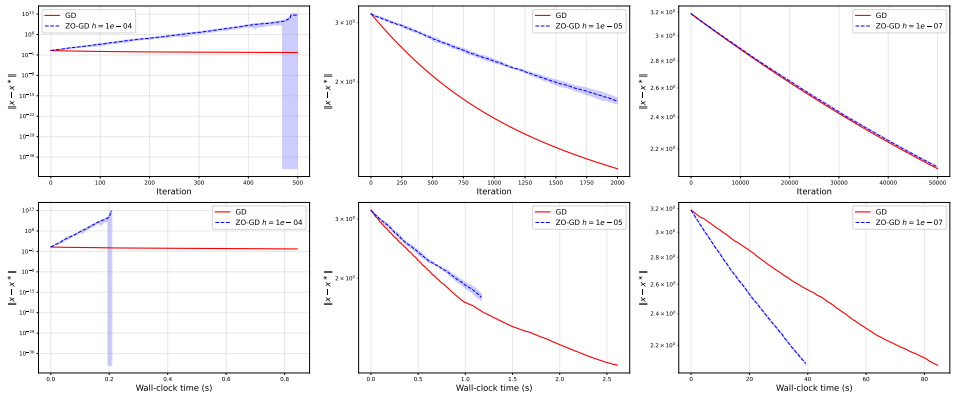

Figure 1: Quadratic objective ($n = 1000$): effect of the step size h on GD (solid) and ZO-GD (dashed). Reducing h shrinks the perturbation and brings ZO-GD closer to GD.

runs, and the small width confirms a small perturbation. We set $h = 1 \times 10^{-3}$ for GD and ZO-GD, $h_1 = 1 \times 10^{-3}$, $h_2 = 5 \times 10^{-4}$ for HB/NAG and their ZO counterparts with smoothing parameter $\mu = 10^{-7}$ and number of sampled directions $t = 7$. The ZO trajectories closely track their FO counterparts, and all methods reach 99.55% training accuracy and 99.40% test accuracy. It takes approximately 4 times longer to run the ZO methods compared to their FO counterpart. Parameter sensitivity experiments confirming the same phenomena as in the quadratic case are provided in Appendix E.2.

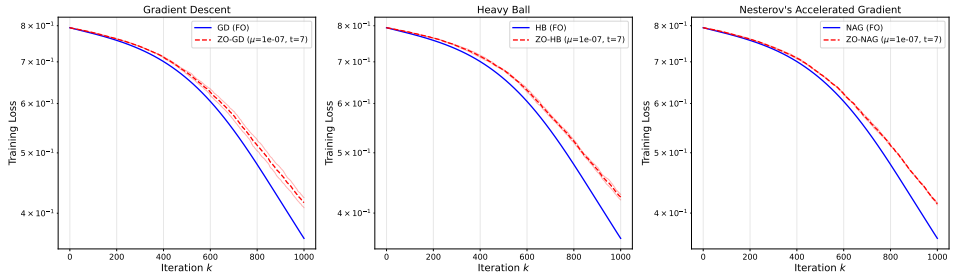

Figure 2: Training loss for GD, HB, NAG and their ZO counterparts on a two-layer neural network ($n = 100{,}609$). Solid: FO; dashed: ZO. All methods achieve 99.55% train / 99.40% test accuracy.

Dimension scaling. To directly validate the absence of extra dimension dependence, we vary $H \in \{16, 32, 64, 128, 256\}$ (n from ≈12,700 to ≈201,000) and run GD and ZO-GD with identical parameters for 1000 iterations. We set $h = 5 \times 10^{-4}$, $t = 10$, and $\mu = 10^{-8}$. Figure 3 shows that ZO-GD tracks GD at the same rate across all dimensions. Table 1 reports iteration counts to reach within 20% of the FO final loss: the ratio ZO/FO stays close to 1. Moreover, Table 1 confirms the per-iteration time ratio remains approximately constant.

Table 1: Dimension scaling: iterations to reach within 20% of FO final loss.

H	n	FO final	ZO final	Iter ZO/FO	Time ZO/FO
16	12,577	0.6922	0.6920	0.99	5.5
32	25,153	0.6720	0.6721	1	4.9
64	50,305	0.6761	0.6798	1.03	5.4
128	100,609	0.6529	0.6584	1.03	5.7
256	201,217	0.6866	0.7017	1.05	4.5

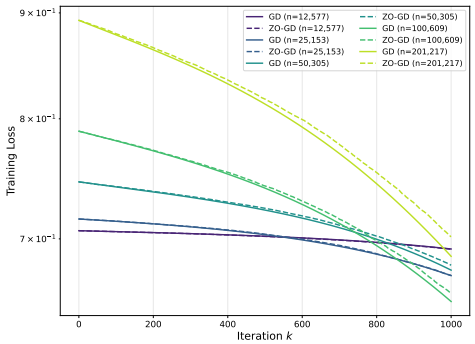

Figure 3: Dimension scaling: Training loss for GD (solid) and ZO-GD (dashed) across five network sizes.

5 Conclusion and future directions

In this work, we revisited the commonly held belief that ZO methods inherently suffer from additional dimension dependence in their iteration complexity compared to their FO counterparts. By adopting a dynamical systems perspective, we showed that, under suitable conditions, the averaged dynamics of ZO algorithms can be interpreted as perturbed versions of their FO counterparts with bounded, controllable perturbations. Leveraging ISS arguments, we proved that ZO methods inherit the same convergence rates as FO methods in expectation, converging to a neighbourhood of the FO fixed point whose size can be made arbitrarily small by tuning the algorithm parameters. These findings suggest that the practical efficiency of ZO methods may be closer to that of FO methods than previously understood, particularly when gradient information is unavailable or expensive.

Limitations and future work. Our results come with a clear scope. First, the guarantees are local and hold in expectation. Remark 1 extends them to the single iterate in mean square, but high-probability bounds require concentration arguments beyond the present analysis. Second, the framework needs contraction on the iterates to keep the perturbation bounded. Here, this is supplied by strong convexity or L_2 regularisation, and over the admissible range the regularised objective is itself strongly convex. Genuinely non-convex objectives, sparsity-promoting regularisers such as L_1, and many motivating applications in their native form therefore lie outside the current theory, even though our experiments suggest the behaviour persists there. Third, while the dimension no longer slows the decay rate, it still shapes the neighbourhood radius, and the number of directions t needed for a target radius scales with σ^2, which may grow with n in the worst case. The price of ZO is thus paid in queries per iteration rather than in iterations, but these queries can run in parallel as the problem is no longer sequential. Natural next steps are to derive contraction from weaker conditions such as incremental stability, to establish high-probability guarantees rather than results in expectation, to handle non-additive perturbations $z^+ = w(z, q)$ and non-smooth objectives, and to treat ZO error as a disturbance to be rejected, using tools from control design to synthesise new ZO algorithms with provable robustness.

Acknowledgments and Disclosure of Funding

This work was supported by the Australian Research Council through a Discovery Project under Grant DP250101763 and the United States Air Force Office of Scientific Research under Grant FA2386-24-1-4014.

References

[BG18] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates. In *Advances in Neural Information Processing Systems*, volume 31, pages 3455–3464, 2018.

[BG22] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, 22(1):35–76, 2022.

[CJSB25] Leilei Cui, Zhong-Ping Jiang, Eduardo D Sontag, and Richard D Braatz. Perturbed gradient descent algorithms are small-disturbance input-to-state stable. *arXiv preprint arXiv:2507.02131*, 2025.

[CRS+18] Krzysztof Choromanski, Mark Rowland, Vikas Sindhwani, Richard Turner, and Adrian Weller. Structured evolution with compact architectures for scalable policy optimization. In *International Conference on Machine Learning*, pages 970–978. PMLR, 2018.

[CSV09] Andrew R Conn, Katya Scheinberg, and Luis N Vicente. *Introduction to derivative-free optimization*. SIAM, 2009.

[CZS+17] 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.

[DGG21] Pavel Dvurechensky, Eduard Gorbunov, and Alexander Gasnikov. An accelerated directional derivative method for smooth stochastic convex optimization. *European Journal of Operational Research*, 288(2):601–621, 2021.

[DJWW15] John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

[DP18] Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

[FPB+25a] Amir Ali Farzin, Yuen-Man Pun, Philipp Braun, Antoine Lesage-Landry, Youssef Diouane, and Iman Shames. Min-max optimisation for nonconvex-nonconcave functions using a random zeroth-order extragradient algorithm. *Transactions on Machine Learning Research*, 2025.

[FPB+25b] Amir Ali Farzin, Yuen-Man Pun, Philipp Braun, Tyler Summers, and Iman Shames. Minimisation of submodular functions using gaussian zeroth-order random oracles. *arXiv preprint arXiv:2510.15257*, 2025.

[FPB+26] Amir Ali Farzin, Yuen-Man Pun, Philipp Braun, Tyler Summers, and Iman Shames. Solving the offline and online min-max problem of non-smooth submodular-concave functions: A zeroth-order approach. *arXiv preprint arXiv:2601.21243*, 2026.

[FPBS25a] Amir Ali Farzin, Yuen-Man Pun, Philipp Braun, and Iman Shames. Minimisation of quasar-convex functions using random zeroth-order oracles. *arXiv preprint arXiv:2505.02281*, 2025.

[FPBS25b] Amir Ali Farzin, Yuen-Man Pun, Philipp Braun, and Iman Shames. Properties of fixed points of generalised extra gradient methods applied to min-max problems. *IEEE Control Systems Letters*, 2025.

[FPBS26] Amir Ali Farzin, Yuen-Man Pun, Philipp Braun, and Iman Shames. On the stability connection between discrete-time algorithms and their resolution odes: Applications to min-max optimisation. *arXiv preprint arXiv:2603.01430*, 2026.

[FS24] Amir Ali Farzin and Iman Shames. Minimisation of polyak-łojasewicz functions using random zeroth-order oracles. In *2024 European Control Conference (ECC)*, pages 3207–3212. IEEE, 2024.

[GL13] Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[HJ12] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge University Press, 2012.

[JW01] Zhong-Ping Jiang and Yuan Wang. Input-to-state stability for discrete-time nonlinear systems. *Automatica*, 37(6):857–869, 2001.

[KB23] Christopher M Kellett and Philipp Braun. *Introduction to Nonlinear Control: Stability, Control Design, and Estimation*. Princeton University Press, 2023.

[Kel14] Christopher M Kellett. A compendium of comparison function results. *Mathematics of Control, Signals, and Systems*, 26(3):339–374, 2014.

[KMR+23] David Kozak, Cesare Molinari, Lorenzo Rosasco, Luis Tenorio, and Silvia Villa. Zeroth-order optimization with orthogonal random directions: D. kozak et al. *Mathematical Programming*, 199(1):1179–1219, 2023.

[LKC+18] Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Paishun Ting, Shiyu Chang, and Lisa Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

[LLC+20] Sijia Liu, Songtao Lu, Xiangyi Chen, Yao Feng, Kaidi Xu, Abdullah Al-Dujaili, Mingyi Hong, and Una-May O'Reilly. Min-max optimization without gradients: Convergence and applications to black-box evasion and poisoning attacks. In *International Conference on Machine Learning*, pages 6282–6293. PMLR, 2020.

[LRP16] Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. In *SIAM Journal on Optimization*, volume 26, pages 57–95. SIAM, 2016.

[MGN+23] Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems*, 36:53038–53075, 2023.

[MMS+18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[MMSN21] Alejandro I Maass, Chris Manzie, Iman Shames, and Hayato Nakada. Zeroth-order optimization on subsets of symmetric matrices with application to MPC tuning. *IEEE Transactions on Control Systems Technology*, 30(4):1654–1667, 2021.

[Nes18] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[NS17] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

[Pol64] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[SHC+17] 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.

[SLA12] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, 25, 2012.

[Son89] Eduardo D. Sontag. Smooth stabilization implies coprime factorization. *IEEE Transactions on Automatic Control*, 34(4):435–443, 1989.

[Son08] Eduardo D Sontag. Input to state stability: Basic concepts and results. In *Nonlinear and optimal control theory: lectures given at the CIME summer school held in Cetraro, Italy June 19–29, 2004*, pages 163–220. Springer, 2008.

[SW96] Eduardo D Sontag and Yuan Wang. New characterizations of input-to-state stability. *IEEE transactions on automatic control*, 41(9):1283–1294, 1996.

[WBMR20] Zhongruo Wang, Krishnakumar Balasubramanian, Shiqian Ma, and Meisam Razaviyayn. Zeroth-order algorithms for nonconvex minimax problems with improved complexities. *arXiv preprint arXiv:2001.07819*, 2020.

[WDBS18] Yining Wang, Simon S. Du, Sivaraman Balakrishnan, and Aarti Singh. Stochastic zeroth-order optimization in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 1356–1365. PMLR, 2018.

[WRJ21] Ashia C. Wilson, Benjamin Recht, and Michael I. Jordan. A lyapunov analysis of accelerated methods in optimization. *Journal of Machine Learning Research*, 22(113):1–34, 2021.

[YHF+18] 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.

[YYFL23] Pengyun Yue, Long Yang, Cong Fang, and Zhouchen Lin. Zeroth-order optimization with weak dimension dependency. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4429–4472. PMLR, 2023.

[ZLH+24] Yihua Zhang, Pingzhi Li, Junyuan Hong, Jiaxiang Li, Yimeng Zhang, Wenqing Zheng, Pin-Yu Chen, Jason Lee, Wotao Yin, Mingyi Hong, et al. Revisiting zeroth-order optimization for memory-efficient llm fine-tuning: A benchmark. In *International Conference on Machine Learning*, 2024.

A Related works and Basic definitions

In this section, we provide a detailed comparison of our results with prior work on zeroth-order optimisation, followed by the basic definitions used throughout the paper.

A.1 Comparison with prior work

Table 2 summarises the convergence guarantees of representative ZO methods and highlights how our framework differs from existing analyses. In the works of Table 2, the iteration complexity carries an explicit factor depending on the problem dimension n (or an effective dimension n_{eff}), and convergence is to the exact optimiser. By contrast, our ISS-based framework eliminates the dimension factor from the iteration complexity entirely; the dimension dependence instead appears in the radius of the convergence neighbourhood, which can be made arbitrarily small by tuning the algorithm parameters μ, h, and t. Moreover, our framework applies uniformly to ZO-GD, ZO-HB, and ZO-NAG, whereas each prior result requires a separate, algorithm-specific analysis. We further note that in prior analyses such as [NS17], the prescribed step size for ZO-GD scales as $O(1/n)$, which is not consistent with standard FO gradient descent. By contrast, our framework permits the same step size as the FO method, with no dependence on the dimension (see Remark 2).

A.2 Basic definitions

We now recall the standard regularity and convexity conditions assumed throughout the paper.

Definition 3 (Lipschitz continuity and gradients)**.** *A continuous function $f : \mathbb{R}^n \to \mathbb{R}$ is globally Lipschitz with constant $L_0 > 0$ if $\|f(x) - f(y)\| \leq L_0\|x - y\|$ for all $x, y \in \mathbb{R}^n$. If f is C^1, the gradient of f is globally Lipschitz with constant $L_1 > 0$ if*

$$\|\nabla f(x) - \nabla f(y)\| \leq L_1\|x - y\|, \qquad \forall\, x, y \in \mathbb{R}^n. \tag{38}$$

Definition 4 (Strong convexity)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a C^1 function. Then f is said to be β-strongly convex with constant $\beta > 0$ if there exists a constant $\beta > 0$ such that*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2}\|x - y\|^2, \qquad \forall\, x, y \in \mathbb{R}^n. \tag{39}$$

Table 2: Comparison of convergence guarantees for ZO methods.

Reference	Assumptions on f	Iter. dim. dep.	Queries/iter.	Total query dim. dep.	Target
[NS17]	convex / s.c.	$O(n)$	2	$O(n)$	Minimiser
[DJWW15]	convex	$O(n)$	2	$O(n)$	Minimiser
[GL13]	nonconvex	$O(n)$	2^*	$O(n)$	$\|\nabla f\|^2 \le \varepsilon$
[YYFL23]	convex / s.c.; Hessian regularity	$O(n_{\text{eff}})$	2	$O(n_{\text{eff}})$	Minimiser
[WDBS18]	convex; sparse gradients	$O(s \log n)$	2	$O(s \log n)$	Minimiser
[KMR+23]	convex / PL	$O(n/\ell)$	$\ell + 1$	$O(n)$	Minimiser
This work	s.c. (Thms. 3–5) n.c. + L_2 reg. (Thm. 6)	**No explicit**	$2t$	worst case via t (σ^2)	Neighbourhood of FO fixed point

s.c. = strongly convex, n.c. = non-convex, 2^*: diminishing step sizes, ℓ: number of orthogonal directions.

B Additional details and proofs of Section 2

In this section, we present additional explanations and proofs of the results in Section 2.

Lemma 6 ([Kel14, Lem. 10])**.** *Let $\alpha : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ be a class-$\mathcal{K}$ function. Then for any $a, b \ge 0$ and any $\varepsilon > 0$,*

$$\alpha(a+b) \le \alpha((1+\varepsilon)a) + \alpha\big((1+\tfrac{1}{\varepsilon})b\big)\,. \tag{40}$$

Proof of Theorem 1. Let $R > 0$ be such that (6)–(7) hold on $\mathcal{B}_R(z^e)$. Since V is locally Lipschitz on $\mathcal{B}_R(z^e)$, there exist constants $L_V > 0$ satisfying

$$|V(x) - V(y)| \le L_V \|x - y\|$$

for all $x, y \in \mathcal{B}_R(z^e)$. Write $\lambda := 1 - c \in (0,1)$ and $\delta := \sup_{j \ge 0} \|q_j\|$. Let $\bar r, \bar q > 0$, let $z_0 \in \mathcal{B}_{\bar r}(z^e)$, let $\{z_k\}_{k\in\mathbb{N}}$ be defined through $z_{k+1} = w(z_k) + q_k$ for $k \in \mathbb{N}$ and $\{q_k\}_{k\in\mathbb{N}} \subset \mathcal{B}_{\bar q}(z^e)$, and let $\bar r$, $\bar q$ be defined such that $\{z_k\}_{k\in\mathbb{N}} \subset \mathcal{B}_{\bar r}(z^e)$. (The existence of $\bar r, \bar q > 0$ with this property will be shown later.) Using the Lipschitz property of V and (7), for the perturbed successor $z_{k+1} = w(z_k) + q_k$, the estimate

$$V(z_{k+1}) = V\big(w(z_k) + q_k\big) \le V(w(z_k)) + L_V \|q_k\| \le \lambda\, V(z_k) + L_V \|q_k\| \tag{41}$$

is obtained. Iterating (41) from $k = 0$ yields

$$V(z_k) \le \lambda^k V(z_0) + L_V \sum_{j=0}^{k-1} \lambda^{k-1-j} \|q_j\| \le \lambda^k V(z_0) + L_V \delta \sum_{j=0}^{k-1} \lambda^j \le \lambda^k V(z_0) + \frac{L_V}{c}\,\delta,$$

where the geometric series was bounded by $\sum_{j=0}^{\infty} \lambda^j = 1/c$. Applying $c_1 \|z_k - z^e\|^2 \le V(z_k)$ and the sandwich bound $V(z_0) \le c_2 \|z_0 - z^e\|^2$, we obtain

$$\|z_k - z^e\|^2 \le \lambda^k c_1^{-1}\, c_2 \|z_0 - z^e\|^2 + \tfrac{L_V}{c c_1}\,\delta$$

or

$$\|z_k - z^e\| \le \sqrt{\frac{c_2}{c_1}} \|z_0 - z^e\| \lambda^{\frac{k}{2}} + \gamma\Big(\sup_{0 \le j \le k} \|q_j\|\Big), \qquad k \ge 0,$$

where $\gamma(\delta) = \sqrt{\frac{L_v}{c c_1}}\sqrt{\delta}$, which is a $\mathcal{K}$ function.

It remains to choose $\bar r > 0$ and $\bar q > 0$ so that $\{z_k\}_{k\in\mathbb{N}} \subset \mathcal{B}_R(z^e)$. For $z \in \mathbb{R}^n$ such that $V(z) \le \ell = c_1 R^2$, (6) implies that $\|z - z^e\|^2 \le R^2$, i.e., $z \in \mathcal{B}_R(z^e)$, and we can conclude $\{z \in \mathbb{R}^n |\ V(z) \le c_1 R^2\} \subset \mathcal{B}_R(z^e)$. From (41), the set $\{z \in \mathbb{R}^n |\ V(z) \le \ell\}$ is forward-invariant for the perturbed system whenever $\lambda\,\ell + L_V \bar q \le \ell$, i.e. whenever $\bar q \le c\,\ell / L_V$. Setting

$$\bar r := \sqrt{\frac{\ell}{c_2}}, \qquad \bar q := \frac{c\,\ell}{L_V}$$

ensures that $\|z_0 - z^e\| \le \bar r$ implies $V(z_0) \le \ell$, and the sublevel set $\{z \in \mathbb{R}^n |\ V(z) \le \ell\}$ is forward-invariant under the perturbed dynamics, so the Lipschitz bounds and the decrease condition remain valid for all $k \ge 0$. This completes the proof. □

Proof of Lemma 1. **Item (i):** Since φ is continuous with $\varphi(0) = 0$ and $\varphi(s) < s$ for $s > 0$, the iterates $\varphi^{(k)}(s) := \varphi \circ \cdots \circ \varphi(s)$ (k times) satisfy $\varphi^{(k)}(s) \downarrow 0$ as $k \to \infty$ for each fixed $s > 0$. Define $\hat{\beta}(s, k) := \varphi^{(k)}(s)$. This is continuous, non-decreasing in s, decreasing to 0 in k, and $\hat{\beta}(0, k) = 0$; so it can be upper-bounded by a class-$\mathcal{KL}$ function, which we denote $\hat{\beta}$. This gives (i).

Item (ii): Fix $\varepsilon > 0$. Since $\varphi(\varepsilon) < \varepsilon$, set $\mu := \varepsilon - \varphi(\varepsilon) > 0$ and take $\bar{d}(\varepsilon) := \mu/2$. If $d \le \bar{d}(\varepsilon)$ and $v_k \le \varepsilon$, then $v_{k+1} \le \varphi(\varepsilon) + d \le \varphi(\varepsilon) + \mu/2 = \varepsilon - \mu/2 < \varepsilon$. This in particular implies that $\varphi(s) \in [0, \varepsilon]$ for all $s \in [0, \varepsilon]$. Meanwhile, if $v_k > \varepsilon$ we have $v_{k+1} \le \varphi(v_k) + d < v_k - \mu + d \le v_k - \mu/2$ (using $\mu \le s - \varphi(s)$ for $s \ge \varepsilon$ by continuity and compactness on $[\varepsilon, v_0]$). Hence, for $v_k > \varepsilon$, v_k decreases by at least $\mu/2$ per step and there exists $K \in \mathbb{N}$ such that $v_K \in [0, \varepsilon]$.

Item (iii): Fix $R > 0$ throughout and let $\hat{\beta} \in \mathcal{KL}$ be the function from part (i). Since $\mathrm{Id} - \varphi : [0, \infty) \to [0, b)$ is class-$\mathcal{K}$ and $d < b$, its inverse is strictly increasing and is defined at d.

Define

$$\sigma(d) := (\mathrm{Id} - \varphi)^{-1}(d),$$

so that $\varphi(\sigma(d)) + d = \sigma(d)$. We record two facts, both following directly from $\mathrm{Id} - \varphi$ being strictly increasing. We will analyse v_{k+1}, for the cases where $v_k \in [0, \sigma(d)]$ or $v_k \in (\sigma(d), \infty)$.

Forward invariance of $[0, \sigma(d)]$ under the mapping $\phi(\cdot)+d$: if $v_k \le \sigma(d)$, then $v_{k+1} \le \varphi(\sigma(d))+d = \sigma(d)$, so $[0, \sigma(d)]$ is forward-invariant.

Strict decrease of v_k on $(\sigma(d), \infty)$: if $v_k > \sigma(d)$, then $v_k - \varphi(v_k) > \sigma(d) - \varphi(\sigma(d)) = d$, hence $\phi(v_k) + d < v_k$ and $v_{k+1} \le \varphi(v_k) + d < v_k$.

Combining these two properties, the sequence $\{v_k\}$ is bounded above by $\max(v_0, \sigma(d))$ for all $k \ge 0$: it cannot increase above v_0 once it starts decreasing (which happens immediately if $v_0 > \sigma(d)$), and it is trapped once it enters $[0, \sigma(d)]$. In particular, for $v_0 \in [0, R]$:

$$v_k \le \max\big(R,\, \sigma(d)\big) \quad \text{for all } k \ge 0. \tag{42}$$

We claim that for all $v_0 \in [0, R]$, $d \ge 0$, and $k \ge 0$,

$$v_k \le \hat{\beta}(v_0, k) + \eta_R(d), \tag{43}$$

for some function η_R with $\eta_R(0) = 0$ that depends only on d (and R), not on v_0 or k. To see this, write $v_k = \hat{\beta}(v_0, k) + e_k$ where $e_k := v_k - \hat{\beta}(v_0, k)$ is the excess of the perturbed trajectory over the unperturbed bound. We need to show that e_k is bounded by a function of d (uniformly over $v_0 \in [0, R]$ and $k \ge 0$). Consider the unperturbed comparison sequence $u_0 = v_0$, $u_{k+1} = \varphi(u_k)$, so $u_k = \hat{\beta}(v_0, k)$ by part (i). Both sequences satisfy:

$$\begin{aligned} v_{k+1} &\le \varphi(v_k) + d, & v_0 &= s \in [0, R], \\ u_{k+1} &= \varphi(u_k), & u_0 &= s. \end{aligned}$$

Subtracting and using $\varphi(v_k) - \varphi(u_k) \le v_k - u_k$ (which follows from $\mathrm{Id} - \varphi$ being non-decreasing, so $v_k - \varphi(v_k) \ge u_k - \varphi(u_k)$ when $v_k \ge u_k$):

$$v_{k+1} - u_{k+1} \le (v_k - u_k) + d \quad \text{whenever } v_k \ge u_k.$$

Since $v_0 - u_0 = 0$, iterating this inequality gives $v_k - u_k \le kd$ whenever $v_k \ge u_k$ persists for all steps up to k.

At first glance, kd grows without bound. But recall that v_k eventually enters and remains in $[0, \sigma(d) + \delta]$ for any fixed $\delta > 0$. Let K^* be the first time $v_k \le \sigma(d) + \delta$. The strict decrease above $\sigma(d) + \delta$ gives a minimum decrease per step: for $v_k \in [\sigma(d) + \delta, \max(R, \sigma(d) + \delta)]$, we have

$$v_k - v_{k+1} \ge v_k - \varphi(v_k) - d \ge (\mathrm{Id} - \varphi)(v_k) - d.$$

On the compact set $[\sigma(d) + \delta, \max(R, \sigma(d) + \delta)]$, where $\delta > 0$ can be arbitrary small (fixed and bounded away from zero), the continuous function $(\mathrm{Id} - \varphi)(s) - d$ is strictly positive (being zero only at $s = \sigma(d)$). Hence, when $R > \sigma(d) + \delta$, the entry time satisfies the existence of K^* to have $v_{K^*} \le \sigma(d) + \delta$ (and $K^* = 0$ otherwise). In particular, K^* depends only on R and d (not on v_0 beyond the constraint $v_0 \le R$). Now we bound the excess for all k:

- For $k \le K^*$: using $v_k \le \max(R, \sigma(d)+\delta)$ from (42) and $\hat{\beta}(v_0, k) \ge 0$, we get $e_k \le \max(R, \sigma(d)+\delta)$. But more precisely, by the iterated bound above: $e_k = v_k - u_k \le K^* d$.
- For $k > K^*$: $v_k \le \sigma(d) + \delta$ and $\hat{\beta}(v_0, k) \ge 0$, so $e_k \le \sigma(d) + \delta$.

Therefore, for all $k \ge 0$:

$$e_k \le \max\big(K^* d,\ \sigma(d) + \delta\big) =: \eta_{R,\delta}(d).$$

Since $\delta > 0$ is arbitrary, for any $d > 0$ we can choose $\delta = d$ (or any class-$\mathcal{K}$ function of d), giving

$$e_k \le \max\big(K^* d,\ \sigma(d) + d\big) =: \eta_R(d).$$

Since $\sigma(0) = 0$ and $K^* = 0$ when $d = 0$, we have $\eta_R(0) = 0$. The function η_R is non-negative, finite, nondecreasing in d, and satisfies $\eta_R(0) = 0$. By regularisation, there exists $\hat{\gamma}_R \in \mathcal{K}$ with $\eta_R(d) \le \hat{\gamma}_R(d)$ for all $d \in [0, b)$.

Substituting into (43) and using $u_k = \hat{\beta}(v_0, k)$:

$$v_k \le \hat{\beta}(v_0, k) + \hat{\gamma}_R(d), \qquad k \ge 0,$$

for all $v_0 \in [0, R]$ and $d \ge 0$. □

Lemma 7 ([JW01, Lemma B.1])**.** *For any $\mathcal{K}_\infty$ function α, there exists a $\mathcal{K}_\infty$ function $\hat{\alpha}$ such that the following holds:*

- $\hat{\alpha}(r) \le \alpha(r)$ *for all $r \ge 0$; and,*
- $\mathrm{Id} - \hat{\alpha} \in \mathcal{K}$.

Proof of the Theorem 2. Let $R > 0$ be chosen so that (6)–(7) hold on $\mathcal{B}_R(z^e)$, and let $L_V > 0$ be a Lipschitz constant for V on $\mathcal{B}_R(z^e)$. Write $\delta := \sup_{j \ge 0} \|q_j\|$. Provided the trajectory stays in $\mathcal{B}_R(z^e)$, the Lipschitz property of V gives

$$V(z_{k+1}) = V(w(z_k) + q_k) \le V(w(z_k)) + L_V \|q_k\|. \tag{44}$$

Combining (44) with the decrease condition (7) and the upper sandwich bound (6) yields

$$V(z_{k+1}) \le V(z_k) - \alpha_3(\|z_k - z^e\|) + L_V \delta \le V(z_k) - \alpha_3\big(\alpha_2^{-1}(V(z_k))\big) + L_V \delta, \tag{45}$$

where the second inequality uses $\|z_k - z^e\| \ge \alpha_2^{-1}(V(z_k))$ from the right-hand side of (6). Define $\rho := \alpha_3 \circ \alpha_2^{-1}$, which is class-$\mathcal{K}_\infty$, and set $v_k := V(z_k)$. Then (45) reads

$$v_{k+1} \le v_k - \rho(v_k) + L_V \delta.$$

Define $\varphi(s) := s - \rho(s)$. Since ρ is class-$\mathcal{K}_\infty$, using Lemma 7, without loss of generality we can say $\varphi \in \mathcal{K}$, thus φ is continuous, $\varphi(0) = 0$, and $\mathrm{Id} - \varphi \in \mathcal{K}_\infty$. Thus, we have

$$v_{k+1} \le \varphi(v_k) + L_V \delta,$$

which is exactly the setting of Lemma 1 with $d = L_V \delta$. Since Theorem 2 is concerned with properties on a compact set, let $v_0 \in [0, \bar{R}]$, where $\bar{R} > 0$ such that $\{z : V(z) \le \bar{R}\} \subset \mathcal{B}_R(z^e)$. Applying part (iii) of the lemma gives a class-$\mathcal{KL}$ function $\hat{\beta}$ and a class-$\mathcal{K}$ function $\hat{\gamma}$ such that

$$V(z_k) \le \hat{\beta}(V(z_0),\, k) + \hat{\gamma}_{\bar{R}}(L_V \delta).$$

According to the sandwich bounds (9) the inequalities $V(z_0) \le \alpha_2(\|z_0 - z^e\|)$ and $\|z_k - z^e\| \le \alpha_1^{-1}(V(z_k))$ are satisfied. Using Lemma 6 with $\varepsilon > 0$ fixed, we obtain

$$\|z_k - z^e\| \le \alpha_1^{-1}\big((1+\epsilon)\,\hat{\beta}(\alpha_2(\|z_0 - z^e\|),\, k)\big) + \alpha_1^{-1}\big((1 + \tfrac{1}{\epsilon})\hat{\gamma}_{\bar{R}}(L_V \delta)\big).$$

With the definitions

$$\beta(s, k) := \alpha_1^{-1}\big((1+\epsilon)\hat{\beta}(\alpha_2(s),\, k)\big), \qquad \gamma(s) := \alpha_1^{-1}\big((1 + \tfrac{1}{\epsilon})\hat{\gamma}_{\bar{R}}(L_V s)\big),$$

we see that β is class-$\mathcal{KL}$ (as a composition of class-$\mathcal{K}_\infty$ and class-$\mathcal{KL}$ functions) and γ is class-$\mathcal{K}$, giving

$$\|z_k - z^e\| \le \beta(\|z_0 - z^e\|,\, k) + \gamma\Big(\sup_{0 \le j \le k} \|q_j\|\Big).$$

To complete the proof, we need trajectories to remain in $\{z \in \mathbb{R}^n : V(z) \leq \bar{R}\} \subset \mathcal{B}_R(z^e)$. From the proof of Lemma 1(iii), we know that $v_k \leq \max(v_0, \sigma(L_V\delta))$, so forward invariance of $\{z \in \mathbb{R}^n : V(z) \leq \bar{R}\}$ holds provided $\sigma(L_V\bar{q}) \leq \bar{R}$, i.e., $\bar{q} \leq \rho(\bar{R})/L_V$. Setting

$$\bar{r} := \alpha_2^{-1}(\bar{R}), \qquad \bar{q} := \frac{\rho(\bar{R})}{L_V},$$

we have that $\|z_0 - z^e\| \leq \bar{r}$ implies $V(z_0) \leq \bar{R}$, and $\sup_k \|q_k\| \leq \bar{q}$ guarantees $\sigma(L_V\bar{q}) = \bar{R}$, so the sublevel set $\{z \in \mathbb{R}^n : V(z) \leq \bar{R}\}$ is forward-invariant under the perturbed dynamics. Hence, the proof is complete. □

Lemma 8. *Let f be β-strongly convex. Then f_μ is β-strongly convex.*

Proof. From the strong convexity of f, we have $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\beta}{2}\|y - x\|^2$. Replacing x and y with $x + \mu u$ and $y + \mu u$, respectively, and taking expectation with respect to u, we obtain

$$f_\mu(y) \geq f_\mu(x) + \langle \nabla f_\mu(x), y - x\rangle + \tfrac{\beta}{2}\|y - x\|^2,$$

which completes the proof. □

C Proof of lemmas and theorems of Section 3

Proof of Lemma 2. Consider q_k given in (17), which satisfies

$$\|q_k\| \leq h\|\nabla f(\bar{z}_k) - \nabla f_\mu(\bar{z}_k)\| + h\|\nabla f_\mu(\bar{z}_k) - E_{U_k}[\nabla f_\mu(\tilde{z}_k)]\|. \tag{46}$$

Using [NS17, Lemma 3], we know that $\|\nabla f(\bar{z}_k) - \nabla f_\mu(\bar{z}_k)\| \leq \frac{\mu}{2}L_1(n+3)^{3/2}$. Thus, the first term on the right-hand side (46) is bounded and can be manipulated through the choice of h and μ. For the second term on the right-hand side of (46), we have

$$\begin{aligned}\|E_{U_k}[\nabla f_\mu(\tilde{z}_k)] - \nabla f_\mu(\bar{z}_k)\| &\leq \|E_{U_k}[\nabla f_\mu(\tilde{z}_k) - \nabla f_\mu(\bar{z}_k)]\| \leq E_{U_k}[\|\nabla f_\mu(\tilde{z}_k) - \nabla f_\mu(\bar{z}_k)\|] \\ &\leq L_1 E_{U_k}[\|\tilde{z}_k - \bar{z}_k\|] \end{aligned} \tag{47}$$

Now let $e_k = \tilde{z}_k - \bar{z}_k$, i.e., we need to find an upper bound on $E[\|e_k\|]$ or $E[\|e_k\|^2]$ to prove the assertion. Combining (14), (16) and (17), we know that

$$\begin{aligned} e_{k+1} &= e_k - hg_\mu(\tilde{z}_k) + hE_{U_k}[\nabla f_\mu(\tilde{z}_k)] \\ &= e_k - h(g_\mu(\tilde{z}_k) - \nabla f_\mu(\tilde{z}_k)) - h(\nabla f_\mu(\tilde{z}_k) - E_{U_k}[\nabla f_\mu(\tilde{z}_k)]). \end{aligned}$$

To proceed, we denote $v_k = g_\mu(\tilde{z}_k) - \nabla f_\mu(\tilde{z}_k)$ and $l_k = \nabla f_\mu(\tilde{z}_k) - E_{U_k}[\nabla f_\mu(\tilde{z}_k)]$ and thus $e_{k+1} = e_k - h(v_k + l_k)$. Considering Assumption 1, the variance of the random oracle is bounded, and we have $E[\|v_k\|^2] \leq \frac{\sigma^2}{t}$. Moreover, for l_k we have

$$\begin{aligned} E_{U_k}[\|l_k\|] &= E_{U_k}[\|\nabla f_\mu(\tilde{z}_k) - \nabla f_\mu(\bar{z}_k) + \nabla f_\mu(\bar{z}_k) - E_{U_k}[\nabla f_\mu(\tilde{z}_k)]\|] \\ &\leq E_{U_k}[\|\nabla f_\mu(\tilde{z}_k) - \nabla f_\mu(\bar{z}_k)\|] + E_{U_k}[\|\nabla f_\mu(\bar{z}_k) - \nabla f_\mu(\tilde{z}_k)\|] \\ &\leq 2L_1 E_{U_k}[\|e_k\|] \end{aligned} \tag{48}$$

Thus we get

$$E[\|e_{k+1}\|] \leq (1 + 2L_1 h)E[\|e_k\|] + h\frac{\sigma}{\sqrt{t}}, \tag{49}$$

which does not provide a uniform bound on $E_{U_k}[\|\tilde{z}_k - \bar{z}_k\|]$ yet. To proceed, we use the strong convexity properties of f. Focusing on $E_{U_k}[\|e_k\|^2]$ and using the fact that $E[v_k] = 0$, we get

$$E_{U_k}[\|e_{k+1}\|^2] = E_{U_k}[\|e_k\|^2] + h^2 E_{U_k}[\|v_k + l_k\|^2] - 2hE_{U_k}[\langle e_k, l_k\rangle]. \tag{50}$$

Moreover, we have

$$E_{U_k}[\langle e_k, l_k\rangle] = E_{U_k}[\langle e_k, \nabla f_\mu(\tilde{z}_k)\rangle]$$

since $E_{U_k}[\langle e_k, E_{U_k}[\nabla f_\mu(\tilde{z}_k)]\rangle] = 0$. Now, the strong convexity of f and consequently strong convexity of f_μ (Lemma 8) implies that

$$f_\mu(\bar{z}_k) \geq f_\mu(\tilde{z}_k) + \langle \nabla f_\mu(\tilde{z}_k), \bar{z}_k - \tilde{z}_k\rangle + \frac{\beta}{2}\|\bar{z}_k - \tilde{z}_k\|^2$$

or

$$E_{U_k}[\langle e_k, \nabla f_\mu(\tilde{z}_k)\rangle] \geq E_{U_k}[f_\mu(\tilde{z}_k)] - f_\mu(\bar{z}_k) + \frac{\beta}{2} E_{U_k}[\|e_k\|^2].$$

Moreover, combining strong convexity of f_μ and Jensen's inequality, leads to the estimates

$$E_{U_k}[f_\mu(\tilde{z}_k)] \geq f_\mu(E_{U_k}[\tilde{z}_k]) \geq f_\mu(\bar{z}_k),$$

and hence

$$E_{U_k}[\langle e_k, \nabla f_\mu(\tilde{z}_k)\rangle] \geq \frac{\beta}{2} E_{U_k}[\|e_k\|^2]. \tag{51}$$

Thus, substituting (51) in (50), we have

$$E_{U_k}[\|e_{k+1}\|^2] \leq (1 - h\beta) E_{U_k}[\|e_k\|^2] + h^2 E_{U_k}[\|v_k + l_k\|^2]. \tag{52}$$

We know that $\|v_k + l_k\|^2 \leq 2\|v_k\|^2 + 2\|l_k\|^2$ and $E[\|v_k\|^2] \leq \frac{\sigma^2}{t}$. Similar to the estimates in (48), we can hence conclude the following chain of equalities and inequalities:

$$\begin{aligned} E[\|l_k\|^2] &= E_{U_k}[\|\nabla f_\mu(\tilde{z}_k) - \nabla f_\mu(\bar{z}_k) + \nabla f_\mu(\bar{z}_k) - E_{U_k}[\nabla f_\mu(\tilde{z}_k)]\|^2] \\ &= E[\|\nabla f_\mu(\tilde{z}_k) - \nabla f_\mu(\bar{z}_k)\|^2 + \|\nabla f_\mu(\bar{z}_k) - E_{U_k}[\nabla f_\mu(\tilde{z}_k)]\|^2] \\ &\quad - 2E[\langle \nabla f_\mu(\tilde{z}_k) - \nabla f_\mu(\bar{z}_k), E_{U_k}[\nabla f_\mu(\tilde{z}_k)] - \nabla f_\mu(\bar{z}_k)\rangle] \\ &\leq E[\|\nabla f_\mu(\tilde{z}_k) - \nabla f_\mu(\bar{z}_k)\|^2] \\ &\leq L_1^2 E[\|e_k\|^2]. \end{aligned} \tag{53}$$

Thus

$$E_{U_k}[\|e_{k+1}\|^2] \leq \rho E_{U_k}[\|e_k\|^2] + \frac{2h^2\sigma^2}{t} \tag{54}$$

and where ρ is defined as $\rho = 1 - h\beta + 2L_1^2 h^2$. To have $\rho < 1$, we require $h < \frac{\beta}{2L_1^2}$. Since $e_0 = 0$, by induction we get

$$E_{U_k}[\|e_k\|^2] \leq \frac{2h^2\sigma^2}{t(1-\rho)} = \frac{2h^2\sigma^2}{t(h\beta - 2L_1^2 h^2)} = \frac{2h\sigma^2}{t(\beta - 2L_1^2 h)}. \tag{55}$$

Hence, combining all the estimates above, we get

$$\sup_k \|q_k\| \leq h\frac{\mu}{2} L_1 (n+3)^{3/2} + hL_1 \sqrt{\frac{2h\sigma^2}{t(\beta - 2L_1^2 h)}}$$

which completes the proof. □

Proof of Theorem 3. By Lemma 2, the averaged ZO-GD dynamics (16) take the form of the perturbed system (4) with perturbation satisfying the bound (18). It remains to verify that the unperturbed GD dynamics satisfy the hypotheses of Theorem 1.

For β-strongly convex f with L_1-Lipschitz gradients and $h \in \big(0, \frac{2}{\beta + L_1}\big)$, from [Nes18, Thm. 2.1.15] we have

$$\|z_{k+1} - z^*\|^2 \leq \Big(1 - \frac{2h\beta L_1}{\beta + L_1}\Big)\|z_k - z^*\|^2,$$

where $z^* = \arg\min_{x \in \mathbb{R}^n} f(x)$. Consider the Lyapunov function $V(z) = \|z - z^*\|^2$. Then (6) holds with $c_1 = c_2 = 1$, and the above contraction gives (7) with $c = \frac{2h\beta L_1}{\beta + L_1} \in (0, 1)$. All hypotheses of Theorem 1 are therefore satisfied.

The ISS bound (8) then guarantees that ZO-GD converges to a neighbourhood of z^* with the same exponential rate $(1-c)^{k/2}$ as GD. The radius of this neighbourhood is governed by $\gamma\big(\sup_k \|q_k\|\big)$, which, combined with the perturbation bound from Lemma 2, yields (19). Since the bound (18) can be made smaller than $\bar{q}$ by choosing μ sufficiently small and t sufficiently large, such parameter choices always exist. □

Proof of Lemma 3. Consider q_k given in (24) for HB, which satisfies

$$\|q_k\| \;\le\; h_1\big\|E_{U_k}[\nabla f_\mu(x_k) - \nabla f(x_k)]\big\| + h_1\big\|E_{U_k}[\nabla f(x_k)] - \nabla f(\bar{x}_k)\big\|. \tag{56}$$

Using [NS17, Lem. 3], we know that $\|\nabla f_\mu(x) - \nabla f(x)\| \le \frac{\mu}{2}\, L_1\, (n+3)^{3/2}$ uniformly in x. Thus, the first term in the right-hand side (56) is bounded and can be ensured to be arbitrarily small by selecting h_1 and μ appropriately. For the second term in the right-hand side of (56), by L_1-Lipschitz continuity of ∇f and Jensen's inequality, we have

$$\big\|E_{U_k}[\nabla f(x_k)] - \nabla f(\bar{x}_k)\big\| \;\le\; L_1\, E_{U_k}[\|x_k - \bar{x}_k\|]. \tag{57}$$

Now let $e_k = z_k - \bar{z}_k$ so that $x_k - \bar{x}_k = Te_k$ where $T = [0 \;\; I_n]$. Thus we are interested in bounding $E[\|Te_k\|]$ or $E[\|Te_k\|^2]$. Subtracting the averaged dynamics (23) (for HB) from the ZO-HB update gives

$$e_{k+1} = A\, e_k - h_1 \begin{bmatrix} 0 \\ v_k + l_k \end{bmatrix}, \tag{58}$$

where $v_k = g_\mu(x_k) - \nabla f_\mu(x_k)$ with $E[v_k] = 0$, and $l_k = \nabla f_\mu(x_k) - E_{U_k}[\nabla f_\mu(x_k)]$. As a next step, let $x_k^e = Te_k$. From the lower block of (58) we have

$$x_{k+1}^e = (1+h_2)\, x_k^e - h_2\, x_{k-1}^e - h_1(v_k + l_k). \tag{59}$$

Let $a_k = E[\|x_k^e\|^2]$. Squaring (59), taking expectations, and using $E[\langle y, v_k\rangle] = 0$ for any y independent of u_k, leads to the expression

$$a_{k+1} = E\big[\|(1+h_2)x_k^e - h_2 x_{k-1}^e\|^2\big] + h_1^2\, E[\|v_k + l_k\|^2] - 2h_1\, E\big[\langle (1+h_2)x_k^e - h_2 x_{k-1}^e,\; l_k\rangle\big]. \tag{60}$$

We derive bounds for each term in (60) separately. For the deterministic quadratic term, expanding and applying Young's inequality provides the upper bound

$$\begin{aligned}\|(1+h_2)x_k^e - h_2 x_{k-1}^e\|^2 &= (1+h_2)^2\|x_k^e\|^2 + h_2^2\|x_{k-1}^e\|^2 - 2h_2(1+h_2)\langle x_{k-1}^e, x_k^e\rangle \\ &\le (1+h_2)(1+2h_2)\,\|x_k^e\|^2 + h_2(1+2h_2)\,\|x_{k-1}^e\|^2. \end{aligned}\tag{61}$$

For the inner product, since $E[x_k^e] = 0$ and $E_{U_k}[\nabla f_\mu(x_k)]$ is deterministic, we get $E[\langle x_k^e, l_k\rangle] = E[\langle x_k - \bar{x}_k, \nabla f_\mu(x_k)\rangle]$. By β-strong convexity of f_μ (from Lemma 8), the inequality

$$\langle \nabla f_\mu(x_k),\, x_k - \bar{x}_k\rangle \;\ge\; f_\mu(x_k) - f_\mu(\bar{x}_k) + \frac{\beta}{2}\,\|x_k - \bar{x}_k\|^2$$

is thus satisfied. Taking expectations and applying Jensen's inequality to $E[f_\mu(x_k)] \ge f_\mu(\bar{x}_k)$ then yields

$$E[\langle x_k^e, l_k\rangle] \;\ge\; \tfrac{\beta}{2}\, a_k. \tag{62}$$

For the cross term $E[\langle x_{k-1}^e, l_k\rangle]$, using $E[\|l_k\|] \le 2L_1\, E[\|x_k^e\|]$ (from L_1-smoothness of f_μ and the triangle inequality) and Young's inequality, we know that

$$|E[\langle x_{k-1}^e, l_k\rangle]| \;\le\; L_1\big(a_{k-1} + a_k\big). \tag{63}$$

Since we assume that the variance of the random oracle is bounded by $E[\|v_k\|^2] \le \sigma^2/t$, for l_k, similar to (53), it holds that $E[\|l_k\|^2] \le L_1^2\, a_k$ and thus $E[\|v_k + l_k\|^2] \le \frac{2\sigma^2}{t} + 2L_1^2\, a_k$.

Substituting all bounds into (60) and letting

$$\phi = (1+h_2)(1+2h_2) - h_1(1+h_2)\beta + 2h_1h_2L_1 + 2h_1^2L_1^2, \tag{64}$$

$$c_0 = h_2(1+2h_2) + 2h_1h_2L_1, \tag{65}$$

we get the two-step recursion

$$a_{k+1} \;\le\; \phi\, a_k + c_0\, a_{k-1} + \frac{2h_1^2\sigma^2}{t}. \tag{66}$$

As a next step, we define $m_k = \max\{a_k, a_{k-1}\}$. Since $a_k \le m_k$ and $a_{k-1} \le m_k$, (66) gives $a_{k+1} \le (\phi + c_0)\, m_k + \frac{2h_1^2\sigma^2}{t}$. Since additionally $a_k \le m_k$ holds, m_{k+1} can be upper bounded by

$$m_{k+1} = \max\{a_{k+1}, a_k\} \;\le\; \max\left\{(\phi + c_0)\, m_k + \frac{2h_1^2\sigma^2}{t}, m_k\right\}, \tag{67}$$

From $m_0 = 0$ (as $e_0 = 0$), we get

$$m_{k+1} = \max\{a_{k+1}, a_k\} \;\le\; (\phi + c_0)\, m_k + \frac{2h_1^2\sigma^2}{t}, \tag{68}$$

which is a standard one-step contraction provided $\rho_{\mathrm{HB}} := \phi + c_0 < 1$. Computing ρ_{HB} explicitly, we get

$$\rho_{\mathrm{HB}} = \phi + c_0 = (1 + 2h_2)^2 - h_1(1 + h_2)\beta + 4h_1h_2L_1 + 2h_1^2L_1^2. \tag{69}$$

Thus $\rho_{\mathrm{HB}} < 1$ requires

$$h_1(1 + h_2)\beta > 4h_2 + 4h_2^2 + 4h_1h_2L_1 + 2h_1^2L_1^2 \tag{70}$$

to hold. To find a sufficient condition for (70) to be satisfied let $h_2 \in (0, ch_1)$. Then it is sufficient to have

$$h_1\beta > 2h_1^2L_1^2 + 4ch_1 + 4c^2h_1^2 + 4ch_1^2L_1 \quad \text{or equivalently} \quad \beta > 2h_1L_1^2 + 4c + 4c^2h_1 + 4ch_1L_1.$$

Thus we need $c \in (0, \frac{\beta}{4})$ and $h_1 \in (0, \frac{\beta - 4c}{4c^2 + 4cL_1 + 2L_1^2})$ to guarantee that (70) is satisfied, providing the bounds stated in Lemma 3.

Since $m_0 = 0$ (as $e_0 = 0$), iterating (68) leads to

$$\sup_k\, a_k \;\le\; \sup_k\, m_k \;\le\; \frac{2h_1^2\sigma^2}{t(1 - \rho_{\mathrm{HB}})}$$

and therefore, by Jensen's inequality we have

$$\sup_k\, E[\|x_k - \bar{x}_k\|] \;\le\; \sup_k \sqrt{a_k} \;\le\; \frac{h_1\sigma}{\sqrt{t}} \sqrt{\frac{2}{1 - \rho_{\mathrm{HB}}}}. \tag{71}$$

Combining (56), the smoothing bias bound, (57), and (71) we can thus conclude that

$$\begin{aligned}\sup_k\, \|q_k\| &\le \frac{h_1\mu\, L_1(n+3)^{3/2}}{2} + h_1L_1 \cdot \frac{h_1\sigma}{\sqrt{t}} \sqrt{\frac{2}{1 - \rho_{\mathrm{HB}}}} \\ &= \frac{h_1\mu\, L_1(n+3)^{3/2}}{2} + \frac{h_1^2L_1\,\sigma}{\sqrt{t}} \sqrt{\frac{2}{1 - \rho_{\mathrm{HB}}}}\,,\end{aligned} \tag{72}$$

which completes the proof. □

Proof of Theorem 4. By Lemma 3, the averaged ZO-HB dynamics (23) take the form of the perturbed system (4) with bounded perturbation. For β-strongly convex f with L_1-Lipschitz gradients and parameters satisfying the hypotheses of Lemma 3, HB locally converges to the minimiser z^* exponentially fast [Pol64]. By the converse Lyapunov theorem for exponentially stable discrete-time systems (see, e.g., [KB23, Thm. 5.6]), there exists a Lyapunov function V satisfying (6)–(7) on some ball $\mathcal{B}_R(z^e)$, $R > 0$, in the augmented state space $\mathbb{R}^{2n}$. All hypotheses of Theorem 1 are therefore satisfied, and the conclusion follows by the same argument as in the proof of Theorem 3. □

Proof of Lemma 4. Consider q_k given in (24) for NAG, which satisfies

$$\|q_k\| \le h_1 \big\| E_{U_k}[\nabla f_\mu(TAz_k) - \nabla f(TAz_k)] \big\| + h_1 \big\| E_{U_k}[\nabla f(TAz_k)] - \nabla f(TA\bar{z}_k) \big\|. \tag{73}$$

Using [NS17, Lem. 3], we know that $\|\nabla f_\mu(x) - \nabla f(x)\| \le \frac{\mu}{2} L_1(n+3)^{3/2}$ uniformly in x. Thus, the first term in the right-hand side (73) is bounded and can be controlled by the choice of h_1 and μ. For the second term in the right-hand side of (73), by L_1-Lipschitz continuity of ∇f and Jensen's inequality, we have

$$\big\| E_{U_k}[\nabla f(TAz_k)] - \nabla f(TA\bar{z}_k) \big\| \le L_1\, E_{U_k}[\|TAz_k - TA\bar{z}_k\|] = L_1\, E_{U_k}[\|TAe_k\|], \tag{74}$$

where $e_k = z_k - \bar{z}_k$. Thus, we need to derive an upper bound on $E[\|TAe_k\|]$ or $E[\|TAe_k\|^2]$. Let $x_k^e = Te_k$ and define the lookahead error $y_k^e = TAe_k = (1 + h_2)x_k^e - h_2x_{k-1}^e$. Subtracting the averaged dynamics (23) (for NAG) from the ZO-NAG update gives

$$e_{k+1} = A\, e_k - h_1 \begin{bmatrix} 0 \\ v_k + l_k \end{bmatrix}, \tag{75}$$

where $v_k = g_\mu(TAz_k) - \nabla f_\mu(TAz_k)$ with $E[v_k] = 0$, and $l_k = \nabla f_\mu(TAz_k) - E_{U_k}[\nabla f_\mu(TAz_k)]$. From the lower block of (75) we obtain

$$x^e_{k+1} = y^e_k - h_1(v_k + l_k). \tag{76}$$

Let $a_k = E[\|x^e_k\|^2]$ and $b_k = E[\|y^e_k\|^2]$. Squaring (76), taking expectations, and using the fact that $E[\langle y^e_k, v_k\rangle] = 0$ leads to the equation

$$a_{k+1} = b_k + h_1^2\, E[\|v_k + l_k\|^2] - 2h_1\, E[\langle y^e_k,\, l_k\rangle]. \tag{77}$$

To proceed, we derive upper bounds for each term in the right-hand side. For the inner product, since $E[y^e_k] = 0$ and $E_{U_k}[\nabla f_\mu(TAz_k)]$ is deterministic, we get $E[\langle y^e_k, l_k\rangle] = E[\langle TAe_k, \nabla f_\mu(TAz_k)\rangle]$. By β-strong convexity of f_μ (Lemma 8), it additionally holds that

$$\langle \nabla f_\mu(TAz_k),\, TAz_k - TA\bar{z}_k\rangle \geq f_\mu(TAz_k) - f_\mu(TA\bar{z}_k) + \frac{\beta}{2}\|TAe_k\|^2.$$

Taking expectations and applying Jensen's inequality to $E[f_\mu(TAz_k)] \geq f_\mu(E[TAz_k]) = f_\mu(TA\bar{z}_k)$ allows us to write

$$E[\langle y^e_k, l_k\rangle] \geq \frac{\beta}{2}\, b_k. \tag{78}$$

This is the key advantage of the NAG structure compared to the HB structure. The strong convexity contraction acts directly on $b_k = E[\|y^e_k\|^2]$, which is the same quantity that appears with a positive sign in (77). There is no separate momentum cross term to bound.

For the noise, since $l_k = \nabla f_\mu(TAz_k) - E_{U_k}[\nabla f_\mu(TAz_k)]$, using L_1-smoothness and similar to the process in (48) and (53), $E[\|l_k\|] \leq 2L_1\, E[\|y^e_k\|]$, so $E[\|l_k\|^2] \leq L_1^2\, b_k$ and with $E[\|v_k\|^2] \leq \sigma^2/t$, we get $E[\|v_k + l_k\|^2] \leq \frac{2\sigma^2}{t} + 2L_1^2\, b_k$. Substituting these estimates into (77) provides the upper bound

$$a_{k+1} \leq (1 - h_1\beta + 2h_1^2 L_1^2)\, b_k + \frac{2h_1^2\sigma^2}{t}. \tag{79}$$

Let $\psi = 1 - h_1\beta + 2h_1^2L_1^2$. For the quadratic term $b_k = E[\|y^e_k\|^2]$, expanding and applying Young's inequality exactly as in (61) yields

$$b_k \leq (1 + h_2)(1 + 2h_2)\, a_k + h_2(1 + 2h_2)\, a_{k-1}. \tag{80}$$

Substituting (80) into (79) and letting

$$\begin{aligned} \phi &= \psi\,(1 + h_2)(1 + 2h_2), \\ c_0 &= \psi\, h_2(1 + 2h_2), \end{aligned}$$

we get the two-step recursion

$$a_{k+1} \leq \phi\, a_k + c_0\, a_{k-1} + \frac{2h_1^2\sigma^2}{t}. \tag{81}$$

Next, we define $m_k = \max\{a_k, a_{k-1}\}$. Since $a_k \leq m_k$ and $a_{k-1} \leq m_k$, (81) gives $a_{k+1} \leq (\phi + c_0)\, m_k + \frac{2h_1^2\sigma^2}{t}$ and from $a_k \leq m_k$ we have

$$m_{k+1} = \max\{a_{k+1}, a_k\} \leq \max\big\{(\phi + c_0)\, m_k + \tfrac{2h_1^2\sigma^2}{t},\; m_k\big\}. \tag{82}$$

Since $m_0 = 0$ (as $e_0 = 0$), we get

$$m_{k+1} \leq (\phi + c_0)\, m_k + \frac{2h_1^2\sigma^2}{t}, \tag{83}$$

which is a standard one-step contraction provided $\rho_{\mathrm{NAG}} := \phi + c_0 < 1$ and computing ρ_{NAG} explicitly yields

$$\rho_{\mathrm{NAG}} = \psi\,(1 + 2h_2)^2 = (1 - h_1\beta + 2h_1^2L_1^2)(1 + 2h_2)^2. \tag{84}$$

Thus $\rho_{\mathrm{NAG}} < 1$ requires $\psi(1 + 2h_2)^2 < 1$, i.e.,

$$(1 + 2h_2)^2 < \frac{1}{1 - h_1\beta + 2h_1^2L_1^2}. \tag{85}$$

Choosing $h_1 \in (0, \frac{\beta}{2L_1^2})$ ensures $\psi \in (0,1)$, and then any $h_2 \in (0, \frac{1}{4}(\psi^{-1/2}-1))$ guarantees (85).

Since $m_0 = 0$, iterating (83), we then have

$$\sup_k a_k \leq \sup_k m_k \leq \frac{2h_1^2\sigma^2}{t(1-\rho_{\mathrm{NAG}})}.$$

To bound $E[\|TAe_k\|]$, note that $\|TAe_k\|^2 = \|y_k^e\|^2 = b_k$, and from (80) we obtain

$$\sup_k E[\|y_k^e\|^2] \leq (1+2h_2)^2 \sup_k m_k \leq \frac{2(1+2h_2)^2 h_1^2 \sigma^2}{t(1-\rho_{\mathrm{NAG}})}.$$

Finally, by Jensen's inequality,

$$\sup_k E[\|TAe_k\|] \leq \frac{(1+2h_2)h_1\sigma}{\sqrt{t}}\sqrt{\frac{2}{1-\rho_{\mathrm{NAG}}}}, \tag{86}$$

and combining (73), the smoothing bias bound, (74), and (86) it holds that

$$\begin{aligned}\sup_k \|q_k\| &\leq \frac{h_1\mu\, L_1(n+3)^{3/2}}{2} + h_1 L_1 \cdot \frac{(1+2h_2)h_1\sigma}{\sqrt{t}}\sqrt{\frac{2}{1-\rho_{\mathrm{NAG}}}} \\ &= \frac{h_1\mu\, L_1(n+3)^{3/2}}{2} + \frac{(1+2h_2)h_1^2 L_1\sigma}{\sqrt{t}}\sqrt{\frac{2}{1-\rho_{\mathrm{NAG}}}}\,, \end{aligned} \tag{87}$$

which completes the proof. □

Proof of Theorem 5. By Lemma 4, the averaged ZO-NAG dynamics (23) take the form of the perturbed system (4) with bounded perturbation. For β-strongly convex f with L_1-Lipschitz gradients and parameters satisfying the hypotheses of Lemma 4, NAG converges to the minimiser z^e exponentially fast [Nes18]. By the converse Lyapunov theorem for exponentially stable discrete-time systems (see, e.g., [KB23, Thm. 5.6]), there exists a Lyapunov function V satisfying (6), (7) on some ball $\mathcal{B}_R(z^e)$, $R>0$, in the augmented state space $\mathbb{R}^{2n}$. All hypotheses of Theorem 1 are therefore satisfied, and the conclusion follows by the same argument as in the proof of Theorem 3. □

Proof of Lemma 5. Consider q_k given in (35), which satisfies

$$\|q_k\| \leq h\|\nabla f(\bar{z}_k) - \nabla f_\mu(\bar{z}_k)\| + h\|\nabla f_\mu(\bar{z}_k) - E_{U_k}[\nabla f_\mu(\tilde{z}_k)]\| \tag{88}$$

Using [NS17, Lem. 3], we know that $\|\nabla f(\bar{z}_k) - \nabla f_\mu(\bar{z}_k)\| \leq \frac{\mu}{2}L_1(n+3)^{3/2}$. Thus, the first term in the right-hand side of (88) is bounded and can be manipulated by choosing h and μ appropriately. For the second term in the right-hand side of (88), we have

$$\begin{aligned}\|E_{U_k}[\nabla f_\mu(\tilde{z}_k)] - \nabla f_\mu(\bar{z}_k)\| &\leq \|E_{U_k}[\nabla f_\mu(\tilde{z}_k) - \nabla f_\mu(\bar{z}_k)]\| \\ &\leq E_{U_k}[\|\nabla f_\mu(\tilde{z}_k) - \nabla f_\mu(\bar{z}_k)\|] \\ &\leq L_1 E_{U_k}[\|\tilde{z}_k - \bar{z}_k\|] \end{aligned} \tag{89}$$

Now, let $e_k = \tilde{z}_k - \bar{z}_k$. As a next step, we derive an upper bound on $E[\|e_k\|^2]$. We know that

$$\begin{aligned} e_{k+1} &= ce_k - hg_\mu(\tilde{z}_k) + hE_{U_k}[\nabla f_\mu(\tilde{z}_k)] \\ &= e_k - h(g_\mu(\tilde{z}_k) - \nabla f_\mu(\tilde{z}_k)) - h(\nabla f_\mu(\tilde{z}_k) - E_{U_k}[\nabla f_\mu(\tilde{z}_k)]) \end{aligned}$$

To proceed, we use the notation $v_k = g_\mu(\tilde{z}_k) - \nabla f_\mu(\tilde{z}_k)$ and $l_k = \nabla f_\mu(\tilde{z}_k) - E_{U_k}[\nabla f_\mu(\tilde{z}_k)]$ and thus $e_{k+1} = ce_k - h(v_k + l_k)$. We assume the variance of the random oracle is bounded and get $E[\|v_k\|^2] \leq \frac{\sigma^2}{t}$. Moreover, for l_k the following chain of inequalities is satisfied:

$$E[\|l_k\|] \leq E[\|\nabla f_\mu(\tilde{z}_k) - \nabla f_\mu(\bar{z}_k)\|] + E[\|\nabla f_\mu(\bar{z}_k) - \nabla f_\mu(\tilde{z}_k)\|] \leq 2L_1 E[\|e_k\|]. \tag{90}$$

From the fact that $E[v_k] = 0$, we get the following bound on $E_{U_k}[\|e_{k+1}\|^2]$,

$$E_{U_k}[\|e_{k+1}\|^2] = c^2 E_{U_k}[\|e_k\|^2] + h^2 E_{U_k}[\|v_k + l_k\|^2] - 2ch E_{U_k}[\langle e_k, l_k\rangle] \tag{91}$$

$$\leq c^2 E_{U_k}[\|e_k\|^2] + h^2 E_{U_k}[\|v_k + l_k\|^2] + 2ch E_{U_k}[\|e_k\|\|l_k\|] \tag{92}$$

$$\leq (c^2 + 4cL_1 h) E_{U_k}[\|e_k\|^2] + h^2 E_{U_k}[\|v_k + l_k\|^2], \tag{93}$$

where the last inequality uses Jensen's inequality and steps similar to same as (90). Moreover, we know that $\|v_k + l_k\|^2 \leq 2\|v_k\|^2 + 2\|l_k\|^2$ and $E[\|v_k\|^2] \leq \frac{\sigma^2}{t}$. Similar to the process in (90) and considering the fact that $E[\|X - E[X]\|^2] \leq E[\|X\|^2]$, we have $E[\|l_k\|^2] \leq L_1^2 E[\|e_k\|^2]$ and thus

$$E_{U_k}[\|e_{k+1}\|^2] \leq \rho E_{U_k}[\|e_k\|^2] + \frac{2h^2\sigma^2}{t}, \tag{94}$$

where $\rho = c^2 + 4cL_1h + 2L_1^2h^2$. For $c \in [0, 1)$, to have $\rho < 1$, we require $h \in \big(0, \frac{\sqrt{2(1+c^2)}-2c}{2L_1}\big)$. Since $e_0 = 0$, by induction we get

$$E_{U_k}[\|e_k\|^2] \leq \frac{2h^2\sigma^2}{t(1-\rho)} = \frac{2h^2\sigma^2}{t(1-c^2-4cL_1h-2L_1^2h^2)}. \tag{95}$$

Hence, combining the derivations above, we get

$$\sup_k \|q_k\| \leq h\frac{\mu}{2}L_1(n+3)^{3/2} + hL_1\sqrt{\frac{2h^2\sigma^2}{t(1-c^2-4cL_1h-2L_1^2h^2)}} \tag{96}$$

which completes the proof. □

Proof of Theorem 6. By Lemma 5, the averaged ZO-GD dynamics (34) take the form of the perturbed system (4) with perturbation satisfying the bound (36). It remains to verify that the unperturbed regularised GD dynamics (32) satisfy the hypotheses of Theorem 2.

By assumption, GD applied to the regularised problem converges to a fixed point z^e. Since we have convergence for GD, the converse Lyapunov theorem for asymptotically stable discrete-time systems (see, e.g., [KB23, Thm. 5.5 and Thm. 5.6]) guarantees the existence of a Lyapunov function V satisfying (9)–(10) on some ball $\mathcal{B}_R(z^e)$, $R > 0$. All hypotheses of Theorem 2 are therefore satisfied.

The ISS bound (11) then guarantees that ZO-GD converges to a neighbourhood of z^e with the same transient decay profile as GD. The radius of this neighbourhood is governed by $\gamma\big(\sup_k \|q_k\|\big)$, which, combined with the perturbation bound from Lemma 5, yields (37). Since the bound (36) can be made smaller than $\bar{q}$ by choosing μ sufficiently small and t sufficiently large, such parameter choices always exist. □

D Complementary remarks on the results of Section 3

The following remarks provide additional technical details on the results presented in Section 3.

Remark 2. *While the first term in* (19) *can be made small by choosing μ small, essentially at no cost, the second term is governed by h and t. It is at most $\varepsilon > 0$ if and only if*

$$t \geq \frac{2h^3L_1^2\sigma^2}{\varepsilon^2(\beta - 2L_1^2h)},$$

so t scales with σ^2 rather than directly with n. Since the classical bound on the second moment of g_μ grows with n [NS17], t may inherit a dimension dependence in the worst case. These t queries are, however, independent and can be evaluated in parallel, which is the favourable regime for ZO methods. When the optimal step size of GD is of order $\sigma^{-2/3}$, a small value of t suffices for ZO-GD to track GD closely with the same computational cost. Moreover, Theorem 3 requires $h \in \big(0, \frac{\beta}{2L_1^2}\big) = \big(0, \frac{1}{2\kappa L_1}\big)$ with $\kappa = L_1/\beta$, whereas GD allows $h \in \big(0, \frac{2}{\beta+L_1}\big)$ [Nes18]. Both ranges are dimension-free and of the same order when $\beta \sim L_1$, but for ill-conditioned problems ($\kappa \gg 1$) the ZO range is more conservative by a factor of order κ.

Remark 3. *We note that the expected performance of ZO-GD for strongly convex objective functions in minimisation problems has been analysed in [NS17, Thm. 8]. Compared to Theorem 3, the result in [NS17, Thm. 8] exhibits an explicit dependence of the iteration complexity on the dimension, and does not account for the variance bound σ or the number of sampled directions t. Moreover, the prescribed step size scales inversely with n, which is not consistent with standard GD. In contrast, Theorem 3 shows that ZO-GD, in expectation, achieves the same decay rate as GD and converges to a neighbourhood of the GD fixed point, whose size depends on σ and can be manipulated via h, μ, and t. Furthermore, for any step size consistent with GD, one can select μ and t such that the perturbation is arbitrarily small and closely track GD.*

Remark 4 (Relaxation of Assumption 1)**.** *Assumption 1 requires the variance of the oracle g_μ defined in* (12) *to be uniformly bounded by a constant σ^2. In some scenarios, the variance of the zeroth-order oracle scales with the magnitude of the gradient at the query point. Here, we show analogous results to Lemma 2 and Theorem 3 under the following relaxed assumption. Analogous results for other lemmas and theorems can be derived following similar steps.*

Assumption 2. *The variance of the oracle g_μ defined in* (12) *satisfies*

$$\mathbb{E}_u\Big[\|g_\mu(x) - \nabla f_\mu(x)\|^2\Big] \le \frac{\sigma_0^2}{t} + \frac{\sigma_1^2}{t}\|\nabla f(x)\|^2, \tag{97}$$

where $\sigma_0, \sigma_1 \ge 0$.

Note that Assumption 1 is recovered by setting $\sigma_1 = 0$ and $\sigma_0 = \sigma$. Under this relaxed assumption, the proof of Lemma 2 proceeds as follows. Consider the recursion for $\mathbb{E}_{U_k}[\|e_{k+1}\|^2]$ derived in (52)*. With $\mathbb{E}[\|v_k\|^2] \le \sigma^2/t$, equation (55) in the proof of Lemma 2 gives*

$$\mathbb{E}_{U_k}\big[\|e_{k+1}\|^2\big] \le \rho\,\mathbb{E}_{U_k}\big[\|e_k\|^2\big] + \frac{2h^2\sigma^2}{t},$$

where $\rho = 1 - h\beta + 2L_1^2h^2$. Under (97)*, the additive term becomes state-dependent. Using the L_1-Lipschitz continuity of ∇f and $\nabla f(z^*) = 0$, we have*

$$\|\nabla f(\tilde{z}_k)\|^2 = \|\nabla f(\tilde{z}_k) - \nabla f(z^*)\|^2 \le L_1^2\|\tilde{z}_k - z^*\|^2 \le 2L_1^2\|e_k\|^2 + 2L_1^2\|\bar{z}_k - z^*\|^2,$$

where the last inequality uses $\|\tilde{z}_k - z^*\|^2 \le 2\|e_k\|^2 + 2\|\bar{z}_k - z^*\|^2$. Defining $\Sigma_1 := \sigma_1^2L_1^2$, in place of (55) we obtain

$$\mathbb{E}_{U_k}\big[\|e_{k+1}\|^2\big] \le \left(\rho + \frac{4h^2\Sigma_1}{t}\right)\mathbb{E}\big[\|e_k\|^2\big] + \frac{2h^2\sigma_0^2}{t} + \frac{4h^2\Sigma_1}{t}\|\bar{z}_k - z^*\|^2. \tag{98}$$

Unlike the case of Assumption 1, bounding $\mathbb{E}[\|e_k\|^2]$ alone is no longer sufficient, since (98) *couples the error $e_k = \tilde{z}_k - \bar{z}_k$ to the averaged trajectory $\bar{z}_k$. We therefore analyse the joint evolution of $a_k := \mathbb{E}[\|e_k\|^2]$ and $b_k := \|\bar{z}_k - z^*\|^2$. Equation* (98) *gives*

$$a_{k+1} \le \rho'\,a_k + \frac{2h^2\sigma_0^2}{t} + \frac{4h^2\Sigma_1}{t}\,b_k, \tag{99}$$

where $\rho' := \rho + 4h^2\Sigma_1/t$. For the b_k recursion, recall the averaged dynamics (16): $\bar{z}_{k+1} = \bar{z}_k - h\nabla f(\bar{z}_k) + q_k$. Expanding, we have

$$\begin{aligned} b_{k+1} &= \|\bar{z}_k - h\nabla f(\bar{z}_k) + q_k - z^*\|^2 \\ &= \|\bar{z}_k - z^* - h\nabla f(\bar{z}_k)\|^2 + \|q_k\|^2 + 2\,\langle \bar{z}_k - z^* - h\nabla f(\bar{z}_k),\, q_k\rangle \\ &\le \left(\sqrt{(1-c)\,b_k} + \|q_k\|\right)^2, \end{aligned}$$

where $c = 2h\beta L_1/(\beta + L_1) \in (0,1)$ is the GD contraction constant derived in [Nes18, Thm 2.1.15]. Using Young's inequality, for any $\varepsilon > 0$,

$$b_{k+1} \le (1+\varepsilon)(1-c)\,b_k + \big(1 + \tfrac{1}{\varepsilon}\big)\,\|q_k\|^2. \tag{100}$$

From (46) *and Jensen's inequality, $\|q_k\|^2 \le 2\,\mathrm{bias}^2 + 2h^2L_1^2\,a_k$, where $\mathrm{bias} := h\mu L_1(n+3)^{3/2}/2$ denotes the smoothing bias term. Substituting into* (100) *yields*

$$b_{k+1} \le (1+\varepsilon)(1-c)\,b_k + 2\big(1 + \tfrac{1}{\varepsilon}\big)\,h^2L_1^2\,a_k + 2\big(1 + \tfrac{1}{\varepsilon}\big)\,\mathrm{bias}^2. \tag{101}$$

Combining (99) *and* (101) *and considering their worst-case setting, i.e., equality, the coupled system can be written in matrix form as*

$$\begin{bmatrix} a_{k+1} \\ b_{k+1} \end{bmatrix} = M\begin{bmatrix} a_k \\ b_k \end{bmatrix} + \begin{bmatrix} 2h^2\sigma_0^2/t \\ 2(1+1/\varepsilon)\,\mathrm{bias}^2 \end{bmatrix}, \tag{102}$$

where

$$M = \begin{bmatrix} \rho' & \dfrac{4h^2\Sigma_1}{t} \\ 2\left(1 + \dfrac{1}{\varepsilon}\right)h^2L_1^2 & (1+\varepsilon)(1-c) \end{bmatrix}. \tag{103}$$

If spectral radius of M is less than 1, the coupled system (102) converges to a neighbourhood of the origin determined by $(I - M)^{-1}$ applied to the constant vector, and both $\sup_k a_k$ and $\sup_k b_k$ are uniformly bounded, which means that $sup_k\|q_k\|$ is uniformly bounded. A sufficient condition for spectral radius of M being less than 1 is that both diagonal entries are strictly less than 1 and

$$(1-\rho')\big(1-(1+\varepsilon)(1-c)\big) > \frac{4h^2\Sigma_1}{t}\cdot 2\big(1+\tfrac{1}{\varepsilon}\big)\,h^2L_1^2, \tag{104}$$

This follows from standard bounds on the spectral radius of a nonnegative matrix [HJ12, Ch 1, 5, and 8]. The condition $\rho' < 1$ is satisfiable for any $\Sigma_1 > 0$ by choosing t sufficiently large or h sufficiently small. The condition $(1+\varepsilon)(1-c) < 1$ is satisfied for $\varepsilon < \frac{1}{1-c} - 1$. The cross condition (104) has a right-hand side of order $O(h^4\Sigma_1 L_1^2/t)$ versus a left-hand side of order $O(h^2)$, which is satisfiable for h small enough or t large enough.

Once uniform boundedness of a_k and b_k is established, the perturbation bound takes the form

$$\sup_k \|q_k\| \le \text{bias} + hL_1 \sup_k \sqrt{a_k},$$

which is finite and controllable through the parameters h, μ, and t. The remainder of the ISS argument (Theorem 3) then applies without modification, yielding the same exponential decay rate and convergence to a neighbourhood whose radius depends on the perturbation bound. The key difference compared to Assumption 1 is that t must now be chosen large enough or h be chosen small enough to ensure that the spectral radius of M is less than 1, introducing a coupling between the variance parameter $\Sigma_1 = \sigma_1^2 L_1^2$ and the GD contraction rate c.

Remark 5. *Setting $h_2 = 0$ reduces the heavy ball method to gradient descent. In this case, ρ_{HB} reduces to $\rho_{\mathrm{GD}} = 1 - h_1\beta + 2h_1^2L_1^2$, and the bound recovers the result of the GD case in Section 3.1.*

Remark 6. *The bound (26) consists of two terms: the first term, $\frac{h_1\mu\, L_1(n+3)^{3/2}}{2}$, is the smoothing bias and which depends on the smoothing parameter μ; the second term captures the variance of the ZO oracle amplified through the HB dynamics and can be manipulated through the number of samples t, the step size h_1, and the momentum parameter h_2 (through ρ_{HB}).*

Remark 7. *Setting $h_2 = 0$ reduces NAG to gradient descent. In this case, ρ_{NAG} reduces to $\rho_{\mathrm{GD}} = 1 - h_1\beta + 2h_1^2L_1^2$, and the bound recovers the result of the GD case in Section 3.1.*

Remark 8. *The bound (28) has the same two-term structure as the HB bound (26): the first term is the smoothing bias (depending on μ), and the second term captures the ZO oracle variance amplified through the dynamics (and depending on t, h_1, and h_2 through ρ_{NAG}).*

Remark 9. *We note that the step-size condition in Lemma 5, combined with the parametrisation $\lambda = \frac{1-c}{2h}$, implies $\lambda > \frac{L_1}{2}$, so the regularised objective is strongly convex across the entire admissible parameter range. Nevertheless, the L_2 regularisation framework remains valuable: it provides a systematic mechanism for applying the ISS analysis to objectives that are not originally strongly convex, with explicit control over the trade-off between the regularisation bias and the convergence neighbourhood radius through the parameters c, h, μ, and t. Extending the analysis to parameter regimes where the regularised objective is not strongly convex is an interesting direction for future work.*

Remark 10. *The same approach of Section 3.3 applies to the regularised problem (31) with HB or NAG: one can show that their ZO counterparts are perturbed versions of the corresponding FO algorithms with bounded and controllable perturbations, yielding results analogous to Theorem 6. The details are omitted for brevity.*

E Numerical examples complementary details

In this section, we present complementary experiments and results for the numerical examples given in Section 4. All the tests have been run on a Dell Latitude 7430 Laptop with a 12th Gen Intel Core i7 CPU.

E.1 Additional experiments for the quadratic objective

This section complements the quadratic experiment in Section 4.1 with additional parameter studies.

Effect of the smoothing parameter μ. We fix $t = 1$ and $h = 10^{-5}$ and vary $\mu \in \{10^{-1}, 4 \times 10^{-3}, 10^{-6}, 10^{-8}\}$. The results are shown in Figure 4. For large μ, ZO-GD fails to converge; reducing μ restores convergence. However, once μ is sufficiently small, further reduction has negligible effect, since the dominant term in the perturbation bound no longer depends on μ. This is in contrast to h, which controls both terms in the perturbation bound.

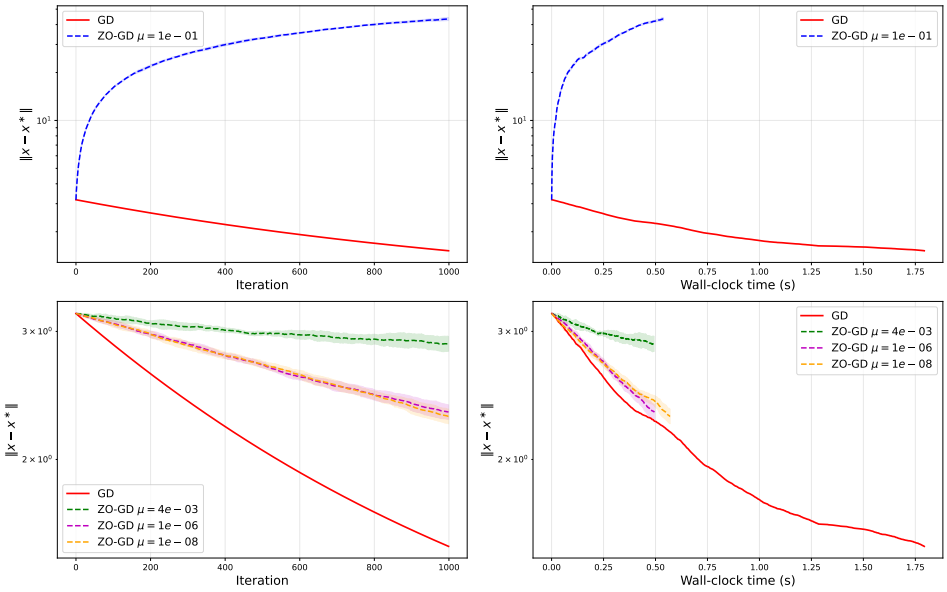

Figure 4: Parameter analysis: smoothing parameter.

Effect of the number of sampled directions t. We fix $\mu = h = 10^{-5}$ and vary $t \in \{1, 5, 10, 20\}$. The results are shown in Figure 5. Increasing t reduces the perturbation, and ZO-GD converges to a neighbourhood of GD that shrinks with t.

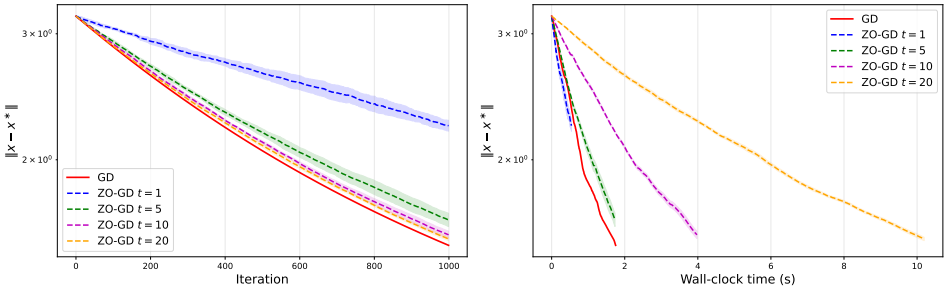

Figure 5: Parameter analysis: number of sampled directions.

Parameter compensation. We revisit the cases in which GD converged, but ZO-GD diverged, demonstrating that the perturbation can always be controlled by adjusting the remaining parameters. First, consider the divergent case $h = 10^{-4}$, $\mu = 10^{-5}$, $t = 1$ from Figure 1. As shown in the first row of Figure 6, increasing the number of sampled directions to $t = 25$ and reducing $\mu = 10^{-7}$ brings the perturbation within the admissible range. Second, consider the divergent case $\mu = 0.01$, $t = 1$, $h = 10^{-5}$ from Figure 4. Since μ and t do not interact directly in the perturbation bound, we instead reduce h to 10^{-7}. The result, shown in the second row of Figure 6, confirms that ZO-GD again tracks GD.

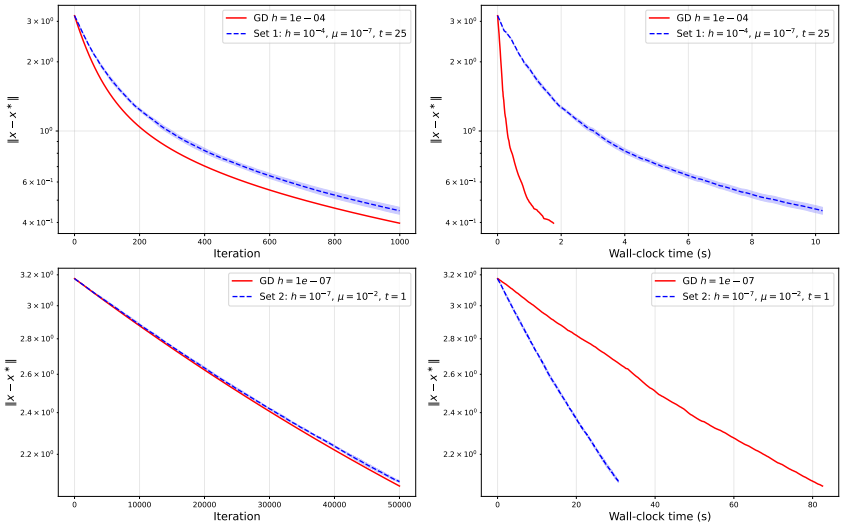

Figure 6: Parameter analysis: compensation for divergent cases.

E.2 Additional experiments for neural network classification

This section complements the neural network experiment in Section 4.2 by providing a parameter sensitivity analysis.

Parameter sensitivity. We conduct parameter sweeps for ZO-GD analogous to those in Appendix E.1. Figure 7 shows the effect of varying h, μ, and t individually while holding the other two fixed, for the network with $100,609$ parameters. The same phenomena observed for the quadratic objective are reproduced: when h is too large, ZO-GD diverges, and reducing h brings ZO-GD progressively closer to GD (Figure 7, left). Reducing μ restores convergence when it is initially too large, but further reduction has diminishing effect once the variance term dominates the perturbation bound (Figure 7, middle). Increasing the number of sampled directions t shrinks the neighbourhood, with ZO-GD converging closer to GD for larger t (Figure 7, right). These observations are consistent with the two-term structure of the perturbation bound in Lemma 5: h and μ control the smoothing bias term, while h and t control the variance term.

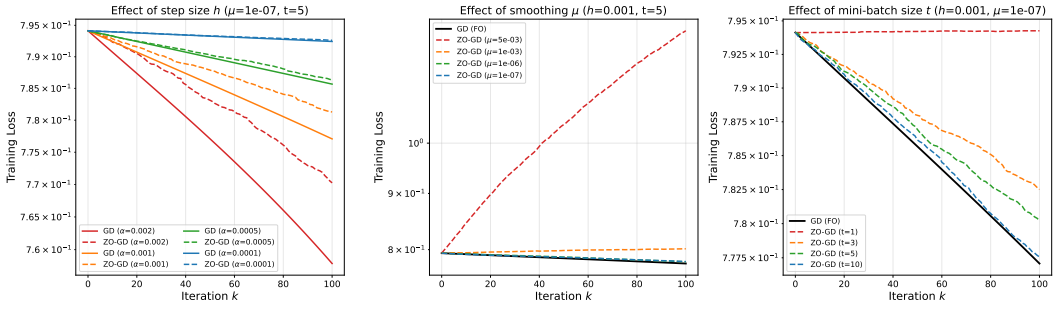

Figure 7: Parameter sensitivity for ZO-GD on the neural network. Left: effect of h. Centre: effect of μ. Right: effect of t. Solid black: GD (FO); dashed coloured: ZO-GD.

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: All the claims made in the abstract and introduction is supported via lemmas, theorems, and proofs in Sections 2 and 3, Appendices B and C.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: The main limitations of the results of this work have been discussed in Sections 3 and 5 and Appendix D.

   Guidelines:

   - The answer [N/A] 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 assumptions used in the main results and the proofs of lemmas and theorems are presented in Sections 2 and 3, Appendices B and C.

Guidelines:

- The answer [N/A] 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: All the needed information for reproducibility of the numerical experiments are given in Section 4 and Appendix E. Moreover, the codes used for numerical experiments are attached as supplementary files.

Guidelines:

- The answer [N/A] 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: The codes used for numerical experiments are attached as supplementary files and provided according to the NeurIPS guidelines.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: All the information of the numerical experiments are given in Section 4 and Appendix E.

   Guidelines:

   - The answer [N/A] 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: [Yes]

   Justification: The theoretical results are in expectation. The numerical experiments report empirical averages over multiple independent runs, with shaded regions indicating standard deviation in the figures.

   Guidelines:

   - The answer [N/A] 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: The results are reported in CPU time as well as number of iterations. Moreover, the detail of the hardware used for numerical experiments is given in Appendix E.

   Guidelines:

   - The answer [N/A] 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: The NeurIPS code of ethics have been reviewed and followed by the authors.

   Guidelines:

   - The answer [N/A] 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: [N/A]

    Justification: This is a theoretical contribution in optimisation that establishes convergence properties of zeroth-order methods. It does not introduce a new application or deployment scenario, and we do not foresee specific societal impacts beyond the general advancement of optimisation theory.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work is theoretical and does not release any data, models, or software that could pose a risk for misuse.

    Guidelines:

    - The answer [N/A] 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: [Yes]

    Justification: The only external dataset used is MNIST, which is publicly available and widely used as a benchmark in the machine learning literature. All prior works upon which our analysis builds are properly cited.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This paper does not introduce new assets such as datasets, models, or code libraries.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This work does not involve human subjects or participants.

    Guidelines:

- The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as a component of the core methodology in this research.

    Guidelines:

    - The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.