Escaping Saddle Points in Zeroth-order Optimization: The Power of Two-point Estimators

Zhaolin Ren [1] **Yujie Tang** [2] **Na Li** [1]

Abstract

Two-point zeroth order methods are important in many applications of zeroth-order optimization, such as robotics, wind farms, power systems, online optimization, and adversarial robustness to black-box attacks in deep neural networks, where the problem may be high-dimensional and/or time-varying. Most problems in these applications are nonconvex and contain saddle points. While existing works have shown that zeroth-order methods utilizing $\Omega(d)$ function valuations per iteration (with d denoting the problem dimension) can escape saddle points efficiently, it remains an open question if zeroth-order methods based on two-point estimators can escape saddle points. In this paper, we show that by adding an appropriate isotropic perturbation at each iteration, a zeroth-order algorithm based on $2m$ (for any $1 \leq m \leq d$) function evaluations per iteration can not only find ϵ-second order stationary points polynomially fast, but do so using only $\tilde{O}(d/m\epsilon^2\bar{\psi})$ function evaluations, where $\bar{\psi} \geq \tilde{\Omega}(\sqrt{\epsilon})$ is a parameter capturing the extent to which the function of interest exhibits the strict saddle property.

1. Introduction

Two-point estimators, which approximate the gradient using two function evaluations per iteration, have been widely studied by researchers in the zeroth-order optimization literature, in convex (Nesterov & Spokoiny, 2017; Duchi et al., 2015; Shamir, 2017), nonconvex (Nesterov & Spokoiny, 2017), online (Shamir, 2017), as well as distributed settings (Tang et al., 2019). A key reason for doing so is that for applications of zeroth-order optimization arising in robotics (Li et al., 2022), wind farms (Tang et al., 2020a), power systems (Chen et al., 2020), online (time-varying) optimization (Shamir, 2017), learning-based control (Malik et al., 2019; Li et al., 2021), and improving adversarial robustness to black-box attacks in deep neural networks (Chen et al., 2017), it may be costly or impractical to wait for $\Omega(d)$ (where d denotes the problem dimension) function evaluations per iteration to make a step. This is especially true for high-dimensional and problems with time-varying noise. See Appendix A for more discussion.

However, despite the advantages of zeroth-order methods with two-point estimators, there has been a lack of existing work studying the ability of two-point estimators to escape saddle points in nonconvex optimization problems. Since nonconvex problems arise often in practice, it is crucial to know if two-point algorithms can efficiently escape saddle points of nonconvex functions and converge to second-order stationary points (see Definition 1 for a definition).

To motivate the challenges of escaping saddle points using two-point zeroth-order methods, we begin with a review of escaping saddle points using first-order methods. The problem of efficiently escaping saddle points in deterministic first-order optimization (with exact gradients) has been carefully studied in several earlier works (Jin et al., 2017; 2018b). A key idea in these works is the injection of an isotropic perturbation whenever the gradient is small, facilitating escape from a saddle if a negative curvature direction exists even without actively identifying the direction. However, the analysis of efficient saddle point escape for stochastic gradient methods is often more complicated. In general, the behavior of the stochastic gradient near the saddle point can be difficult to characterize. Hence, strong concentration assumptions are typically made on the stochastic gradients being used, such as subGaussianity, boundedness of the variance or a bounded gradient estimator (Ge et al., 2015; Daneshmand et al., 2018; Xu et al., 2018; Fang et al., 2019; Roy et al., 2020; Vlaski & Sayed, 2021b), creating an analytical issue when such idealized assumptions fail to hold.

Indeed, though zeroth-order methods can be viewed as stochastic gradient methods, common zeroth order estimators, such as two-point estimators (Nesterov & Spokoiny, 2017), are not subGaussian, and can have unbounded vari-

[1] John A. Paulson School of Engineering and Applied Sciences, Harvard University [2] Department of Industrial Engineering & Management at Peking University. Correspondence to: Zhaolin Ren <zhaolinren@g.harvard.edu>, Na Li <nali@seas.harvard.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

ance. For instance, it can be shown that the variance of the two-point estimator is on the order of $\Omega(d\|\nabla f(x)\|^2)$ (Nesterov & Spokoiny, 2017), with both a dependence on the problem dimension d as well as on the norm of the gradient, which can be unbounded. Due to non-subGaussianity and unboundedness, it is tricky to bound the effect of such zeroth-order estimators and establish tight concentration inequalities that facilitate its escape near saddle points. In addition, the large variance of the zeroth-order estimator is also an issue in non-saddle regions, i.e. when the gradient is large. While this is not an issue to show function improvement in expectation, as we discuss later, this becomes an issue when guaranteeing high probability bounds.

Due to these difficulties, previous works on escaping saddle points in zeroth-order optimization have exclusively focused on approaches requiring $\Omega(d)$ function evaluations per iteration to accurately estimate the gradient (Jin et al., 2018a; Bai et al., 2020; Vlatakis-Gkaragkounis et al., 2019), or in some cases negative curvature directions (Zhang et al., 2022; Lucchi et al., 2021) or the Hessian itself (Balasubramanian & Ghadimi, 2022), reducing in a sense the zeroth-order problem back to a first-order one. However, as explained earlier, two-point zeroth-order algorithms are important for high-dimensional and/or time-varying problems in many applications areas. This raises an important question:

Can two-point zeroth-order methods escape saddle points and reach approximate second order stationary points efficiently?

Our Contribution. In this work, we show that by adding an appropriate isotropic perturbation at each iteration, a zeroth-order algorithm based on *any* number m of pairs ($1 \leq m \leq d$) of function evaluations per iteration can not only find $(\epsilon, \sqrt{\epsilon})$-second order stationary points (cf. the definition later in Definition 1) polynomially fast, but do so using only $\tilde{O}(\mathrm{polylog}(\frac{1}{})d/\epsilon^{2:5})$ function evaluations, with a probability of at least $1-\delta$. In particular, this proves that using a single two-point zeroth-order estimator at each iteration (with appropriate perturbation) suffices to efficiently escape saddle points in zeroth-order optimization, with high probability. Moreover, for functions that are $(\epsilon, \psi, O(\sqrt{\epsilon}))$ strict-saddle (see Definition 3 for a definition of strict saddle functions), our results become $\tilde{O}(\mathrm{polylog}(\frac{1}{})d/\psi\epsilon^2)$, which is a significant improvement when $\psi \gg \epsilon$; strict saddle functions have been identified as an important class of functions in nonconvex optimization, with several well-known examples such as tensor decomposition (Ge et al., 2015), dictionary learning and phase retrieval (Sun et al., 2015). A comparison of our results with existing zeroth-order and first-order methods is shown in Table 1. We also provide numerical results in Section 4 showing that our proposed two-point algorithm requires fewer total function evaluations to converge than zeroth order methods that use $2d$ function evaluations per iteration, for a nonconvex test function proposed in (Du et al., 2017).

To overcome the theoretical challenges that were discussed earlier, we i) first show, via a careful analysis, that zeroth order methods can make function value improvement across iterates with large gradients with high probability, even when only a single two-point estimator (which can have significant variance at large gradients) is used per iteration. ii) Second, near saddle points, we overcome issues caused by the unbounded variance and non-subGaussinity of zeroth-order gradient estimators by developing new technical tools, including novel martingale concentration inequalities involving Gaussian vectors, to tightly bound such terms. In turn, this allows us to show that the noise emanating from the zeroth-order estimators will not overwhelm the effect of the additional isotropic perturbative noise, facilitating escape along negative curvature directions. To the best of our knowledge, both analyses are novel, and may be independent contributions on their own.

Related Work. Due to space considerations, we defer a full discussion of related work to Appendix A.

2. Problem Setup

We make the following assumptions on the class of functions $f : \mathsf{R}^d \to \mathsf{R}$ which we consider.

Assumption 1 (Properties of f)**.** *We suppose that $f : \mathsf{R}^d \to \mathsf{R}$ satisfies the following properties:*

1. *f is twice-differentiable and lower bounded, i.e. $f \ :=$ $\min_x f(x) > -\infty$.*
2. *f is L-gradient Lipschitz, i.e.*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\| \ \forall x, y \in \mathsf{R}^d.$$

3. *f is ρ-Hessian Lipschitz, i.e.*

$$\nabla^2 f(x) - \nabla^2 f(y) \ \ \leq \rho\|x-y\| \ \forall x, y \in \mathsf{R}^d.$$

In our work, we focus on finding approximate second order stationary points, defined below.

Definition 1. A point $x \in \mathsf{R}^d$ is an (ϵ, φ)-second order stationary point if

$$\|\nabla f(x)\| < \epsilon, \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(x)) > -\varphi.$$

We define an (ϵ, φ)-approximate saddle point as follows.

Definition 2. A point $x \in \mathsf{R}^d$ is an (ϵ, φ)-approximate saddle point, if

$$\|\nabla f(x)\| < \epsilon, \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(x)) \leq -\varphi.$$

		Iteration Complexity	Fun. Evaluations. per iter
First-order	(Jin et al., 2017) (deterministic)	$\tilde{O} \ \frac{1}{\epsilon^2}$	—
	(Fang et al., 2019) (SGD)	$\tilde{O} \ \frac{1}{\epsilon^{3:5}}$	—
Zeroth-order	(Jin et al., 2018a)	$\tilde{O} \ \frac{1}{\epsilon^2}$	$\tilde{O} \ \frac{d^2}{\epsilon^3}$
	(Bai et al., 2020)	$\tilde{O} \ \frac{1}{\epsilon^2}$	$\tilde{O} \ \frac{d^2}{\epsilon^8}$
	(Vlatakis-Gkaragkounis et al., 2019)	$\tilde{O} \ \frac{1}{\epsilon^2}$	$\tilde{O}\,(d)$
	(Balasubramanian & Ghadimi, 2022)	$\tilde{O} \ \frac{1}{\epsilon^{1:5}}$	$\tilde{O} \ \frac{d}{\epsilon^2} + \frac{d^4}{\epsilon}$
	(Lucchi et al., 2021)^{y}	$\tilde{O} \ \frac{1}{\epsilon^2}$	$\tilde{O} \ \frac{d}{\epsilon^{2=3}}$
	(Zhang et al., 2022)	$\tilde{O} \ \frac{1}{\epsilon^2}$	$\tilde{O}(d)$
	Algorithm 1 (this paper, $1 \leq m \leq d$)‡	$\tilde{O} \ \frac{d}{\epsilon^2 \bar{\psi} m}$	$2m$

Table 1: Selected comparison of convergence results to $(\epsilon, O(\sqrt{\epsilon})$-second order stationary points in smooth, nonconvex functions; for ^{y}, the convergence is to $(\epsilon, \epsilon^{2/3})$-second order stationary points. For ^{z}, the term $\bar{\psi}$ in the denominator is (i) ψ when the function f is $(\epsilon, \psi, O(\sqrt{\epsilon}))$-strict saddle for a $\psi > O(\sqrt{\epsilon})$ (see Definition 3 for a definition) and (ii) $O(\sqrt{\epsilon})$ if otherwise.

Following past convention (Jin et al., 2019a), we will focus in particular on escaping $(\epsilon, \sqrt{\rho\epsilon})$-saddle points. For notational simplicity, in following text, we refer to $(\epsilon, \sqrt{\rho\epsilon})$-saddle points simply as ϵ-saddle points and $(\epsilon, \sqrt{\rho\epsilon})$-second order stationary points as ϵ-second order stationary points. Beyond the definition of ϵ-approximate saddle points above, it is known that many nonconvex functions with saddle points, such as orthogonal tensor decomposition (Ge et al., 2015), phase retrieval and dictionary learning (Sun et al., 2015), satisfy what is known as a *strict saddle* condition (Ge et al., 2015). For the Hessians of the saddle points of such functions, there is always a strict negative eigenvalue whose magnitude is bounded from below. We provide a precise definition below.

Definition 3. A twice-differential function $f(x)$ is $(\epsilon, \psi, \varrho)$-strict saddle for any $\psi > \varrho > 0$, if for any point x, either

1. $\|\nabla f(x)\| \geq \epsilon$ holds,
2. or when $\|\nabla f(x)\| < \epsilon$ holds, either
   (a) $\lambda_{\min}(\nabla^2 f(x)) \leq -\psi$, or
   (b) $\lambda_{\min}(\nabla^2 f(x)) > -\varrho$.

In our work, we consider the following batch symmetric two-point zeroth-order estimator.

Definition 4 ((Batch) two-point zeroth-order estimator with perturbation)**.** We define a m-batch two-point zeroth order estimator as follows:

$$g_u^{(m)}(x) := \frac{1}{m} \sum_{i=1} \frac{f(x + uZ_i) - f(x - uZ_i)}{2u} Z_i, \quad (1)$$

where $Z_i \overset{i.i.d}{\sim} N(0, I)$, and $u > 0$ is a smoothing radius.

Algorithm 1 Zeroth-order perturbed gradient descent (ZOPGD)

```
input: $x_0$, horizon $T$, step-size $\eta$, smoothing radius $u$, perturbation radius $r$, batch size $m$
for step $t = 0, \ldots, T$ do
    Sample $Z^{(m)} = \{Z_{t,i}\}_{i=1}^m \sim N(0, I)$ to compute $g_u^{(m)}(x_t))$, defined in Eq. (1).
    Update $x_{t+1} = x_t - \eta \ \ g_u^{(m)}(x_t) + Y_t$ , where $Y_t \sim N(0, \frac{r^2}{d} I)$
```

Such $2m$ zeroth-order gradient estimators have frequently been studied in zeroth-order optimization works (see e.g. (Nesterov & Spokoiny, 2017)). To facilitate efficient escape from saddle points, our proposed Algorithm 1 adds isotropic perturbation at each iteration.

We now state an informal version of our main result, and follow that with a few remarks.

Theorem 1 (Main result, informal version of Theorem 2)**.** *Consider running Algorithm 1. Let $\tilde{O}$ hide polylogarithmic terms in δ and other parameters. Suppose $\delta \in (0, 1/e]$. Suppose $\sqrt{\rho\epsilon} \leq \min\{1, L\}$[1], such that $\bar{\psi} \leq \min\{1, L\}$, where*

:= (min f ; 1; L g *if* 9 > p— *s.t.* f ( ) *is* (; ; p—)*-strict saddle*
p— *if otherwise:*

(2)

[1] In our paper, we focus on the case p— L; otherwise, by the L-Lipschitz assumption, min(r 2 f (x)) L for all x 2 R^d, which implies -first order stationary points are also -second order stationary points.

Suppose

$$u = O \quad \frac{\min f^{p}{}_{-}; {}^{p}{}_{-}\bar{r}g}{p^{-}d} \quad ; r = O(\ ); \quad = O \quad \frac{m}{d \max f L; L^2 g} \quad ;$$

Then, in

$$T = \tilde{\Omega} \quad \frac{(f(x_0) - f\ \ )}{\eta\epsilon^2} + \frac{\rho^2(f(x_0) - f\ \ )}{\eta\bar{\psi}^4}$$
$$= \tilde{\Omega} \quad \frac{d \max\{L, L^2\}\rho^2(f(x_0) - f\ \ )}{m\bar{\psi}\epsilon^2}$$

iterations (with each iteration using $2m$ function evaluations), with probability at least $1 - \delta$, at least half the iterates are $(\epsilon, \sqrt{\rho\epsilon})$-second-order stationary points.

Remark 1. As the choice of η in Proposition 4 (Appendix D) and Theorem 2 (Appendix F) respectively imply, the $\tilde{\Omega} \quad \frac{f(x_0)\ \ f}{\eta\epsilon^2}$ term in the sample complexity comes from the large gradient iterations (Proposition 4), whereas the $\tilde{\Omega} \quad \frac{\rho^2(f(x_0)\ \ f\ \ )}{\eta\bar{\psi}^4}$ term comes from the escape saddle point phase.

Remark 2. As a corollary of Theorem 1, for functions f which are $(\epsilon, \psi, \sqrt{\rho\epsilon})$ strict saddle, assuming that $\psi \geq \sqrt{\rho\epsilon}$, the sample complexity of our algorithm scales as $\tilde{\Omega} \quad \frac{d \max f L^2, L g(f(x_0)\ \ f\ \ )}{m\epsilon^2\psi}$, which scales as $\tilde{\Omega} \quad \frac{d}{m\epsilon^2}$ when ψ is of size $\Omega(1)$. Thus, in this setting, for two-point estimators, where $m = 1$, the dependence on d and ϵ in our sample complexity (as measured by function evaluations) matches that achieved by the algorithms in (Vlatakis-Gkaragkounis et al., 2019; Zhang et al., 2022), which have to use $2d$ function evaluations per iteration to estimate the gradient.

Comparison to gradient-based methods. For first-order escape saddle point algorithms, standard perturbation-based methods (without acceleration) can find a $(\epsilon, O(\sqrt{\epsilon}))$-second-order stationary point using $\tilde{O}(1/\epsilon^2)$ iterations for deterministic GD (Jin et al., 2019a), while for standard SGD the best-known rates are slower at $\tilde{O}(1/\epsilon^{3.5})$ (Fang et al., 2019). In contrast, our sample complexity (as measured by the total number of function evaluations) is $\tilde{O} \quad \frac{d}{\epsilon^2\bar{\psi}}$, where $\bar{\psi}$ is defined in Eq. (2). The extra (linear) dependence on d is typical for zeroth-order algorithms (see e.g. (Nesterov & Spokoiny, 2017)); intuitively, gradient calculation for d-dimensional functions requires $O(d)$ calculations agnostically, so it makes sense that zeroth-order algorithms requires d times more iterations. For general non strict-saddle functions, our dependence on ϵ sits between that of the deterministic methods and SGD methods, and suggests the benefit of a specialized treatment of zeroth-order methods over considering them simply as a subclass of SGD methods. Moreover, for $(\epsilon, \psi, \sqrt{\rho\epsilon})$- strict-saddle functions where $\psi = \Omega(1)$, our sample complexity becomes $\tilde{O}(\frac{d}{\epsilon^2})$, with an ϵ dependence that matches that of the best existing sample complexity for non-accelerated first-order escape saddle point methods (Jin et al., 2017)

Comparison to existing zeroth-order methods. As Table 1 suggests, our sample complexity significantly outperforms that of (Jin et al., 2018a), (Bai et al., 2020), (Balasubramanian & Ghadimi, 2022), and also that in (Lucchi et al., 2021), which is a random search method. We note that the sample complexity in (Vlatakis-Gkaragkounis et al., 2019; Zhang et al., 2022) outperform our method, with a function evaluation complexity of $\tilde{O} \quad \frac{d}{\epsilon^2}$. However, for for $(\epsilon, \psi, \sqrt{\rho\epsilon})$- strict-saddle functions where $\psi = \Omega(1)$, our sample complexity becomes $\tilde{O}(\frac{d}{\epsilon^2})$, which matches the sample complexity in (Vlatakis-Gkaragkounis et al., 2019; Zhang et al., 2022). Moreover, a key limitation of their methods is a requirement to use $\Omega(d)$ function evaluations to estimate the gradient at each iteration, which may not be practical in realistic applications when d is large. In contrast, our method supports any number of function evaluations at each iteration between 1 to d. Moreover, numerically, we found that for a test nonconvex function proposed in (Du et al., 2017), our method (with two-point estimators) takes fewer function evaluations to escape saddle points and converge to the global minimum than the methods in (Vlatakis-Gkaragkounis et al., 2019; Zhang et al., 2022); see Section 4 for details.

3. Proof strategy and key challenges in the zeroth-order setting

Broadly speaking, our proof include two major parts, i) characterizing the progress made in iterations when the gradient is large (which we can define to be iterations t where $\|\nabla f(x_t)\| \geq \epsilon$) (Section 3.1), ii) and iterations when we are at an ϵ-approximate saddle point (where progress may be made along the negative eigendirection of the Hessian matrix) (Section 3.2). While the approach is similar to the first-order case (e.g. (Jin et al., 2019a)), the zeroth-order setting brings forth several unique challenges. In the rest of this section, we explain these challenges, sketch out our high-level proof outlines, and provide statements of the main technical results. Due to limited space, we defer the full proof to the Appendix.

3.1. Showing function decrease when gradients are large

Challenge. Due to the noise in two-point (or $2m$ where m is a small constant) zeroth-order gradient, even when the gradient is large, it may not always be possible to make progress at each iteration, especially when $m < d$ is used in the gradient estimation equation in Eq. (1). While it is tempting to use an expectation-based argument to handle this issue, it is known that expectation-based function decrease arguments are insufficient for the purpose of escaping saddle points (see e.g. Proposition 1 in (Ziyin et al., 2021)). We tackle this issue by using high-probability arguments instead; we note that achieving these high-probability bounds is highly nontrivial due to the large variance of the two-point zeroth-order estimator (scaling with d times the squared norm of the gradient). Hence, any single iteration of the zeroth-order method may in fact lead to a function increase rather than decrease.

High-level proof outline. (i) We first characterize the function value change for our proposed algorithm (Lemma 1). **(ii)** Next, we tackle the issue of the possibility that the function value might increase for any given iteration. The key idea here is that across any small consecutive number of iterations, there will be one iteration where the zeroth-order estimator is sufficiently aligned with the gradient direction (Lemma 14 in Appendix D). **(iii)** Along with a series of other technical results in Appendix D, we then show that the function makes sufficient progress across the duration of the algorithm, with high probability (Proposition 1). To more concretely illustrate the key analytical challenge, we next introduce the following function decrease lemma, proved in Appendix D.

Lemma 1 (Function decrease for batch zeroth-order optimization)**.** *Suppose at each time t, the algorithm performs the update step (with batch-size parameter $1 \le m \le d$)*

$$x_{t+1} = x_t - \eta \quad g_u^{(m)}(x_t) + Y_t \quad ,$$

where

$$g_u^{(m)}(x_t) = \frac{1}{m}\sum_{i=1}^{m}\frac{f(x_t+uZ_{t,i}) - f(x_t-uZ_{t,i})}{2u}Z_{t,i},$$

where each $Z_{t,i}$ is drawn i.i.d from $N(0,I)$, $u > 0$ is the smoothing radius, and $Y_t \sim N(0, \frac{r^2}{d}I)$ with $r > 0$ denoting the perturbation radius.

Then, there exist absolute constants $c_1 > 0, C_1 \ge 1$ such that, for any $T \in \mathsf{Z}^+$ and $T \ge \tau > 0$, $\alpha > 0$ and $\delta \in (0, 1/e]$, upon defining $\mathcal{H}_{0,\tau}(\delta)$ to be the event on which the inequality

$$f(x_\tau) \quad f(x_0) \tag{3}$$

$$\frac{3}{4}\sum_{t=0}^{\tau\;1}\frac{1}{m}\sum_{i=1}^{m} Z_{t,i}^{\top}\nabla f(x_t) \quad ^2 \tag{4}$$

$$+ \quad - + \frac{c_1 L \;\; ^2 \;\; ^3 d}{m}\sum_{t=0}^{\tau\;1}\|\nabla f(x_t)\|^2$$

$$+ \quad u^4 \;\; ^2 \;\; c_1 d^3 \;\; \log\frac{T}{} \;\; ^3 + \;\; L \;\; ^2 u^4 \;\; ^2 \;\; c_1 d^4 \;\; \log\frac{T}{} \;\; ^4$$

$$+ \;\; c_1 r^2(\;\; + \;\; L)\log\frac{T}{} + \;\; c_1 L \;\; ^2 r^2 \tag{5}$$

is satisfied (where $\chi := \log(C_1 dmT/\delta)$), we have

$$\mathrm{P}(H_{0,\tau}(\;)) \quad 1 \quad \frac{(\;\; + 4)}{T}; \qquad \mathrm{P}(\cap_{\tau=1}^{\tau^0} H_{0,\tau}(\;)) \quad 1 \quad \frac{5 \;\; ^0}{T}$$

for any $0 \le \tau^0 \le T$.

Our goal is to show that we can arrive at a contradiction $f(x_T) < \min_x f(x)$ when there is a large number of steps at which $\|\nabla f(x_t)\| \ge \epsilon$ (Proposition 1). As we can see from Eq. (5), this implies that we need to prove a lower bound of the form

$$\sum_{t=0}^{T\;1}\frac{1}{m}\sum_{i=1}^{m} Z_{t,i}^{\top}\nabla f(x_t) \quad ^2 \qquad \frac{1}{} + \frac{c_1 L \quad ^3 d}{m}\sum_{t=0}^{T\;1}\|\nabla f(x_t)\|^2 \tag{6}$$

for some α which is not too large (an example would be picking α such that it only scales logarithmically in the problem parameters). However, it is tricky to prove such a lower-bound in the zeroth-order setting. In particular, for small batch-sizes m, $\frac{1}{m}\sum_{i=1}^{m} Z_{t,i}^{\top}\nabla f(x_t) \quad ^2$ could be small even as $\|\nabla f(x_t)\|^2$ is large; this is because for each $i \in [m]$, $Z_{t,i}$ could have a negligible component in the $\nabla f(x_t)$ direction. This necessitates a more delicate analysis to prove a bound similar to Eq. (6). Due to space reasons, we defer our more detailed proof approach outline to Appendix D (see the discussion immediately following Lemma 1). The results in Appendix D culminates in the following result which limits the number of large-gradient.

Proposition 1 (Bound on number of iterates with large gradients, informal version of Proposition 4)**.** *Let $\delta \in (0, 1/e]$ be arbitrary. Letting $\tilde{O}$ hide polylogarithmic dependencies on δ (and other parameters), consider choosing u, r, η and T such that*

$$u = O \quad \frac{\sqrt{\;}}{\sqrt{\;\;}d} \quad ; \; r = O(\;); \qquad = O \quad \frac{m}{dL} \quad ;$$

$$T = \tilde{\;} \quad \frac{(f(x_0) \quad f \;\;) + \;\; ^2/L)}{^2} \Big) \quad .$$

Then, with probability at least $1 - O(\delta)$, there are at most $T/4$ iterations for which $\|\nabla f(x_t)\| \ge \epsilon$.

3.2. Making progress near saddle points

Challenge. The noise in two-point zeroth-order estimators makes the analysis around $\epsilon-$approximate saddle points challenging, because the concentration properties of the (non-subGaussian) noise are hard to characterize. Intuitively, a noisier estimator might facilitate easier escape from saddle point. However, without an appropriate concentration bound, the noise may behave in unpredictable ways, preventing escape from saddle regions. Previous analysis of saddle point escape using stochastic estimators typically requires these estimators to satisfy subGaussian properties (Jin et al., 2019a; Fang et al., 2019), which zeroth-order estimators do not satisfy.

High-level proof outline. (i) We first prove a technical result showing that the travelling distance of the iterates can be bounded in terms of the function value decrease (i.e., Improve or Localize, Lemma 2). **(ii)** Next, at any ϵ-saddle point, we consider a coupling argument and define two sequences running near-identical zeroth-order dynamics, differing only in the sign of their perturbative term along the minimum eigendirection of H , which denotes the Hessian of the saddle (Lemma 3). Using Lemma 2 in point (i), if we assume for contradiction that the two sequences both "get stuck" and make little function value progress, the dynamics of the difference between the two sequences will remain small as both sequences remain close to the saddle point. **iii)** However, since the perturbation vectors of the two sequences differ in the (most) negative direction of H, the norm of the the difference of the two sequences will grow exponentially so long as *a).* the sequences remain close to the saddle point (and thus the Hessian has a negative curvature direction) and *b).* the effect of the zeroth-order stochastic noise can be controlled. This leads to a contradiction, implying that sufficient function decrease must have been made (Proposition 5 in Appendix E.3). **(iv)** To show that the zeroth-order stochastic noise can be controlled, we prove one technical result (Proposition 2), providing a concentration bound for the product of (possibly unbounded) subGaussian random vectors that scales linearly with the dimension d. This enables us to control the effect of the zeroth-order noise near saddle points, and is essential in showing that the eventual sample complexity scales linearly with d.

We provide a more detailed proof sketch below, where we elaborate more on our analytical challenges and ideas. We first introduce an informal statement of a key technical result that bounds, with high probability, the travelling distance of the iterates in terms of the function value decrease.

Lemma 2 (Improve or Localize, informal version of Lemma 23)**.** *Consider the perturbed zeroth-order update Algorithm 1. Let $\delta \in (0, 1/e]$ be arbitrary. Consider any $T_s = \tilde\Omega\left(\frac{1}{m}\log(1/\delta)\right)$, and any $t_0 \geq 0$. For any $F > 0$, suppose $f(x_{T_s+t_0}) - f(x_{t_0}) > -F$, i.e. $f(x_{t_0}) - f(x_{T_s+t_0}) < F$. Letting $\tilde O$ hide polylogarithmic terms involving δ, suppose*

$$u = O\left(\frac{\min\{\sqrt{\epsilon}, \sqrt{\rho}\, r\}}{\sqrt{\rho}\, d}\right); \quad r = O\left(\min\left\{\epsilon; \frac{F}{T_s}\right\}\right);$$
$$\eta = O\left(\frac{m\sqrt{\ }}{dL}\right):$$

Then, with probability at least $1 - O\left(\frac{T_s\delta}{T}\right)$ (here $T \geq T_s$ denotes the total number of iterations), for each $\tau \in \{0, 1, \ldots, T_s\}$, we have that

$$\|x_{t_0+\tau} - x_{t_0}\|^2 \leq \phi_{T_s}(\delta, F), \quad \textit{where}$$
$$\phi_{T_s}(\delta, F) = \tilde O\left(\max\left\{T_s, \frac{d}{m}\right\}\right)\eta F + \tilde O(\eta^2\epsilon^2).$$

Intuitively, the above result shows that if little function value improvement has been made, then the algorithm's iterates have not moved much, such that it remains approximately in a saddle region if it started out in a saddle region. Next, Lemma 3 formally introduces the coupling we have mentioned, setting the stage for the rest of our arguments. For notational convenience, in this section, unless otherwise specified, we will assume that the initial iterate x_0 is an ϵ-saddle point.

Lemma 3. *Suppose x_0 is an ϵ-approximate saddle point. Without loss of generality, suppose that the minimum eigendirection of $H := \nabla^2 f(x_0)$ is the e_1 direction (i.e. the first basis vector in $\mathbb{R}^d$), and let γ to denote $-\lambda_{\min}(\nabla^2 f(x_0))$ (note $\gamma \geq \sqrt{\rho\epsilon}$). Consider the following coupling mechanism, where we run the zeroth-order gradient dynamics, starting with x_0, with two isotropic noise sequences, Y_t and Y'_t respectively, where $(Y_t)_1 = -(Y_t)'_1$, and $(Y_t)_j = (Y_t)'_j$ for all other $j \neq 1$. Suppose that the sequence $\{Z_{t,i}\}_{t\in T, i\in[m]}$ is the same for both sequences. Let $\{x_t\}$ denote the sequence with the $\{Y_t\}$ noise sequence, and let the $\{x'_t\}$ denote the sequence with the $\{Y'_t\}$ noise sequence, where $x'_0 = x_0$, and*

$$x'_{t+1} = x'_t - \eta\left(\frac{\sum_{i=1}^m Z_{t,i}Z_{t,i}^\top \nabla f(x'_t) + \frac{u}{2}Z_{t,i}Z_{t,i}^\top \tilde H'_{t,i} Z_{t,i}}{m} + Y'_t\right);$$

and $\tilde H'_{t,i} := \frac{H'_{t,i,+} - H'_{t,i,-}}{2}$, with $H'_{t,i,+} = \nabla^2 f(x'_t + \alpha'_{t,i,+}uZ'_i)$ for some $\alpha'_{t,i,+} \in [0,1]$, and $H'_{t,i,-} = \nabla^2 f(x_t - \alpha'_{t,i,-}uZ'_i)$ for some $\alpha'_{t,i,-} \in [0,1]$. Then, for any $t \geq 0$,

$$\begin{aligned}\hat x_{t+1} &:= x_{t+1} - x'_{t+1} \\ &= \underbrace{-\eta\sum_{\tau=0}^{t}(I-\eta H)^{t-\tau}\hat\xi_{g0}(\tau)}_{W_{g0}(t+1)} \underbrace{-\eta\sum_{\tau=0}^{t}(I-\eta H)^{t-\tau}(H_\tau - H)\hat x_\tau}_{W_H(t+1)} \\ &\quad \underbrace{-\eta\sum_{\tau=0}^{t}(I-\eta H)^{t-\tau}\hat\xi_u(\tau)}_{W_u(t+1)} \underbrace{-\eta\sum_{\tau=0}^{t}(I-\eta H)^{t-\tau}\hat Y_\tau}_{W_p(t+1)}\end{aligned}$$

where

$$\xi_{g0}(t) = \frac{1}{m}\sum_{i=1}^{m}(Z_{t,i}Z_{t,i}^\top - I)\nabla f(x_t);$$
$$\xi'_{g0}(t) = \frac{1}{m}\sum_{i=1}^{m}(Z_{t,i}(Z_{t,i})^\top - I)\nabla f(x'_t);$$
$$\hat\xi_{g0}(t) = \xi_{g0}(t) - \xi'_{g0}(t); \quad \xi_u(t) = \frac{1}{m}\sum_{i=1}^{m}\frac{u}{2}Z_{t,i}Z_{t,i}\tilde H_{t,i}Z_{t,i};$$
$$\xi'_u(t) = \frac{1}{m}\sum_{i=1}^{m}\frac{u}{2}Z_{t,i}Z_{t,i}\tilde H'_{t,i}Z_{t,i}; \quad \hat\xi_u(t) = \xi_u(t) - \xi'_u(t);$$
$$\hat Y_t = Y_t - Y'_t; \quad H_t = \int_0^1 \nabla^2 f(ax_t + (1-a)x'_t)da:$$

Our goal is to show that the dominating term in the evolution of the difference dynamics comes from the W_p term involving the additional perturbation. To this end, we need to bound the remaining terms, W_{g_0}, W_H, W_u. A key technical challenge is to find a precise concentration bound for the $W_{g_0}(t+1)$ term, where

$$W_{g_0}(t+1) = \sum_{\tau=0}^{t} (I - \eta H)^{t-\tau} \left(\frac{\sum_{i=1}^{m} (Z_{\tau,i} Z_{\tau,i}^{\top} - I)(\nabla f(x_\tau) - \nabla f(x_\tau^0))}{m} \right).$$

For the simplicity of discussion, we assume for the time being that $m = 1$, and drop the i index in the subscript of $Z_{\tau,i}$. Since $\mathbb{E}[Z_\tau Z_\tau^{\top}] = I$, heuristically, assuming that $Z_\tau Z_\tau^{\top} - I$ satisfies "nice" concentration properties, utilizing the independence of the Z_τ's across time and the fact that $(I - \eta H) \preceq (1 + \eta\gamma)I$, we would like to show that with high probability,

$$\|W_{g_0}(t)\| \lesssim \eta \sqrt{\sum_{\tau=0}^{t-1} (1+\eta\gamma)^{2(t-1-\tau)} \mathbb{E}\left[\|(Z_\tau Z_\tau^{\top} - I)(\nabla f(x_\tau) - \nabla f(x_\tau^0))\|^2 \mid \mathcal{F}_{\tau-1} \right]} \tag{7}$$

where $\mathcal{F}_{\tau-1}$ is a sigma-algebra containing all randomness up to and including iteration $\tau - 1$, such that x_τ and x_τ^0 are both in $\mathcal{F}_{\tau-1}$, but Z_τ is not. Then, assuming that Eq. (7) holds, since

$$\mathbb{E}\left[\|(Z_\tau Z_\tau^{\top} - I)(\nabla f(x_\tau) - \nabla f(x_\tau^0))\|^2 \mid \mathcal{F}_{\tau-1} \right] = O(d) \|\nabla f(x_\tau) - \nabla f(x_\tau^0)\|^2,$$

it follows that

$$\|W_{g_0}(t)\| \lesssim \eta \sqrt{O(d) \sum_{\tau=0}^{t-1} (1+\eta\gamma)^{2(t-1-\tau)} \|\nabla f(x_\tau) - \nabla f(x_\tau^0)\|^2}$$

With this bound on $\|W_{g_0}(t)\|$, we eventually prove in Proposition 5 in Appendix E.3 that our algorithm escapes any $\epsilon-$saddle point with constant probability and that the $O(d)$ term appearing in the square root term above will eventually lead to an $O(d)$ dependence in the sample complexity[2]. We note that the $O(d)$ dimension dependence matches that of the best-known existing upper bound for finding first-order stationary points in smooth nonconvex zeroth-order optimization (Nesterov & Spokoiny, 2017), and has been conjectured to be the best possible dimension dependence for general smooth nonconvex zeroth-order optimization (Balasubramanian & Ghadimi, 2022).

Key technical challenge The key challenge in the above argument is to show that an equation in the form of Eq. (7) could in fact hold. At first glance, that an inequality such as Eq. (7) should hold is rather non-obvious — this is because while the variable $(Z_\tau Z_\tau - I)(\nabla f(x_\tau) - \nabla f(x_\tau^0)) \mid \mathcal{F}_{\tau-1}$ is mean-zero, it is subExponential rather than subGaussian. In fact, even in the subGaussian case, given a sequence of random vectors $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{t-1}$, such that each $\mathbb{E}[\boldsymbol{x}_\tau \mid \mathcal{F}_{\tau-1}] = 0$, and that each $\boldsymbol{x}_\tau \mid \mathcal{F}_{\tau-1}$ is norm-subGaussian with parameter $\sigma_\tau \in \mathcal{F}_{\tau-1}$ (which is an appropriate generalization of subGaussianity for vectors, proposed in (Jin et al., 2019b)), proving a concentration inequality of the form $\sum_{\tau=0}^{t-1} \boldsymbol{x}_\tau \approx \tilde{O}\left(\sqrt{\sum_{\tau=0}^{t-1} \sigma_\tau^2}\right)$ is a very delicate matter.

In our case, the analogue of $\boldsymbol{x}_\tau$ is $(I - \eta H)^{t-1-\tau}(Z_\tau Z_\tau - I)(\nabla f(x_\tau) - \nabla f(x_\tau^0))$, while the analogue of σ_τ^2 is $(1+\eta\gamma)^{2(t-1-\tau)} \mathbb{E}\left[\|(Z_\tau Z_\tau - I)(\nabla f(x_\tau) - \nabla f(x_\tau^0))\|^2 \mid \mathcal{F}_{\tau-1} \right]$. Existing techniques (cf. (Tropp et al., 2015; Jin et al., 2019b)) rely crucially on subGaussian properties that allow for each τ the moment-generating function $\mathbb{E}[e^{\theta \boldsymbol{Y}} \mid \mathcal{F}_{\tau-1}]$ to be defined for any fixed (and non-random) $\theta > 0$, where $\boldsymbol{Y}_\tau$ takes the form

$$\boldsymbol{Y}_\tau = \begin{pmatrix} 0 & \boldsymbol{x}_\tau^{\top} \\ \boldsymbol{x}_\tau & 0 \end{pmatrix},$$

such that $\mathbb{E}[\boldsymbol{Y}_\tau \mid \mathcal{F}_{\tau-1}] = 0$ (since $\mathbb{E}[\boldsymbol{x}_\tau \mid \mathcal{F}_{\tau-1}] = 0$), and the eigenvalues of $\boldsymbol{Y}_\tau$ are $\pm\|\boldsymbol{x}_\tau\|$. In the case when $\boldsymbol{x}_\tau$ is merely subExponential, the Moment Generating Function (MGF), $\mathbb{E}[e^{\theta \boldsymbol{Y}} \mid \mathcal{F}_{\tau-1}]$, will no longer be well-defined at any fixed (and non-random) $\theta > 0$. This poses a challenge in our setting, since $\boldsymbol{x}_\tau$ takes the form $(I - \eta H)^{t-1-\tau}(Z_\tau Z_\tau^{\top} - I)(\nabla f(x_\tau) - \nabla f(x_\tau^0))$, which is subExponential rather than subGaussian. While it may be possible to force $(I - \eta H)^{t-1-\tau}(Z_\tau Z_\tau^{\top} - I)(\nabla f(x_\tau) - \nabla f(x_\tau^0))$ to be sub-Gaussian, say by normalizing Z_τ to have norm $\sqrt{d}$ (note any bounded random vector is also subGaussian), such that $\|(Z_\tau Z_\tau^{\top} - I)g\|^2 \leq d^2 \|g\|^2$ for any vector $g \in \mathbb{R}^d$, a careful examination of the argument in Proposition 5 would show that this results in a $O(d^2)$ rather than $O(d)$ dependence in the sample complexity, incurring a heavy price on the overall sample complexity (extra factor of d) if d is large.

Our solution To overcome the issue, we build on the following observation: with high probability, for any vector $g \in \mathbb{R}^d$, $|Z_\tau^{\top} g|$ is bounded within some log factor of $\|g\|$. On the event $\{|Z_\tau^{\top} g| = \tilde{O}(\|g\|)\}$, the variable

$$(Z_\tau Z_\tau^{\top} - I)g = Z_\tau (Z_\tau^{\top} g) - g \approx Z_\tau \|g\| - g$$

behaves approximately like a subGaussian random vector since $Z_\tau \sim N(0, I_d)$. Based on this intuition, after some careful analysis, we can show that $(Z_\tau Z_\tau^{\top} - I)(\nabla f(x_\tau) - \nabla f(x_\tau^0)) \mid \mathcal{F}_{\tau-1}$ is subGaussian on the event that $|Z_\tau^{\top} \nabla f(x_\tau)|$ is bounded within some log factor of $\|\nabla f(x_\tau)\|$, which happens with high probability. This then allows us to show that on this event, the corresponding MGF is well-defined for all fixed $\theta > 0$, enabling us

[2] For general $1 \leq m \leq d$, there will also be an $O(1/m)$ dependence in the sample complexity.

to prove a concentration inequality of the form Eq. (7). This intuition is crystallized in the following proposition, which proves a more general bound than what we strictly need. For notational simplicity, we introduce the function $lr(x) := \log(x \log(x))$.

Proposition 2. Let $F_t; t \quad 1$ be a filtration. Let $(Z_t)_{t \quad 0}$ be a sequence of random vectors following the distribution $N(0; I)$ such that $Z_t \; 2 \; F_t$ and is independent of $F_{t \quad 1}$, and let $(v_t)_{t \quad 0}$ be a sequence of random vectors such that $v_t \; 2 \; F_{t \quad 1}$. For each $\quad 0$, let

$$W \quad = \quad X^{1}_{t=0} M_t(Z_t Z_t^{>} \quad I)v_t;$$

where each M_t is a deterministic matrix of appropriate dimension. Then, there exist some absolute constants $c^0; C > 0$ such that for any $\; 2 \; Z^{+}$ and $\; 2 \; (0; 1=e]$, the following statements hold:

1. For any $> \; 0$, with probability at least $1 \quad$, we have

$$kW \; k \quad c^0 \; X^{1}_{t=0} kM_t k_2^2 d(lr(C \; = \;))^2 kv_t k^2 + \frac{1}{} \log(Cd \; = \;):$$

2. For any $B > b > \; 0$, with probability at least $1 \quad$,

$$\text{either} \quad X^{1}_{t=0} kM_t k_2^2 d(lr(C \; = \;))^2 kv_t k^2 \quad B; \; \text{or}$$

$$kW \; k \quad \begin{matrix} v \\ u \\ u \\ t \end{matrix} \overline{\max \left(X^{1}_{t=0} kM_t k_2^2 d(lr(C \; = \;))^2 kv_t k^2 ; b \right)} \; c^0 p \; \overline{(\log(C \; d= \;) + \log(\log(B=b) + 1))}$$

Moreover, as is clear from the bounds above, we may pick $C \quad 1$ such that $\log \frac{C}{} \quad 1; 8 \quad 2 \; (0; \frac{1}{e}]$.

With this result, along with a series of other technical results in Appendix E.3, we can show that the algorithm makes a function decrease of F with $(1) \quad$ probability near an- saddle point (Proposition 5 in Appendix E.3). Armed with Proposition 5, as well as Proposition 1, the main result in Theorem 1 then follows. The complete detailed analysis can be found in Appendix E (escaping saddle point) and Appendix F (main result).

4. Simulations

We test the performance of our proposed algorithm with two-point estimators (ZOPGD-2pt) against existing zeroth-order benchmarks using the octopus function (proposed in (Du et al., 2017))[3]. It is known that the octopus function defined on R^d, which chains d saddle points sequentially, takes exponential (in d) time for exact gradient descent to escape; it has thus emerged as a popular benchmark to evaluate algorithms that seek to escape saddle points. In our experiments, we compare the performance of our two-point estimator algorithm (ZOPGD-2pt) with PAGD (Algorithm 1 in (Vlatakis-Gkaragkounis et al., 2019)) and ZO-GD-NCF (see (Zhang et al., 2022)), which are the only two existing zeroth-order algorithms that have (a) $O(d= \;^2)$ sample complexity for escaping saddle points (with the latter algorithm yielding the tightest bounds), and (b) performed the best empirically on escaping saddle points (see the simulation results in (Zhang et al., 2022)). Both PAGD and ZO-GD-NCF have to use $2d$ function evaluations per iteration to estimate the gradient while our algorithm only needs to use 2 function evaluations. We plot the function value against the number of function evaluations.

We tested the algorithms for $d = 10$ and $d = 30$. To account for the stochasticity in the algorithms, for each algorithm, we computed the average and standard deviation over 30 trials, and plotted the mean trajectory with an additional band that represents $1:5$ times the standard deviation. For our algorithm's hyperparameters, we picked

$$= \frac{1}{4dL}; u = 10 \quad ^2; r = 0:05; m = 1:$$

Note $m = 1$ corresponds to using a two-point estimator. For PAGD, we used the hyperparameters listed in their paper, and for ZO-GD-NCF, we used the code from their Neurips submission. For initialization, we chose a random x_0 near the saddle point at the origin, drawn from $N(0; 10 \quad ^3 I_{d \quad d})$

As we see in Fig. 2, our algorithm reaches the global minimum of the octopus function in significantly fewer function evaluations than PAGD and ZO-GD-NCF (approximately 2.5 times faster than ZO-GD-NCF, and approximately 3 times faster than PAGD), despite our algorithm only using 2 function evaluations per iteration compared to $2d$ function evaluations per iteration for both PAGD and ZO-GD-NCF. This suggests that in addition to our theoretical convergence guarantees, there can also be empirical benefits to using two-point estimators versus existing $2d$-point estimators in the zeroth-order escaping saddle point literature.

5. Conclusion

In this paper, we proved that using two function evaluations per iteration suffices to escape saddle points and reach approximate second order stationary points efficiently in zeroth-order optimization. Along the way, we also gave the first analysis of high-probability function change using two (or more)-point zeroth-order gradient estimators, as well as a novel concentration bound for sums of subExponential (but not subGaussian) vectors which are each the products of Gaussian vectors. These technical contributions may be

[3] Our code can be found at https://github.com/rafflesintown/escape-saddle-points-2pt

Figure 1: Performance on toy octopus function, with d = 30

of independent interest to researchers working in zeroth-order optimization as well as general stochastic optimization. Finally, we provided numerical evidence supporting the theoretical convergence results.

6. Acknowledgements

This work is supported by NSF CAREER: ECCS-1553407, NSF CNS: 2003111, NSF AI institute: 2112085, and ONR YIP: N00014-19-1-2217.

References

Adolphs, L., Daneshmand, H., Lucchi, A., and Hofmann, T. Local saddle point optimization: A curvature exploitation approach. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 486–495. PMLR, 2019.

Antonakopoulos, K., Mertikopoulos, P., Piliouras, G., and Wang, X. Adagrad avoids saddle points. In *International Conference on Machine Learning*, pp. 731–771. PMLR, 2022.

Avdiukhin, D. and Yaroslavtsev, G. Escaping saddle points with compressed sgd. *Advances in Neural Information Processing Systems*, 34:10273–10284, 2021.

Avdiukhin, D., Jin, C., and Yaroslavtsev, G. Escaping saddle points with inequality constraints via noisy sticky projected gradient descent. In *11th Annual Workshop on Optimization for Machine Learning*, 2019.

Bai, Q., Agarwal, M., and Aggarwal, V. Escaping saddle points for zeroth-order non-convex optimization using estimated gradient descent. In *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6. IEEE, 2020.

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

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. 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*, pp. 15–26, 2017.

Chen, Y., Bernstein, A., Devraj, A., and Meyn, S. Model-free primal-dual methods for network optimization with application to real-time optimal power flow. In *2020 American Control Conference (ACC)*, pp. 3140–3147. IEEE, 2020.

Criscitiello, C. and Boumal, N. Efficiently escaping saddle points on manifolds. *Advances in Neural Information Processing Systems*, 32, 2019.

Daneshmand, H., Kohler, J., Lucchi, A., and Hofmann, T. Escaping saddles with stochastic gradients. In *International Conference on Machine Learning*, pp. 1155–1164. PMLR, 2018.

Du, S. S., Jin, C., Lee, J. D., Jordan, M. I., Singh, A., and Poczos, B. Gradient descent can take exponential time to escape saddle points. In *Advances in neural information processing systems*, pp. 1067–1077, 2017.

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

Fang, C., Lin, Z., and Zhang, T. Sharp analysis for nonconvex sgd escaping from saddle points. In Conference on Learning Theory, pp. 1192–1234. PMLR, 2019.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. Online convex optimization in the bandit setting: gradient descent without a gradient. In Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 385–394, 2005.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points—online stochastic gradient for tensor decomposition. In Conference on Learning Theory, pp. 797–842, 2015.

Ge, R., Li, Z., Wang, W., and Wang, X. Stabilized svrg: Simple variance reduction for nonconvex optimization. In Conference on learning theory, pp. 1394–1448. PMLR, 2019.

Han, A. and Gao, J. Escape saddle points faster on manifolds via perturbed riemannian stochastic recursive gradient. arXiv preprint arXiv:2010.12191, 2020.

Huang, M. Escaping saddle points for nonsmooth weakly convex functions via perturbed proximal algorithms. arXiv preprint arXiv:2102.02837, 2021.

Huang, M., Ji, K., Ma, S., and Lai, L. Efficiently escaping saddle points in bilevel optimization. arXiv preprint arXiv:2202.03684, 2022.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. In International Conference on Machine Learning, pp. 1724–1732. PMLR, 2017.

Jin, C., Liu, L. T., Ge, R., and Jordan, M. I. On the local minima of the empirical risk. Advances in neural information processing systems, 31, 2018a.

Jin, C., Netrapalli, P., and Jordan, M. I. Accelerated gradient descent escapes saddle points faster than gradient descent. In Conference On Learning Theory, pp. 1042–1085. PMLR, 2018b.

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. arXiv preprint arXiv:1902.04811, 2019a.

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. A short note on concentration inequalities for random vectors with subgaussian norm. arXiv preprint arXiv:1902.03736, 2019b.

Larson, J., Menickelly, M., and Wild, S. M. Derivative-free optimization methods. Acta Numerica, 28:287–404, 2019.

Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. First-order methods almost always avoid strict saddle points. Mathematical programming, 176(1):311–337, 2019.

Li, J., Balasubramanian, K., and Ma, S. Stochastic zeroth-order riemannian derivative estimation and optimization. Mathematics of Operations Research, 2022.

Li, Y., Tang, Y., Zhang, R., and Li, N. Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach. IEEE Transactions on Automatic Control, 2021.

Li, Z. Ssrgd: Simple stochastic recursive gradient descent for escaping saddle points. Advances in Neural Information Processing Systems, 32, 2019.

Liang, G., Tong, Q., Zhu, C., and Bi, J. Escaping saddle points with stochastically controlled stochastic gradient methods. arXiv preprint arXiv:2103.04413, 2021.

Lucchi, A., Orvieto, A., and Solomou, A. On the second-order convergence properties of random search methods. Advances in Neural Information Processing Systems, 34: 25633–25645, 2021.

Malik, D., Pananjady, A., Bhatia, K., Khamaru, K., Bartlett, P., and Wainwright, M. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In The 22nd international conference on artificial intelligence and statistics, pp. 2916–2925. PMLR, 2019.

Mokhtari, A., Ozdaglar, A., and Jadbabaie, A. Escaping saddle points in constrained optimization. Advances in Neural Information Processing Systems, 31, 2018.

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

Panageas, I., Piliouras, G., and Wang, X. First-order methods almost always avoid saddle points: The case of vanishing step-sizes. Advances in Neural Information Processing Systems, 32, 2019.

Reddi, S., Zaheer, M., Sra, S., Poczos, B., Bach, F., Salakhutdinov, R., and Smola, A. A generic approach for escaping saddle points. In International conference on artificial intelligence and statistics, pp. 1233–1242. PMLR, 2018.

Roy, A., Balasubramanian, K., Ghadimi, S., and Mohapatra, P. Escaping saddle-point faster under interpolation-like conditions. Advances in Neural Information Processing Systems, 33:12414–12425, 2020.

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

Staib, M., Reddi, S., Kale, S., Kumar, S., and Sra, S. Escaping saddle points with adaptive gradient methods. In International Conference on Machine Learning, pp. 5956–5965. PMLR, 2019.

Sun, J., Qu, Q., and Wright, J. When are nonconvex problems not scary? arXiv preprint arXiv:1510.06096, 2015.

Sun, T., Li, D., Quan, Z., Jiang, H., Li, S., and Dou, Y. Heavy-ball algorithms always escape saddle points. arXiv preprint arXiv:1907.09697, 2019a.

Sun, Y., Flammarion, N., and Fazel, M. Escaping from saddle points on riemannian manifolds. Advances in Neural Information Processing Systems, 32, 2019b.

Tang, H., Lian, X., Qiu, S., Yuan, L., Zhang, C., Zhang, T., and Liu, J. Deepsqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression. arXiv preprint arXiv:1907.07346, 2019.

Tang, Y., Ren, Z., and Li, N. Zeroth-order feedback optimization for cooperative multi-agent systems. In 2020 59th IEEE Conference on Decision and Control (CDC), pp. 3649–3656. IEEE, 2020a.

Tang, Y., Zhang, J., and Li, N. Distributed zero-order algorithms for nonconvex multiagent optimization. IEEE Transactions on Control of Network Systems, 8(1):269–281, 2020b.

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

Vershynin, R. High-dimensional probability: An introduction with applications in data science, volume 47. Cambridge university press, 2018.

Vladimirova, M., Girard, S., Nguyen, H., and Arbel, J. Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions. Stat, 9(1):e318, 2020.

Vlaski, S. and Sayed, A. H. Distributed learning in nonconvex environments—part ii: Polynomial escape from saddle-points. IEEE Transactions on Signal Processing, 69:1257–1270, 2021a.

Vlaski, S. and Sayed, A. H. Second-order guarantees of stochastic gradient descent in non-convex optimization. IEEE Transactions on Automatic Control, 2021b.

Vlatakis-Gkaragkounis, E.-V., Flokas, L., and Piliouras, G. Efficiently avoiding saddle points with zero order methods: No gradients required. Advances in Neural Information Processing Systems, 32, 2019.

Wang, J.-K., Lin, C.-H., and Abernethy, J. Escaping saddle points faster with stochastic momentum. arXiv preprint arXiv:2106.02985, 2021.

Xu, Y., Jin, R., and Yang, T. First-order stochastic algorithms for escaping from saddle points in almost linear time. Advances in neural information processing systems, 31, 2018.

Zhang, C. and Li, T. Escape saddle points by a simple gradient-descent based algorithm. Advances in Neural Information Processing Systems, 34:8545–8556, 2021.

Zhang, H., Xiong, H., and Gu, B. Zeroth-order negative curvature finding: Escaping saddle points without gradients. arXiv preprint arXiv:2210.01496, 2022.

Ziyin, L., Li, B., Simon, J. B., and Ueda, M. Sgd with a constant large learning rate can converge to local maxima. arXiv preprint arXiv:2107.11774, 2021.

A. Related Work

Two-point methods in zeroth-order optimization. Two-point (or in general $2m$-point, where $1 \leq m < d$ with d being the problem dimension) estimators, which approximate the gradient using two (or $2m$) function evaluations per iteration, have been widely studied by researchers in the zeroth-order optimization literature, in convex (Nesterov & Spokoiny, 2017; Duchi et al., 2015; Shamir, 2017), nonconvex (Nesterov & Spokoiny, 2017), online (Shamir, 2017), as well as distributed settings (Tang et al., 2019). A key reason for doing so is that for applications of zeroth-order optimization arising in robotics (Li et al., 2022), wind farms (Tang et al., 2020a), power systems (Chen et al., 2020), online (time-varying) optimization (Shamir, 2017), learning-based control (Malik et al., 2019; Li et al., 2021), and improving adversarial robustness to black-box attacks in deep neural networks (Chen et al., 2017), it may be costly or impractical to wait for $\Theta(d)$ (where d denotes the problem dimension) function evaluations per iteration to make a step. This is especially true for high-dimensional and/or time-varying problems. Indeed, for high-dimensional problems, two-point estimators can make swift progress even in the initial stage compared to $2d$-point estimator, and can reach a higher-quality solution if computation is limited (Tang et al., 2020b; Chen et al., 2017). For instance, consider the work in (Chen et al., 2017), which studies the use of zeroth-order estimators to perform black-box attacks on deep neural networks, in order to identify (and then defend against) adversarial images that may lead to misclassification. In the paper, the authors employed two-point zeroth-order estimators, due to the high computational cost of using $2d$ function evaluations per iteration for hundreds of iterations (d is the dimension of an image, which in this case is over 20000). The authors showed empirically that their two-point estimators worked well; however there over no accompanying theoretical results.

For online or time-varying environments, two-points estimators also often preferable. Since zeroth-order methods are often used in physical systems whose environment drifts or changes over time, this leads naturally to a time-varying or online optimization. For these problems, $2d$-point estimators will not produce a good estimation because the underlying function can drift to a very different problem while waiting for the $2d$ function evaluations. Indeed, the fewer function evaluations an optimization procedure needs, the faster it can catch up with the time-varying environment. In fact, for online optimization, it has been shown that two points estimator is optimal for convex Lipschitz functions (Shamir, 2017). Thus, two-point estimators are a natural fit for time-varying online optimization problems.

Saddle point escape with access to deterministic gradient. While standard gradient descent can escape saddle points asymptotically (Lee et al., 2019; Panageas et al., 2019), it is known that standard gradient descent may take exponential time to escape saddle points (Du et al., 2017). Hence, when access to deterministic gradient is available, research has centered on escaping saddle points with adding perturbation (Jin et al., 2017), momentum/acceleration based methods (Jin et al., 2018b; Sun et al., 2019a; Staib et al., 2019), or gradient-based robust Hessian power/curvature exploitation methods (Zhang & Li, 2021; Adolphs et al., 2019). In addition, there has also been work on escaping saddle points devoted to specific optimization settings, such as constrained optimization (Mokhtari et al., 2018; Avdiukhin et al., 2019), optimization of weakly convex functions (Huang, 2021), bilevel optimization (Huang et al., 2022), as well as on general manifolds (Sun et al., 2019b; Criscitiello & Boumal, 2019; Han & Gao, 2020).

Saddle point escape in stochastic gradient descent (SGD). In practice, only stochastic gradient estimators are available in many problems. While SGD may converge to local maxima in worst-case scenarios (Ziyin et al., 2021), under assumptions such as bounded variance or subGaussian noise, there have been many works that have studied the problem of saddle point escape in SGD (Ge et al., 2015; Daneshmand et al., 2018; Xu et al., 2018; Jin et al., 2019a; Vlaski & Sayed, 2021b). The best existing rate (without considering momentum/variance reduction techniques) appears to belong to that of (Fang et al., 2019), which converges to ϵ-second order stationary points using $O(1/\epsilon^{3.5})$ stochastic gradients. While zeroth-order gradient estimators may also be viewed as stochastic gradients, they typically do not satisfy the bounded/subGaussian noise assumptions that are assumed in these works, making a direct comparison inappropriate. Escaping saddle point via momentum methods in SGD has also been studied (Wang et al., 2021; Antonakopoulos et al., 2022); while we do not consider incorporating momentum in our works, this may be interesting future work. A number of papers has also considered the specialized setting of escaping saddle points in nonconvex finite-sum optimization (Reddi et al., 2018; Liang et al., 2021), with many considering the case where variance-reduction is used (Ge et al., 2019; Li, 2019). While the finite-sum problem is quite different from our problem, the variance reduction approach considered in these works may be a relevant future direction. The saddle point escape problem has also been studied in other specific settings such as compressed optimization (Avdiukhin & Yaroslavtsev, 2021), distributed optimization (Vlaski & Sayed, 2021a), or in the overparameterization case (Roy et al., 2020).

Saddle point escape with zeroth-order information. The problem of escaping saddle points in zeroth-order optimization has been studied less often, and we have already listed all known works comparable to our work in the introduction (Bai et al., 2020; Vlatakis-Gkaragkounis et al., 2019; Balasubramanian & Ghadimi, 2022); a more detailed comparison of these works with our results has been provided in the discussion following the statement of our main result Theorem 1. We would like to mention that (Roy et al., 2020) also includes a convergence result of $\tilde{O}\left(\frac{d^{1.5}}{\epsilon^{4.5}}\right)$ for the case with noisy function evaluations, which is incomparable to our existing work which focuses on the case with exact function evaluation. In addition, (Roy et al., 2020) also makes a subGaussian assumption on the estimator noise, which zeroth-order estimators in our paper do not satisfy. Nonetheless, considering the extension to noisy function evaluations will make for important future work.

Zeroth-order optimization. Our work rests on a line of research in zeroth-order optimization which focuses on constructing gradient estimators using zeroth-order function values (Flaxman et al., 2005; Duchi et al., 2015; Nesterov & Spokoiny, 2017; Shamir, 2017; Larson et al., 2019). As we have discussed, for smooth nonconvex functions, it is known that two-point zeroth-order estimators suffice to find first-order ϵ-stationary points using $O(d/\epsilon^2)$ function evaluations (Nesterov & Spokoiny, 2017). Our work studies the more complicated problem of reaching ϵ-second order stationary points, attaining a rate of $\tilde{O}(d/\epsilon^{2.5})$.

B. Proof Roadmap

We begin by introducing several key concentration inequalities in Appendix C which we will frequently use in our proofs. We then describe in detail (and prove) the sequence of results that lead up to Proposition 4 in Appendix D, showing that there cannot be too many iterations with large gradients. Next, we describe the saddle point argument in detail, and prove Proposition 5 in Appendix E.3. Finally, we combine these results and prove our main result Theorem 2 (whose informal version is Theorem 1) in Appendix F

Throughout our proofs, absolute constants, as denoted by (e.g. c, c', C), may change from line to line. However, within the same proof, for clarity, we try to index different constants differently. We assume $d \ge 2$ and $m \le d$.

Notations. We shall denote the conditional expectation and conditional probability by $\mathbb{E}_{\mathcal{F}}[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}]$ and $\mathbb{P}_{\mathcal{F}}(\cdot) = \mathbb{P}(\cdot \mid \mathcal{F})$ where $\mathcal{F}$ is a sigma-algebra.

C. Concentration inequalities

This section serves to introduce several probabilistic results which will be useful for our main proofs in subsequent sections. We first introduce subGaussian, subExponential and norm-subGaussian random vectors in Appendix C.1. Next, in Appendix C.2, we provide concentration bounds for norm-subGaussian and subExponential random vectors. We then prove a novel concentration inequality involving products of subGaussian random vectors in Appendix C.3. We conclude by stating some concentration bounds for Appendix C.4 random variables.

C.1. subGaussian, subExponential and norm-subGaussian random vectors

We first define subGaussian and subExponential random vectors. A detailed reference for these concepts can be found in (Vershynin, 2018).

Definition 5 (subGaussian and subExponential random vectors). A random vector $x \in \mathbb{R}^d$ is σ-subGaussian (SG(σ)), if there exists $\sigma > 0$ such that for any unit vector $g \in \mathbb{S}^{d-1}$,

$$\mathbb{E}\left[\exp(\lambda \langle g, x - \mathbb{E}[x] \rangle)\right] \le \exp(\lambda^2 \sigma^2 / 2) \quad \forall \lambda \in \mathbb{R}.$$

Meanwhile, a random vector $x \in \mathbb{R}^d$ is σ-subExponential (SE(σ)), if there exists $\sigma > 0$ such that for any unit vector $g \in \mathbb{S}^{d-1}$,

$$\mathbb{E}\left[\exp(\lambda \langle g, x - \mathbb{E}[x] \rangle)\right] \le \exp(\lambda^2 \sigma^2 / 2) \quad \forall |\lambda| \le \frac{1}{\sigma}$$

An alternative concentration property for random vectors revolving around its norm, known as norm-subGaussianity (Jin et al., 2019b), is also relevant.

Definition 6 (norm-subGaussian random vectors). A random vector $x\ 2\ R^d$ is -norm-subGaussian (nSG()), there exists > 0 such that

$$P(kx \quad Ex\,k \quad s) \quad 2e^{\frac{s^2}{2\ 2}} \quad 8s \quad 0:$$

We recall the following result which provides several examples of nSG random vectors. In particular, it tells us a random vector $x\ 2\ R^d$ that is($= \stackrel{p}{\ } \overline{d}$) subGaussian is also -subGaussian.

Lemma 4 (Lemma 1 in (Jin et al., 2019b)) There exists absolute constant c such that the following random vectors are all nSG(c).

1. A bounded random vector $x\ 2\ R^d$ so that $kx\,k$.
2. A random vector $x\ 2\ R^d$, where $x = \ e_1$ and the random variable $2\ R$ is -subGaussian.
3. A random vector $x\ 2\ R^d$ that is($= \stackrel{p}{\ } \overline{d}$) subGaussian

In addition, if $x\ 2\ R^d$ is zero-mean nSG(), its component along a single direction is also subGaussian.

Lemma 5. Suppose $x\ 2\ R^d$ is zero-mean nSG(). Then, for any fixed vector $v\ 2\ R^d$, $hv\,;x\,i$ is zero-mean $kv\,k$ -subGaussian.

Proof. Without loss of generality, we assume that $v\ 2\ S^{d\ \ 1}$ is a unit vector. That $hv\,;x\,i$ is zero-mean follows directly from x being zero-mean and v being fixed. Meanwhile, since $jhv\,;x\,ij \quad k \quad vkkx\,k = \ kx\,k$, for any $s \quad 0$, it follows that

$$P(jhv\,;x\,ij \quad s) \quad P(kx\,k \quad s) \quad 2e^{\frac{s^2}{2\ 2}};$$

where the last inequality follows from the fact that x is zero-mean and also nSG( ). Hence $hv\,;x\,i$ is zero-mean SG(), as desired. □

C.2. Concentration bounds for norm-subGaussian and subExponential random vectors

We begin by giving some concentration bounds for norm-subGaussian random vectors. To do so, we introduce the following condition.

Condition 1. Consider random vectors $x_1;::::;x_n\ 2\ R^d$, and corresponding filtrations F_i generated by $(x_1;::::;x_i)$. We assume $x_i\ j\,F_{i\ \ 1}$ is zero-mean, nSG$(_i)$, with $_i\ 2\,F_{i\ \ 1}$, i.e,

$$E\,[x_i\ j\,F_{i\ \ 1}] = 0\,;$$

and

$$P\,(kx_i\,k \quad s\,j\,F_{i\ \ 1}) \quad 2e^{\frac{s^2}{2\ _i^2}} \quad 8s \quad 0;$$

where $_i$ is a measurable function of $(x_1;::::;x_{i\ \ 1})$ for each i.

For norm subGaussian random vectors satisfying Condition 1, we first have the following bound.

Lemma 6. Suppose $(x_1;::::;x_n)\ 2\ R^d$ satisfy Condition 1, i.e. each $x_i\ j\,F_{i\ \ 1}$ is mean-zero, nSG$(_i)$ with $_i\ 2\,F_{i\ \ 1}$. Let $f\,u_i\,g$ denote a sequence of random vectors such that $u_i\ 2\,F_{i\ \ 1}$ for every $i\ 2\ [n]$. Then, there exists an absolute constant c, such that for any $2\ (0;1)$ and $> \ 0$, with probability at least 1 ,

$$\overset{n}{X}_{i=1} hu_i\,;x_i\,i \quad c \quad \overset{n}{X}_{i=1} ku_i\,k^2 \ \ _i^2 + \frac{1}{\ } \log(1= \ ):$$

Proof. We note that if x_i is mean-zero and nSG$(_i)$, then by Lemma 5 $hu_i\,;x_i\,i\ j\ F_{i\ \ 1}$ is zero-mean and $ku_i\,k\ _i$-subGaussian. The rest of the proof follows from the proof of Lemma 39 in (Jin et al., 2019a) (key idea is exponentiate and then apply Markov's inequality). For completeness, we restate the proof here. Observe that, for any i, since $h u_i ; x_i i$ is $k u_i k \ _i$-subGaussian, for any > 0, we have that

$$E [\exp(\ h u_i ; x_i i) \ j \ F_{i \ 1}] \quad \exp(\ ^2 k u_i k^2 \ _i^2 = 2)$$

For any > 0 and 0, observe that

$$
\begin{aligned}
& P \left(\sum_{i=1}^{n} h u_i ; x_i i \quad {}^2 k u_i k^2 \ {}_i^2 = 2 \quad s \right) \\
= \ & P \left(\exp \left(\sum_{i=1}^{n} h u_i ; x_i i \quad {}^2 k u_i k^2 \ {}_i^2 = 2 \right) \quad \exp(s) \right) \\
& E \left[\exp \left(\sum_{i=1}^{n} h u_i ; x_i i \quad {}^2 k u_i k^2 \ {}_i^2 = 2 \right) \right] \exp(\ s) \\
= \ & E \left[E \left[\exp \left(\sum_{i=1}^{n} h u_i ; x_i i \quad {}^2 k u_i k^2 \ {}_i^2 = 2 \right) \ F_{n \ 1} \right] \right] \exp(\ s) \\
= \ & E \left[\exp \left(\sum_{i=1}^{n \ 1} h u_i ; x_i i \quad {}^2 k u_i k^2 \ {}_i^2 = 2 \right) E \left[\exp \ h u_n ; x_n i \quad {}^2 k u_n k^2 \ {}_n^2 = 2 \ \ F_{n \ 1} \right] \right] \exp(\ s) \\
\overset{(i)}{} \ & E \left[\exp \left(\sum_{i=1}^{n \ 1} h u_i ; x_i i \quad {}^2 k u_i k^2 \ {}_i^2 = 2 \right) \right] \exp(\ s) \qquad \exp(\ s)
\end{aligned}
$$

Above, (i) follows from the fact that $h u_i ; x_i i \ j \ F_{i \ 1}$ is zero-mean and $k u_i k \ _i$-subGaussian for each $i \ 2 \ [n]$. The final result then follows by picking $c = \frac{1}{2}$ and $s = \log(1 = \ )$. □

Assuming Condition 1, the following concentration result also holds for a sequence of nSG random vectors.

Lemma 7 (Lemma 6, Corollary 7 and Corollary 8 in (Jin et al., 2019b) combined) Suppose $(x_1; : : : ; x_n) \ 2 \ R^d$ satisfy Condition 1. Then, there exists an absolute constant c such that for any fixed $2 \ (0; 1)$, > 0, with probability at least 1 ,

$$\sum_{i=1}^{n} x_i \qquad c \quad \sum_{i=1}^{n} \ {}_i^2 + \frac{1}{} \log(2d = \ ) :$$

Moreover, there are two corollaries.

1. (Corollary 7 in (Jin et al., 2019b)) When $f \ _i g$ is deterministic, there exists an absolute constant c such that for any fixed $2 \ (0; 1)$, with probability at least 1 .

$$\sum_{i=1}^{n} x_i \qquad c \sqrt{\log(2d = \ ) \sum_{i=1}^{n} \ {}_i^2}$$

2. (Corollary 8 in (Jin et al., 2019b)) Suppose that the $f \ _i g$ sequence is random. Then, there exists an absolute constant c such that for any fixed $2 \ (0; 1)$ and $B > b > 0$, with probability at least 1 :

$$\text{either} \quad \sum_{i=1}^{n} \ {}_i^2 \quad B \quad \text{or} \quad \sum_{i=1}^{n} x_i \qquad c \sqrt{\max \left(\sum_{i=1}^{n} \ {}_i^2 ; b \right) (\log(2d = \ ) + \log(\log(B = b)))}$$

We state here a Bernstein-type concentration inequality for sub-exponential random variables, which we also need.

Lemma 8 (Bernstein concentration inequality)**.** Consider a sequence of independently distributed σ-subexponential variables $x_1, \dots, x_n \in \mathbb{R}$, with mean $E[x_i] \le c'$ for some $c' > 0$ and each $i \in [n]$. Then, there exists an absolute constant $C > 0$, such that for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\sum_{i=1}^{n} x_i \le C\sigma(n + \log(1/\delta)). \tag{8}$$

Proof. The result of Eq. (8) follows by applying Bernstein's inequality to $\sum_{i=1}^{n} x_i - E[x_i]$ (so each summand is mean-zero). Per Bernstein's inequality, (cf. Theorem 2.8.1 in (Vershynin, 2018)), there exists an absolute constant $c > 0$ such that for any $s \ge 0$,

$$P\left(\sum_{i=1}^{n} (x_i - E[x_i]) \ge s\right) \le \exp\left(-c \min\left(\frac{s^2}{n\sigma^2}, \frac{s}{\sigma}\right)\right).$$

Pick $s = \sigma\left(n + \frac{\log(1/\delta)}{c}\right)$. Then,

$$\min\left(\frac{s^2}{n\sigma^2}, \frac{s}{\sigma}\right) = \min\left(n + 2\frac{\log(1/\delta)}{c} + \frac{(\log(1/\delta))^2}{c^2 n}; n + \frac{\log(1/\delta)}{c}\right) = n + \frac{\log(1/\delta)}{c}.$$

Continuing, we have that

$$P\left(\sum_{i=1}^{n} (x_i - E[x_i]) \ge s\right) \le \exp\left(-c \min\left(\frac{s^2}{n\sigma^2}, \frac{s}{\sigma}\right)\right) \le \exp\left(-c\left(n + \frac{\log(1/\delta)}{c}\right)\right) \le \delta.$$

Thus, it follows that with probability at least $1 - \delta$,

$$\sum_{i=1}^{n} (x_i - E[x_i]) \le \sigma\left(n + \frac{\log(1/\delta)}{c}\right) \implies \sum_{i=1}^{n} x_i \le \sigma\left(n + \frac{\log(1/\delta)}{c}\right) + nc';$$

where implication holds since by assumption $E[x_i] \le c'$ for some $c' > 0$. Then, by setting $C = \max\{1 + c', 1/c\}$, the desired result follows. □

C.3. A novel concentration inequality for the zeroth-order setting

In the zeroth-order setting, we will frequently have to bound the norms of terms of the form

$$W = \sum_{t=0}^{\tau-1} M_t (Z_t Z_t^{\top} - I) v_t; \tag{9}$$

where M_t is a known and fixed quantity, while Z_t is random, and v_t depends on x_0 and the history of previous $\{Z_j\}_{j=0}^{t-1}$'s, and is hence $\mathcal{F}_{t-1}$-measurable. For our purposes, it suffices to consider $Z \sim N(0, I)$.

To see why such a bound will be useful, as mentioned in the main text and as we will see again later in the full proofs, in the analysis of escaping saddle points, we need to bound a term of the form

$$W_{g_0}(\tau) = \sum_{t=0}^{\tau-1} (I - \eta H)^{\tau-1-t} (Z_t Z_t^{\top} - I)(\nabla f(x_t) - \nabla f(x_t'));$$

where $H = \nabla^2 f(x_0)$ (assuming that x_0 is an ϵ-saddle point), and x_t and x_t' are two coupled sequences. Comparing with Eq. (9), we see that for the equation above, we can define $M_t = (I - \eta H)^{\tau-1-t}$ (a fixed and known quantity) and $v_t = \nabla f(x_t) - \nabla f(x_t')$ (clearly, $\nabla f(x_t) - \nabla f(x_t')$ is $\mathcal{F}_{t-1}$-measurable). This motivates why we wish to bound terms of the form Eq. (9).

Observe that each $(Z_t Z_t^{\top} - I) v_t \mid \mathcal{F}_{t-1}$ term is subExponential rather than subGaussian. While it is possible to define norm-subExponential vectors in analogous way to norm-subGaussian vectors, the corresponding moment generating function

(MGF) for subExponential random variables is not defined on the entirety of R. When bounding a sum in the form of $\sum_{t=0}^{\ \ 1}(Z_tZ_t^{>} \quad I)v_t$, this creates a subtle but challenging technical issue.

Following the intuition outlined in the main text, we bypass this difficulty by proving the following result. For notational simplicity, we introduce the function

$$\mathrm{lr}(x) := \log(x\log(x)) : \qquad (10)$$

We now recall Proposition 2 which we first introduced in the main text.

Proposition 2. Let F_t; $t \quad 1$ be a filtration. Let $(Z_t)_{t \quad 0}$ be a sequence of random vectors following the distribution $N(0; I)$ such that $Z_t \ 2\ F_t$ and is independent of $F_{t \quad 1}$, and let $(v_t)_{t \quad 0}$ be a sequence of random vectors such that $v_t \ 2\ F_{t \quad 1}$. For each $\quad 0$, let

$$W \quad = \sum_{t=0}^{\ \ 1} M_t(Z_tZ_t^{>} \quad I)v_t\,;$$

where each M_t is a deterministic matrix of appropriate dimension. Then, there exist some absolute constants $c^0; C > 0$ such that for any $\ 2\ Z^+$ and $\ 2\ (0; 1{=}e]$, the following statements hold:

1. For any $\ > \ 0$, with probability at least $1 \quad$, we have

$$kW \ k \quad c^0 \sum_{t=0}^{\ \ 1} kM_tk_2^2 d(\mathrm{lr}(\ C \ = \ ))^2 kv_tk^2 + \frac{1}{\ } \log(Cd \ = \ ):$$

2. For any $B > b > \ 0$, with probability at least $1 \quad$,

$$\text{either} \quad \sum_{t=0}^{\ \ 1} kM_tk_2^2 d(\mathrm{lr}(\ C \ = \ ))^2 kv_tk^2 \quad B; \text{ or}$$

$$kW \ k \quad \sqrt{\max\left(\sum_{t=0}^{\ \ 1} kM_tk_2^2 d(\mathrm{lr}(\ C \ = \ ))^2 kv_tk^2; b\right)} \; c^0\sqrt{(\log(C \ d{=} \ ) + \log(\log(\ B{=}b) + 1))}$$

Moreover, as is clear from the bounds above, we may pick $C \ 1$ such that $\log \frac{C}{\ } \quad 1; 8 \quad 2\ (0; \frac{1}{e}]$.

Proof. We will focus on proving the first point, since the second follows as a natural corollary of our proof of the first part and the proof of Corollary 8 in (Jin et al., 2019b). For simplicity, we shall assume $v_t \ 6{=}\ 0$ in the intermediate steps; extension to the general case is straightforward.

First of all, for $0 \quad < \ 1$, let

$$g(\ ;\ ) = \sqrt{\frac{2}{\ }} \int^{\sqrt{2\,\mathrm{lr}(1= \ )}} (x^2 \quad 1)e^{\ \ x^2=2}\,dx = \sqrt{\frac{2}{\ }} \quad e^{\ \ ^2=2} \left(\frac{\sqrt{2\,\mathrm{lr}(1= \ )}}{\log(1= \ )} \right) :$$

It's not hard to see that for a fixed $\ 2\ (0; 1{=}e]$, $g(\ ;\ )$ is continuous and strictly increasing over $\ 2\ [0; 1)$. Then, since $\frac{\log x}{x} + 1 \quad x$ for $x \quad 1$, by plugging in $x = \log(1= \ )$, we get

$$\frac{\mathrm{lr}(1= \ )}{(\log(1= \ ))^2} = \frac{\log\log(1= \ ) + \log(1= \ )}{(\log(1= \ ))^2} = \frac{1}{\log(1= \ )} \quad \frac{\log\log(1= \ )}{\log(1= \ )} + 1 \qquad 1;$$

which leads to

$$g(2 \ ;\ ) = \sqrt{\frac{2}{\ }} \quad 2\ e^{\ \ 2 \ ^2} \left(\frac{\sqrt{2\,\mathrm{lr}(1= \ )}}{\log(1= \ )} \right) \quad \sqrt{\frac{2}{\ }} \quad 2e^{\ \ 2=e^2} \quad \sqrt{\frac{\ }{2}} \quad > \ 0$$

for $\ 2\ (0;1=e]$. Furthermore, we obviously have $g(0;\ ) < 0$. Therefore $g(\ ;\ ) = 0$ has a unique solution in $(0;2\ )$, which we denote by $(\ )$.[4] These results imply that, for a random variable Z following the standard normal distribution, we have

$$\text{E}\ \overset{h}{(Z^2\quad 1)1_{(\ )\ j\ Z\ j}}\ \sqrt{2\,\text{lr}(1=\ )}\ \overset{i}{=}\ \overset{r}{\frac{2}{\ }}\int_{(\ )}^{\sqrt{2\,\text{lr}(1=\ )}}(x^2\quad 1)e^{\ x^2=2}\,dx = g(h(\ );\ ) = 0$$

and

$$\begin{aligned}\text{P}(\ (\ )\ \ j\ Z\ j\ \ \sqrt{2\,\text{lr}(1=\ )})\ \ &1\ \ 2\ \ \frac{1}{\sqrt{2}}\int_{\sqrt{2\,\text{lr}(1=\ )}}^{1} e^{\ x^2=2}\,dx + \frac{1}{\sqrt{2}}\int_{0}^{(\ )} e^{\ x^2=2}\,dx \ !\\ &1\ \ 2\ \ \frac{1}{2}\exp\ \ \frac{2\,\text{lr}(1=\ )}{2}\ +\ \frac{(\ )}{\sqrt{2}}\ = 1\ \ 2\ \ \frac{\ }{2\log(1=\ )} + \frac{(\ )}{\sqrt{2}}\\ &1\ \ 2\ \ \frac{\ }{2} + \frac{2}{\sqrt{2}}\ \ \ \ 1\ \ C\end{aligned}$$

for any $\ 2\ (0;1=e]$, where we define the absolute constant $C = 2(1=2 + 2=\sqrt{2}\ )$.

Now we let A_t denote the event

$$A_t = \left(\ (\ )\ \ \frac{Z_t^> v_t}{kv_t k}\ \ \sqrt{2\,\text{lr}(1=\ )}\ \right):$$

Since $Z_t^> v_t = kv_t k$ conditioned on $F_{t\ 1}$ follows the standard normal distribution, we have

$$\text{P}_{F_{t\ 1}}(A_t)\quad 1\quad C\ ; \tag{11}$$

and

$$\text{E}_{F_{t\ 1}}\ v_t^>\ \ Z_t Z_t^>\quad I\ \ v_t 1_{A_t}\ \ = 0:$$

Moreover, for any random vector $u\ 2\ F_{t\ 1}$ that is orthogonal to v_t, we have

$$\text{E}_{F_{t\ 1}}\ u^>\ \ Z_t Z_t^>\quad I\ \ v_t 1_{A_t}\ \ = \text{E}_{F_{t\ 1}}\ u^> Z_t\ \ \ \text{E}_{F_{t\ 1}}\ Z_t^> v_t 1_{A_t}\ \ = 0\ ;$$

where we used the fact that $Z_t^> u$ is independent of $Z_t^> v_t$ conditioned on $F_{t\ 1}$. Therefore

$$\text{E}_{F_{t\ 1}}\ (Z_t Z_t^>\quad I\ )v_t 1_{A_t}\ \ = 0:$$

Consider defining then the random variable Q_t by

$$Q_t \coloneqq (\ Z_t Z_t^>\quad I\ )v_t\ \ 1_{A_t}:$$

We now show that $Q_t\ j\ F_{t\ 1}$ is norm-subGaussian. Let $u\ 2\ \mathbb{R}^d$ with $kuk = 1$ be arbitrary. We have

$$\begin{aligned}u^> Q_t &= u^> (Z_t Z_t^>\quad I\ )v_t\ \ 1_{A_t}\\ &= u^>\ \ \frac{v_t v_t^>}{kv_t k^2} + I\ \ \ \frac{v_t v_t^>}{kv_t k^2}\ \ (Z_t Z_t^>\quad I\ )v_t\ \ 1_{A_t}\\ &= u^> v_t\ \ \frac{jZ_t^> v_t j^2}{kv_t k^2}\ \ \ 1\ \ \ 1_{A_t} + u^>\ \ I\ \ \ \frac{v_t v_t^>}{kv_t k^2}\ \ (Z_t Z_t^>\quad I\ )v_t\ \ 1_{A_t}\\ &= u^> v_t\ \ \frac{jZ_t^> v_t j^2}{kv_t k^2}\ \ \ 1\ \ \ 1_{A_t} + u_?^> Z_t Z_t^> v_t\ \ 1_{A_t};\end{aligned}$$

[4] By letting $W_0(x)$ denote the the principal branch of the Lambert W function, it can be shown that

$$(\ ) = \sqrt{\ \ W_0\ \ \frac{2\ ^2\,\text{lr}(1=\ )}{(\log(1=\ ))^2}}\ :$$

where we denote $u_\perp = \left(I \quad \frac{v_t v_t^\top}{\|v_t\|^2}\right) u$. Since

$$u^\top v_t \left(\frac{|Z_t^\top v_t|^2}{\|v_t\|^2} \quad 1\right) 1_{A_t} \quad |u^\top v_t|(2\,\mathrm{lr}(1/\ ) \quad 1);$$

we see that $u^\top v_t \left(\frac{|Z_t^\top v_t|^2}{\|v_t\|^2} \quad 1\right) 1_{A_t}$ conditioned on $\mathcal{F}_{t-1}$ is $|u^\top v_t|(2\,\mathrm{lr}(1/\ ) \quad 1)$-subGaussian. Furthermore, since $|u_\perp^\top Z_t Z_t^\top v_t \quad 1_{A_t}| \quad |Z_t^\top u_\perp| \sqrt{2\,\mathrm{lr}(1/\ )}\|v_t\|$, we have

$$\mathbb{P}_{\mathcal{F}_{t-1}}\left(|u_\perp^\top Z_t Z_t^\top v_t \quad 1_{A_t}| \quad s\right) \quad \mathbb{P}_{\mathcal{F}_{t-1}}\left(|Z_t^\top u_\perp| \sqrt{2\,\mathrm{lr}(1/\ )}\|v_t\| \quad s\right);$$

and since $Z_t u_\perp / \|u_\perp\| \mid \mathcal{F}_{t-1}$ follows the standard normal distribution, we see that $u_\perp^\top Z_t Z_t^\top v_t \quad 1_{A_t}$ is a $\sqrt{2\,\mathrm{lr}(1/\ )}\|u_\perp\|\|v_t\|$-subGaussian variable. Note that $u^\top Q_t$ is just the sum of $u^\top v_t \left(\frac{|Z_t^\top v_t|^2}{\|v_t\|^2} \quad 1\right) 1_{A_t}$ and $u_\perp^\top Z_t Z_t^\top v_t \quad 1_{A_t}$, we can conclude that $u^\top Q_t$ is subGaussian with parameter

$$\begin{aligned}
&(2\,\mathrm{lr}(1/\ ) \quad 1)|u^\top v_t| + \sqrt{2\,\mathrm{lr}(1/\ )}\|u_\perp\|\|v_t\| \\
&2\,\mathrm{lr}(1/\ )(|u^\top v_t| + \|u_\perp\|\|v_t\|) \quad 2\sqrt{2}\,\mathrm{lr}(1/\ )\sqrt{|u^\top v_t|^2 + \|u_\perp\|^2\|v_t\|^2} \\
= {}& 2\sqrt{2}\,\mathrm{lr}(1/\ )\|v_t\|;
\end{aligned}$$

whenever $\ \in (0, 1/e]$. Consequently, by Lemma 1 in (Jin et al., 2019b), we see that $Q_t \mid \mathcal{F}_{t-1}$ is $8\,\mathrm{lr}(1/\ )\sqrt{d}\|v_t\|$-norm-subGaussian.

It follows easily that $M_t Q_t \mid \mathcal{F}_{t-1}$ is mean-zero and $8\,\mathrm{lr}(1/\ )\|M_t\|_2\|v_t\|\sqrt{d}$-norm-subGaussian. Hence, by Lemma 6 in (Jin et al., 2019a), we know that there exists an absolute constant $c > 0$ such that for any $\ > 0$ and $\ > 0$, we have that with probability at least $1 \quad$,

$$\sum_{t=0}^{\ \ 1} M_t Q_t \quad c\sum_{t=0}^{\ \ 1} d(\mathrm{lr}(1/\ ))^2\|M_t\|_2^2\|v_t\|^2 + \frac{1}{\ }\log(2d/\ ):$$

Now, consider denoting the event

$$A := \bigcup_{t=0}^{\ \ 1} A_t = \left\{ Z_t^\top v_t \in (\ )\|v_t\|, \sqrt{2\,\mathrm{lr}(1/\ )}\|v_t\| \ ; \forall t \in \{0, \ldots, \quad 1\} \right\}$$

By the union bound and Eq. (11), we note that

$$\mathbb{P}(A) \quad 1 \quad C \ :$$

Moreover, note that on the event A, $\sum_{t=0}^{\ \ 1} M_t Q_t = \sum_{t=0}^{\ \ 1} M_t(Z_t Z_t^\top \quad I)v_t$. Hence,

$$\begin{aligned}
&\mathbb{P}\left(\sum_{t=0}^{\ \ 1} M_t(Z_t Z_t^\top \quad I)v_t \quad c\sum_{t=0}^{\ \ 1} d(\mathrm{lr}(1/\ ))^2\|M_t\|_2^2\|v_t\|^2 + \frac{1}{\ }\log(2d/\ )\right) \\
&\quad \mathbb{P}\left(\sum_{t=0}^{\ \ 1} M_t Y_t \quad c\sum_{t=0}^{\ \ 1} d(\mathrm{lr}(1/\ ))^2\|M_t\|_2^2\|v_t\|^2 + \frac{1}{\ }\log(2d/\ ); \text{ and } A \text{ happens}\right) \\
&\quad 1 \quad \mathbb{P}\left(\sum_{t=0}^{\ \ 1} M_t Y_t \quad c\sum_{t=0}^{\ \ 1} d(\mathrm{lr}(1/\ ))^2\|M_t\|_2^2\|v_t\|^2 + \frac{1}{\ }\log(2d/\ )\right) + \mathbb{P}(A^c)\right) \\
&\quad 1 \quad (\ + C\ ):
\end{aligned}$$

Now, by rescaling $\ $ to $\ /(C \quad + 1)$, we get the desired result. Note this C is different from the C in the statement of the lemma by an absolute multiplicative factor. □

C.4. sub-Weibull random variables

In our work, we occasionally require bounding sums of heavy-tailed distribution, e.g. higher powers of $\|Z\|$ where $Z \sim N(0, I)$. To this end, we consider the following definition of sub-Weibull random variables.

Definition 7. We say that a random variable $X \in \mathbb{R}$ is sub-Weibull(K, θ) for some $K, \theta > 0$,

$$P(|X| \geq s) \leq 2\exp(-(s/K)^{1/\theta}) \quad \forall s \geq 0.$$

For instance, the standard normal distribution is sub-Weibull$(1, \frac{1}{2})$. From the way we define the tail parameter θ, the larger the θ, the heavier the tail of the distribution.

In our work, we need to show that the sum of sub-Weibull random variables is again sub-Weibull, which is ensured by the following result

Lemma 9. Suppose X and Y are sub-Weibull(K_X, θ) and sub-Weibull(K_Y, θ) respectively. Then, XY is sub-Weibull$(C(K_X K_Y), 2\theta)$ and $X + Y$ is sub-Weibull$(C(K_X + K_Y), \theta)$ for some absolute constant $C > 0$.

A helpful result is the following, which bounds the sum of identically distributed sub-Weibull random variables.

Lemma 10 (Corollary 3.1 in (Vladimirova et al., 2020)). Suppose $X_1, \ldots, X_n$ are identically distributed (K', θ) sub-Weibull random variables. Then, for some absolute constant $c > 0$, for all $s \geq ncK'$, we have

$$P\left(\sum_{i=1}^{n} X_i \geq s\right) \leq \exp\left(-\left(\frac{s}{ncK'}\right)^{1/\theta}\right)$$

In our work, we frequently need to bound sums of the k-th power of the norm of a standard d-dimensional Gaussian. We do so using Lemma 10.

Lemma 11. Suppose $X_i \overset{i.i.d}{\sim} N(0, I_d)$ for $i \in [n]$. Then, for any $k \in \mathbb{Z}^+$, there exists absolute constants $c, C > 0$ such that for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\sum_{i=1}^{n} \|X_i\|^{2k} \leq nCc^k d^k (1 + (\log(1/\delta))^k).$$

In particular, for any $\delta \in (0, 1/e)$ such that $\log(1/\delta) \geq 1$, it follows that

$$\sum_{i=1}^{n} \|X_i\|^{2k} \leq 2nCc^k d^k (\log(1/\delta))^k.$$

Proof. First, observe that for any $j \in [d]$, $(X_i)_j^2$, being subExponential, is $(1, 1)$-subWeibull. Then, by Lemma 9, $\|X_i\|^2 = \sum_{j=1}^{d} (X_i)_j^2$ is $(cd, 1)$ for some absolute constant c. Now, it follows from definition of sub-Weibullness in Definition 7 that $\|X_i\|^{2k}$ is $(c^k d^k, k)$-subWeibull. Hence, applying Lemma 10, we have that there exists absolute constant $C > 0$ such that for any $s \geq nCc^k d^k$,

$$P\left(\sum_{i=1}^{n} \|X_i\|^{2k} \geq s\right) \leq \exp\left(-\left(\frac{s}{nCc^k d^k}\right)^{1/k}\right)$$

Choosing $s = (1 + (\log(1/\delta))^k) nCc^k d^k$, we arrive then at the desired result. □

C.5. Supermartingale concentration inequalities

We first state and prove a supermartingale-type concentration inequality of the form we later require.

Lemma 12. Consider a filtration of sigma-algebras $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_{n-1} \subseteq \mathcal{F}_n$ and a sequence of random variables $X_1, \ldots, X_n$ such that $X_i \in \mathcal{F}_i$. Suppose that

$$P_{\mathcal{F}_{i-1}}(X_i \leq a) = 1 \quad \text{and} \quad P_{\mathcal{F}_{i-1}}(X_i \geq b) \leq p \tag{12}$$

for some $a, b > 0$ and $0 < p \leq \frac{1}{2}$. Then, for any $0 < \mu \leq b$ such that $|-b+\mu| \geq \frac{1-p}{p}(a+\mu)$, we have

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq -n\mu + s\right) \leq \exp\left(-\frac{s^2}{4n(b-\mu)^2}\right), \quad \forall s > 0.$$

Proof. Observe that by Markov's inequality, for any $\lambda > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq -n\mu + s\right) = \mathbb{P}\left(\exp\left(\lambda\sum_{i=1}^{n}(X_i+\mu)\right) \geq \exp(\lambda s)\right) \leq \frac{\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n}(X_i+\mu)\right)\right]}{\exp(\lambda s)}.$$

Now, observe that

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n}(X_i+\mu)\right)\right] &= \mathbb{E}\left[\mathbb{E}_{\mathcal{F}_{n-1}}\left[\exp\left(\lambda\sum_{i=1}^{n}(X_i+\mu)\right)\right]\right]\\ &= \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n-1}(X_i+\mu)\right)\mathbb{E}_{\mathcal{F}_{n-1}}[\exp(\lambda(X_n+\mu))]\right]. \end{aligned} \tag{13}$$

Let us now compute $\mathbb{E}_{\mathcal{F}_{n-1}}[\exp(\lambda(X_n+\mu))]$:

$$\begin{aligned}&\mathbb{E}_{\mathcal{F}_{n-1}}[\exp(\lambda(X_n+\mu))]\\ &= \int_{(-\infty,-b]}\exp(\lambda(x+\mu))\,\mathbb{P}_{\mathcal{F}_{n-1}}(X_n \in dx) + \int_{(-b,a]}\exp(\lambda(x+\mu))\,\mathbb{P}_{\mathcal{F}_{n-1}}(X_n \in dx)\\ &\leq \mathbb{P}_{\mathcal{F}_{n-1}}(X_n \leq -b)\exp(\lambda(-b+\mu)) + \mathbb{P}_{\mathcal{F}_{n-1}}(-b < X_n \leq a)\exp(\lambda(a+\mu))\\ &\leq p\exp(\lambda(-b+\mu)) + (1-p)\exp(\lambda(a+\mu)).\end{aligned}$$

Then observe that by our choice of μ, $-b+\mu < 0$, and that $|-b+\mu| \geq (a+\mu)\frac{1-p}{p}$. Since we assumed $p \leq \frac{1}{2}$, this means that $\frac{1-p}{p} \geq 1$ and so for any $k \geq 1$,

$$|-b+\mu| \geq (a+\mu)\frac{1-p}{p} \implies |-b+\mu| \geq (a+\mu)\left(\frac{1-p}{p}\right)^{1/k} \implies p|-b+\mu|^k \geq (1-p)(a+\mu)^k.$$

Consequently, by Taylor expansion,

$$\begin{aligned}&p\exp(\lambda(-b+\mu)) + (1-p)\exp(\lambda(a+\mu))\\ &= 1 + \sum_{k=1}^{\infty}\frac{\lambda^k(p(-b+\mu)^k + (1-p)(a+\mu)^k)}{k!} \leq 1 + \sum_{k=1}^{\infty}\frac{\lambda^k(p(-b+\mu)^k + p|-b+\mu|^k)}{k!}\\ &= 1 + \sum_{k=1}^{\infty}\frac{\lambda^{2k}2p|-b+\mu|^{2k}}{(2k)!} \leq 1 + \sum_{k=1}^{\infty}\frac{\lambda^{2k}|-b+\mu|^{2k}}{(k)!}\\ &= \exp(\lambda^2(-b+\mu)^2),\end{aligned}$$

which leads to

$$\mathbb{E}_{\mathcal{F}_{n-1}}[\exp(\lambda(X_n+\mu))] \leq \exp(\lambda^2(-b+\mu)^2).$$

Now, continuing from Eq. (13), we have that

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n}(X_i+\mu)\right)\right] &\leq \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n-1}(X_i+\mu)\right)\mathbb{E}_{\mathcal{F}_{n-1}}1[\exp(\lambda(X_n+\mu))]\right]\\ &\leq \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n-1}(X_i+\mu)\right)\exp(\lambda^2(b-\mu)^2)\right]\end{aligned}$$

$$: : :$$
$$\exp(n\ {}^2(b\ \ )^2):$$

Thus, for any > 0 and $s\ \ 0$,

$$P\left(\sum_{i=1}^n X_i \quad n\ + s\right) \quad \frac{E[\exp(\ (\sum_{i=1}^n (X_i +\ \ )))]}{\exp(\ s)}$$
$$\exp(n\ {}^2(b\ \ )^2\ \ \ s)$$

By finding the minimizing , we find that

$$P\left(\sum_{i=1}^n X_i \quad n\ + s\right) \quad \exp\ \frac{s^2}{4n(b\ \ )^2}\ ;$$

which completes the proof. □

We will later require a weakened form of a supermartingale concentration inequality, as stated and proven below.

Proposition 3 (Weakened supermartingale concentration inequality). Consider a filtration of sigma-algebras $F_0\ \ F_1$ F_n and a sequence of random variables $X_1; : : : ; X_n$ such that $X_i\ 2\ F_i$. Consider for each $i\ 2\ f\ 1; : : : ; ng$ a bad set B_i where $1_{B_i}\ 2\ F_{i\ \ 1}$, and suppose

$$P_{F_{i\ \ 1}}(X_i 1_{B_i^c}\ \ \ a) = 1 \qquad \text{and} \qquad P_{F_{i\ \ 1}}(X_i 1_{B_i^c}\ \ \ b)\ \ \ p$$

for some $a; b > 0$ and $0\ \ \ p\ \ \ 1=2$. Then, for any $0 <\ \ \ b$ such that $j\ \ b +\ \ j\ \ \ \frac{1\ \ p}{p}(a +\ \ )$, we have

$$P\left(\sum_{i=1}^n X_i \quad n\ + s\right) \quad \exp\ \frac{s^2}{4n(b\ \ )^2}\ + \sum_{i=1}^n P(X_i\ 2\ B_i); \qquad 8s > 0:$$

Proof. We define $Q_i := X_i 1_{B_i^c}$. We can then apply Lemma 12 and get

$$P\left(\sum_{i=1}^n Q_i \quad n\ + s\right) \quad \exp\ \frac{s^2}{4n(b\ \ )^2}\ :$$

Since $P(X_i\ 6=\ Q_i$ for some $i\ 2\ [n])\ \ \ \sum_i P(X_i\ 2\ B_i)$, it follows that

$$P\left(\sum_{i=1}^n X_i \quad n\ + s\right) \quad \exp\ \frac{s^2}{4n(b\ \ )^2}\ + \sum_{i=1}^n P(X_i\ 2\ B_i);$$

which completes the proof. □

D. Function decrease in large gradient regime

In this section, we show that sufficient function decrease can be made across the iterations with large gradients. We first restate and prove the function decrease lemma (Lemma 1), first introduced in the main text. We then provide a detailed roadmap of our proof in the subsequent discussion following the proof of Lemma 1.

Lemma 1 (Function decrease for batch zeroth-order optimization). Suppose at each time t, the algorithm performs the update step (with batch-size parameter $1\ \ m\ \ \ d$)

$$x_{t+1} = x_t \qquad g_u^{(m)}(x_t) + Y_t\ \ ;$$

where

$$g_u^{(m)}(x_t) = \frac{1}{m}\sum_{i=1}^m \frac{f(x_t + uZ_{t;i})\ \ \ f(x_t\ \ \ uZ_{t;i})}{2u} Z_{t;i}\ ;$$

where each $Z_{t,i}$ is drawn i.i.d from $\mathcal{N}(0, I)$, $u > 0$ is the smoothing radius, and $Y_t \sim \mathcal{N}(0, \frac{r^2}{d} I)$ with $r > 0$ denoting the perturbation radius.

Then, there exist absolute constants $c_1 > 0, C_1 \geq 1$ such that, for any $\mathcal{T} \in \mathbb{Z}^+$ and $\mathcal{F} > 0$, $\eta > 0$ and $\delta \in (0, 1/e]$, upon defining $H_{0,\tau}(\mathcal{F})$ to be the event on which the inequality

$$\begin{aligned} f(x_\tau) - f(x_0) \leq & -\frac{3\eta}{4}\sum_{t=0}^{\tau-1}\frac{1}{m}\sum_{i=1}^{m}\left(Z_{t,i}^\top \nabla f(x_t)\right)^2 && (3)\\ & + \eta\left(\frac{\delta}{\mathcal{T}} + \frac{c_1 L\eta^2\chi^3 d}{m}\right)\sum_{t=0}^{\tau-1}\|\nabla f(x_t)\|^2 && (4)\\ & + \eta\rho^2 u^4\tau^2 c_1 d^3\left(\log\frac{\mathcal{T}}{\delta}\right)^3 + L\eta^2\rho^2 u^4\tau^2 c_1 d^4\left(\log\frac{\mathcal{T}}{\delta}\right)^4 \\ & + c_1 r^2(\sqrt{\tau} + L\tau)\log\frac{\mathcal{T}}{\delta} + c_1 L\eta^2 r^2 && (5) \end{aligned}$$

is satisfied (where $\chi := \log(C_1 d m \mathcal{T}/\delta)$), we have

$$P(H_{0,\tau}(\mathcal{F})) \geq 1 - \frac{(\tau+4)\delta}{\mathcal{T}}; \qquad P\left(\bigcap_{\tau=1}^{\tau'} H_{0,\tau}(\mathcal{F})\right) \geq 1 - \frac{5\tau'\delta}{\mathcal{T}}$$

for any $0 \leq \tau' \leq \mathcal{T}$.

Proof. First, for each $t \in \{-1, \ldots, \mathcal{T}\}$, we define $\mathcal{F}_t$ to be the sigma-algebra generated by

$$x_0; \quad \sigma(\{Z_{0,i}\}_{i=1}^m, \ldots, \{Z_{t,i}\}_{i=1}^m); \quad \sigma(Y_0, \ldots, Y_t).$$

Note that $\mathcal{F}_{-1}$ is the sigma-algebra generated only by x_0.

By Taylor expansion, for any $x, y \in \mathbb{R}^d$, there exists $\theta \in [0, 1]$ such that $f(x + y) = f(x) + \langle\nabla f(x), y\rangle + \frac{1}{2}y^\top \nabla^2 f(x + \theta y)y$. Therefore

$$\frac{f(x_t + uZ_{t,i}) - f(x_t - uZ_{t,i})}{2u} = \langle\nabla f(x), Z_{t,i}\rangle + \frac{u}{2}Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i}$$

with

$$\tilde{H}_{t,i} = \frac{\nabla^2 f(x + \theta_{i,+}uZ_{t,i}) - \nabla^2 f(x - \theta_{i,-}uZ_{t,i})}{2}$$

for some $\theta_{i,\pm} \in [0, 1]$, and

$$x_{t+1} = x_t - \left(\eta\frac{1}{m}\sum_{i=1}^{m}\left(Z_{t,i}Z_{t,i}^\top \nabla f(x_t) + \frac{u}{2}Z_{t,i}Z_{t,i}^\top \tilde{H}_{t,i}Z_{t,i}\right) + Y_t\right) \tag{14}$$

By the ρ-Hessian Lipschitz property of f, it follows that $\|\tilde{H}_{t,i}\| \leq \rho u\|Z_{t,i}\|$.

Observe that

$$\begin{aligned} f(x_{t+1}) & \overset{(i)}{\leq} f(x_t) + \langle x_{t+1} - x_t, \nabla f(x_t)\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\ & \overset{(ii)}{=} f(x_t) - \eta\frac{1}{m}\sum_{i=1}^{m}\left(Z_{t,i}^\top \nabla f(x_t)\right)^2 - \eta\frac{1}{m}\sum_{i=1}^{m}\frac{u}{2}Z_{t,i}^\top \nabla f(x_t) \cdot Z_{t,i}^\top \tilde{H}_{t,i}Z_{t,i} - \langle\nabla f(x_t), Y_t\rangle \\ & \quad + \frac{L}{2}\left\|\eta\frac{1}{m}\sum_{i=1}^{m}\left(Z_{t,i}Z_{t,i}^\top \nabla f(x_t) + \frac{u}{2}Z_{t,i}Z_{t,i}^\top \tilde{H}_{t,i}Z_{t,i}\right) + Y_t\right\|^2 \\ & \overset{(iii)}{\leq} f(x_t) - \frac{\eta}{m}\sum_{i=1}^{m}\left(Z_{t,i}^\top \nabla f(x_t)\right)^2 + \frac{\eta}{m}\sum_{i=1}^{m}\left(\frac{\left(Z_{t,i}^\top \nabla f(x_t)\right)^2}{4} + \frac{u^2\left(Z_{t,i}^\top \tilde{H}_{t,i}Z_{t,i}\right)^2}{4}\right) - \langle\nabla f(x_t), Y_t\rangle \end{aligned}$$

$$+\frac{L\eta^2}{2}\left(2\left\|\frac{1}{m}\sum_{i=1}^m Z_{t,i}Z_{t,i}^\top\nabla f(x_t)\right\|^2+u^2\left\|\frac{1}{m}\sum_{i=1}^m Z_{t,i}Z_{t,i}^\top\tilde{H}_{t,i}Z_{t,i}\right\|^2+4\|Y_t\|^2\right)$$

$$\overset{(iv)}{\le} f(x_t)-\frac{3\eta}{4m}\sum_{i=1}^m\left\|Z_{t,i}^\top\nabla f(x_t)\right\|^2+\frac{u^2\eta}{m}\sum_{i=1}^m\frac{u^2\rho^2\|Z_{t,i}\|^6}{4}-\eta\langle\nabla f(x_t),Y_t\rangle$$

$$+\frac{L\eta^2}{2}\left(2\left\|\frac{1}{m}\sum_{i=1}^m Z_{t,i}Z_{t,i}^\top\nabla f(x_t)\right\|^2+\frac{u^2}{m}\sum_{i=1}^m u^2\rho^2\|Z_{t,i}\|^8+4\|Y_t\|^2\right)$$

$$\le f(x_t)-\frac{3\eta}{4m}\sum_{i=1}^m\left\|Z_{t,i}^\top\nabla f(x_t)\right\|^2+\frac{u^4\rho^2\eta}{4m}\sum_{i=1}^m\|Z_{t,i}\|^6+\frac{L\rho^2u^4\eta^2}{2m}\sum_{i=1}^m\|Z_{t,i}\|^8-\eta\langle\nabla f(x_t),Y_t\rangle$$

$$+\frac{L\eta^2}{2}\left(2\left\|\frac{1}{m}\sum_{i=1}^m Z_{t,i}Z_{t,i}^\top\nabla f(x_t)\right\|^2+4\|Y_t\|^2\right) \tag{15}$$

Above, to derive (i), we used the L-smoothness of f. To derive (ii), we used the expression for $(x_{t+1}-x_t)$ shown in Eq. (14). To derive (iii), we used the fact that $ab\le(a^2+b^2)/2$ for any $a,b\in\mathbb{R}_{\ge 0}$, as well as two applications of the fact that $\|a+b\|^2\le 2(\|a\|^2+\|b\|^2)$ for any two vectors $a,b\in\mathbb{R}^d$. To derive (iv), we used the fact that $\|\tilde{H}_{t,i}\|\le\rho u\|Z_{t,i}\|$.

To continue from Eq. (15), we first observe that we can rewrite

$$Z_{t,i}Z_{t,i}^\top\nabla f(x_t)=(Z_{t,i}Z_{t,i}^\top-I)\nabla f(x_t)+\nabla f(x_t),$$

so that

$$\left\|\frac{1}{m}\sum_{i=1}^m Z_{t,i}Z_{t,i}^\top\nabla f(x_t)\right\|^2\le 2\left\|\frac{1}{m}\sum_{i=1}^m(Z_{t,i}Z_{t,i}^\top-I)\nabla f(x_t)\right\|^2+2\|\nabla f(x_t)\|^2.$$

Observe that we can apply the bound in Proposition 2 to $\left\|\sum_{i=1}^m(Z_{t,i}Z_{t,i}^\top-I)\nabla f(x_t)\right\|$, and since $Z_{t,i}$ is independent of $\mathcal{F}_{t-1}$ for all i, we know there exist absolute constants $c_1>0, C_1\ge 1$ such that for any $\delta\in(0,1/e]$ and $\theta>0$, with probability at least $1-\delta$ conditioned on $\mathcal{F}_{t-1}$,

$$\left\|\sum_{i=1}^m(Z_{t,i}Z_{t,i}^\top-I)\nabla f(x_t)\right\|\le c_1\sum_{i=1}^m\left(\theta d(\ln(C_1m/\delta))^2\|\nabla f(x_t)\|^2+\frac{1}{\theta}\log(C_1dm/\delta)\right)$$

$$=c_1\left(\theta md(\ln(C_1m/\delta))^2\|\nabla f(x_t)\|^2+\frac{1}{\theta}\log(C_1dm/\delta)\right). \tag{16}$$

Moreover, since $C_1\ge 1$, $\log(C_1dm/\delta)$ and $\ln(C_1m/\delta)$ both are at least 1 as long as $\delta\in(0,1/e]$. Observe that conditioned on $\mathcal{F}_{t-1}$, $\nabla f(x_t)$ is fixed. Hence, we can pick

$$\theta=\frac{1}{\sqrt{c_1md\ln(C_1dm/\delta)}\,\|\nabla f(x_t)\|}$$

which is $\mathcal{F}_{t-1}$-measurable, and plug it into Eq. (16) to find that the probability conditioned on $\mathcal{F}_{t-1}$ of the following event

$$\left\|\sum_{i=1}^m(Z_{t,i}Z_{t,i}^\top-I)\nabla f(x_t)\right\|\le 2\sqrt{c_1}(\ln(C_1dm/\delta))^{3/2}\sqrt{md}\,\|\nabla f(x_t)\| \tag{17}$$

is at least $1-\delta$. By taking the total expectation, it follows that the event has a total probability at least $1-\delta$ Thus, with probability at least $1-\delta$,

$$\left\|\frac{1}{m}\sum_{i=1}^m Z_{t,i}Z_{t,i}^\top\nabla f(x_t)\right\|^2\le 2\left\|\frac{1}{m}\sum_{i=1}^m(Z_{t,i}Z_{t,i}^\top-I)\nabla f(x_t)\right\|^2+2\|\nabla f(x_t)\|^2$$

$$
\begin{aligned}
&4c_1(\mathrm{lr}(\ C_1 dm=\ ))^3\frac{d}{m}\mathrm{kr}\ f(x_t)\mathrm{k}^2+2\,\mathrm{kr}\ f(x_t)\mathrm{k}^2\\
&c_2(\mathrm{lr}(\ C_1 dm=\ ))^3\frac{d}{m}\mathrm{kr}\ f(x_t)\mathrm{k}^2; \qquad (18)
\end{aligned}
$$

where the last inequality comes from the fact that $\mathrm{lr}(\ C_1 dm=\ )\quad 1$, our assumption at the outset of the appendix that $d\quad m$, and denoting $c_2 := 4c_1 + 2$.

Denote the event $H_{0;}\ (\ )$ as the event that

$$
\begin{aligned}
f(x\ )\quad f(x_0)\quad & \mathrm{X}^{\ 1}_{t=0}\frac{3}{4m}\mathrm{X}^{m}_{i=1}\ Z^{>}_{t;i}\ r\ f(x_t)\ ^2+L\ ^2\frac{c_2 d(\mathrm{lr}(\ C_1 dm=\ ))^3}{m}\mathrm{X}^{\ 1}_{t=0}\mathrm{kr}\ f(x_t)\mathrm{k}^2\\
&+\frac{u^4\ ^2}{4m}\mathrm{X}^{\ 1}_{t=0}\mathrm{X}^{m}_{i=1}\mathrm{k}Z_{t;i}\ \mathrm{k}^6+\frac{L\ ^2u^4\ ^2}{2m}\mathrm{X}^{\ 1}_{t=0}\mathrm{X}^{m}_{i=1}\mathrm{k}Z_{t;i}\ \mathrm{k}^8\\
&\mathrm{X}^{\ 1}_{t=0}\mathrm{hr}\ f(x_t);Y_t\mathrm{i}+2L\ ^2\mathrm{X}^{\ 1}_{t=0}\mathrm{k}Y_t\mathrm{k}^2 \qquad (19)
\end{aligned}
$$

holds.

Now, continuing from Eq. (15), and using the bound in Eq. (18), summing over the iterations from $t = 0$ to $\quad 1$, we find using the union bound that $P(\backslash^{\ 0}_{\ =1}\ H_{0;}\ (\ ))\quad 1\quad\ ^0$, $P(H_{0;}\ (\ ))\quad 1\quad$:

Now, by Lemma 6, for any $\ 2\ (0;1);\ >\ 0$, with probability at least $1\quad$, there exists an absolute constant $c_3 > 0$ such that

$$
\mathrm{X}^{\ 1}_{t=0}\mathrm{hr}\ f(x_t);Y_t\mathrm{i}\qquad \frac{1}{\ }\mathrm{X}^{\ 1}_{t=0}\mathrm{kr}\ f(x_t)\mathrm{k}^2+c_3\ r\ ^2\log(1=\ )\ ^{!}\ : \qquad (20)
$$

Meanwhile, since $Y_t\quad N(0;(r\ ^2=d)I)$, $\mathrm{k}Y_t\mathrm{k}^2$ is sub-exponential with sub-exponential norm cr^2 for some absolute constant $c > 0$, and by Bernstein's inequality (Lemma 8), there exists some absolute constant $c_4 > 0$ such that

$$
\mathrm{X}^{\ 1}_{t=0}\mathrm{k}Y_t\mathrm{k}^2\quad c_4 r^2(\quad+\log(1=\ )) \qquad (21)
$$

with probability at least $1\quad$.

To bound $\mathrm{P}^{\ 1}_{t=0}\frac{1}{m}\mathrm{P}^{m}_{i=1}\mathrm{k}Z_{t;i}\ \mathrm{k}^6$ and $\mathrm{P}^{\ 1}_{t=0}\frac{1}{m}\mathrm{P}^{m}_{i=1}\mathrm{k}Z_{t;i}\ \mathrm{k}^8$, both sums of heavy tailed Gaussian moments, we use Lemma 11, which states that for any $k\ 2\ Z^+$ and $\ 2\ (0;1)$, with probability at least $1\quad$,

$$
\frac{1}{m}\mathrm{X}^{\ 1}_{t=0}\mathrm{X}^{m}_{i=1}\mathrm{k}Z_{t;i}\ \mathrm{k}^{2k}\quad c_5\ (c_6)^k d^k(1+(\log(1=\ ))^k) \qquad (22)
$$

for some absolute constants $c_5, c_6 > 0$. As in the statement of the proof, using $\ = \mathrm{lr}(\ C_1 dm=\ )$ to ease the notation, denote the event that

$$
\begin{aligned}
f(x\ )\quad f(x_0)\quad & \frac{3}{4}\mathrm{X}^{\ 1}_{t=0}\frac{1}{m}\mathrm{X}^{m}_{i=1}\ Z^{>}_{t;i}\ r\ f(x_t)\ ^2+\quad\frac{}{\ }+\frac{c_2L\ ^2\ ^3d}{m}\mathrm{X}^{\ 1}_{t=0}\mathrm{kr}\ f(x_t)\mathrm{k}^2\\
&+\frac{u^4\ ^2}{2}\ c_5c_6^3d^3\quad\log\frac{1}{\ }\ ^3+\ L\ ^2u^4\ ^2\ c_5c_6^4d^4\quad\log\frac{1}{\ }\ ^4\\
&+\ (c_3\ r\ ^2+2c_4\ Lr\ ^2)\log\frac{1}{\ }+2c_4L\ ^2\ r\ ^2
\end{aligned}
$$

holds as $H_{0;}\ (\ )$.

Plugging Eq. (20), Eq. (21), and Eq. (22) into Eq. (19), by union bound, we see that

$$P(\backslash_{=1}^{0} H_{0;}\ (\ ))\quad 1\quad (\ ^0 + 4\ ^0) = 1\quad 5\ ^0; \qquad P(H_{0;}\ )\quad 1\quad (\ +4)\ :$$

The final result then follows by rescaling to $\overline{T}$ and denoting $c_1 := \max f\, c_2; c_3; 2c_4; c_5c_6^3 = 2; c_5c_6^4 g.$ □

Outline of proof approach. Similar to the first-order setting, our goal is to show that we can arrive at a contradiction $f(x_T) < \min_x f(x)$ when there is a large number of steps at which $kr\ f(x_t)k$. Roughly speaking, as Eq. (5) shows, we need to prove a lower bound of the form

$$\sum_{t=0}^{T\ 1} \frac{1}{m}\sum_{i=1}^{m} Z_{t;i}^{>} r\ f(x_t)\ ^2 \qquad \frac{1}{} + \frac{c_1 L\ \ ^3 d}{m} \sum_{t=0}^{T\ 1} kr\ f(x_t)k^2 \tag{23}$$

for some which is not too large (an example would be picking such that it only scales logarithmically in the problem parameters). However, it is tricky to prove such a lower-bound in the zeroth-order setting. In particular, for small batch-sizes m, $\frac{1}{m}\sum_{i=1}^{m} Z_{t;i}^{>} r\ f(x_t)\ ^2$ could be small even as $kr\ f(x_t)k^2$ is large; this is because for each $i\ 2\ [m]$, $Z_{t;i}$ could have a negligible component in the $r\ f(x_t)$ direction. This necessitates a more careful analysis to prove a bound similar to Eq. (23). We do so using the following approach.

1. Intuitively, whilst for each individual iteration, $\frac{1}{m}\sum_{i=1}^{m} Z_{t;i}^{>} r\ f(x_t)\ ^2$ could be small even as $kr\ f(x_t)k^2$ is large, in a small number of (consecutive) iterations $f\, t_0; :::; t_0 + t_f\ g$, with high probability, there will be at least one iteration t within $f\, t_0; :::; t_0 + t_f\quad 1g$, such that $\frac{1}{m}\sum_{i=1}^{m} Z_{t;i}^{>} r\ f(x_t)\ ^2 = (\ kr\ f(x_t)k^2)$. We formalize this intuition in Lemma 14. Thus, we consider breaking the time-steps into chunks where each chunk has t_f consecutive iterations.

2. Consider any such interval $f\, t_0; :::; t_0 + t_f\quad 1g$. There are two cases to consider.

   (a) The first case is when the gradient throughout all t_f iterations is large enough to dominate the perturbation terms. Intuitively, in this case, it is not hard to see that given appropriate parameter choices, the gradient will change little throughout the t_f iterations. In fact, as we formalize in Lemma 16, for an appropriate choice of t_f and , we can show that

   $$\frac{1}{2}kr\ f(x_{t_0})k\quad kr\quad f(x_t)k\quad 2kr\ f(x_{t_0})k \qquad 8t\ 2\ f\, t_0; :::; t_0 + t_f\quad 1g:$$

   As a result, combined with point 1, we see that

   $$\sum_{t=t_0}^{t_0+t_f\ \ 1} \frac{1}{m}\sum_{i=1}^{m} Z_{t;i}^{>} r\ f(x_t)\ ^2 \qquad (\ kr\ f(x_{t_0})k^2):$$

   Thus, by choosing and judiciously, for such intervals, it is possible to show that

   $$\begin{aligned}\sum_{t=t_0}^{t_0+t_f\ \ 1} \frac{1}{m}\sum_{i=1}^{m} Z_{t;i}^{>} r\ f(x_t)\ ^2 \quad (\ kr\ f(x_{t_0})k^2) \qquad & \frac{1}{} + \frac{c_1 L\ \ ^3 d}{m} \sum_{t=t_0}^{t_0+t_f\ \ 1} kr\ f(x_t)k^2 \\ = \quad & \frac{1}{} + \frac{c_1 L\ \ ^3 d}{m} \qquad t_f\, kr\ f(x_{t_0})k^2\end{aligned}$$

   Thus, in these intervals, it is possible to obtain function improvement on the order of $(\ kr\ f(x_{t_0})k^2)$.

   (b) The remaining case is when the gradient is small and dominated by the perturbation terms in any one of the t_f iterations. In this case, as we show in Lemma 17, for each of the t_f iterations, the gradient will be small and on the same scale as the perturbation terms. In turn, by choosing r, t_f and appropriately, we can make the perturbation terms small. Thus, whilst these intervals may not contribute to function decrease, they also contribute little in the way of function increase.

3. When there are at least $T/4$ iterations with large gradient (i.e. $\|\nabla f(x_t)\| \geq \epsilon$), assuming t_f divides T, it follows that there are at least $T/(4t_f)$ intervals of length t_f where one iteration in the interval contains a large gradient. By choosing u, r and η appropriately such they are dominated by ϵ, it is possible to show that with high probability, such an interval cannot belong to the second case above, and must instead be from the first case. Since $\|\nabla f(x_t)\| \approx \|\nabla f(x_{t_0})\|$ for each $t \in \{t_0, \ldots, t_0 + t_f - 1\}$ in this case, and we know that one of the iterations has a gradient with size at least ϵ, it follows that we make function decrease progress of at least $\Omega(\epsilon^2)$ for such intervals. By appropriately choosing u and r to limit the effects of the intervals of the second form, we can then show a contradiction of the form $f(x_T) < f^*$. We demonstrate this formally in Proposition 4.

We formalize our approach in the following series of results. First, for analytical convenience, we prove the following result showing that for any t, the perturbation terms $\|Y_t\|$ and $\frac{1}{m}\sum_{i=1}^m \|Z_{t,i}\|^4$ are bounded with high probability.

Lemma 13. There exists an absolute constant $c_3 > 0$ such that, for any $t \in \mathbb{N}$, the event

$$G_t(\delta) := \left\{ \|Y_t\|^2 \leq c_3^2 r^2 \left(1 + \frac{\log(T/\delta)}{d}\right) \quad \text{and} \quad \frac{1}{m}\sum_{i=1}^m \|Z_{t,i}\|^4 \leq 2c_3 d^2 \left(\log\frac{T}{\delta}\right)^2 \right\}$$

has probability at least $1 - 2\delta/T$ for any $\delta \in (0, 1/e]$.

Proof. Noting that $Y_t \sim \mathcal{N}(0, (r^2/d)I)$, by applying Bernstein's inequality (Lemma 8), it can be shown that with probability at least δ/T,

$$\|Y_t\|^2 \leq c_3^2 r^2 \left(1 + \frac{\log(T/\delta)}{d}\right);$$

where $c_3 > 0$ is some absolute constant. Then by using Lemma 11, applying the union bound, and redefining the constant c_3, we complete the proof. □

Next, in Lemma 14, we show that in a small number of iterations, with high probability, there exists some iteration t such that $\frac{1}{m}\sum_{i=1}^m \left(Z_{t,i}^\top \nabla f(x_t)\right)^2 \geq \frac{1}{2}\|\nabla f(x_t)\|^2$.

Lemma 14. There exists an absolute constant $c_2 \geq 1$ such that, upon defining

$$t_f(\delta) = \frac{c_2}{m}\log\frac{T}{\delta}; \qquad \delta > 0;$$

and defining the event

$$B_{t_0}(\delta; k) := \bigcup_{t=t_0}^{t_0+k-1} \left\{ \frac{1}{m}\sum_{i=1}^m \left(Z_{t,i}^\top \nabla f(x_t)\right)^2 \geq \frac{1}{2}\|\nabla f(x_t)\|^2 \right\};$$

we have

$$P(B_{t_0}(\delta; k)) \geq 1 - \frac{\delta}{T}:$$

for any $\delta \in (0, 1)$, $t_0 \in \mathbb{N}$ and $k \geq t_f(\delta)$.

Proof. Denote the event

$$E_t = \left\{ \frac{1}{m}\sum_{i=1}^m |Z_{t,i}^\top \nabla f(x_t)|^2 < \frac{1}{2}\|\nabla f(x_t)\|^2 \right\}:$$

Observe that, conditioned on $\mathcal{F}_{t-1}$, the set of random variables $\left\{\|\nabla f(x_t)\|^2 - |Z_{t,i}^\top \nabla f(x_t)|^2\right\}_{i=1}^m$ are independent, mean-zero, and subexponential with subexponential norm $c\|\nabla f(x_t)\|^2$ for some absolute constant $c > 0$. Hence

$$\begin{aligned} P_{\mathcal{F}_{t-1}}(E_t) &= P_{\mathcal{F}_{t-1}}\left(\frac{1}{m}\sum_{i=1}^m |Z_{t,i}^\top \nabla f(x_t)|^2 < \frac{1}{2}\|\nabla f(x_t)\|^2 \right) \\ &= P_{\mathcal{F}_{t-1}}\left(\sum_{i=1}^m \left(\|\nabla f(x_t)\|^2 - \left(Z_{t,i}^\top \nabla f(x_t)\right)^2 \right) > \frac{m}{2}\|\nabla f(x_t)\|^2 \right) \end{aligned}$$

$$\le \exp(-c'm);$$

where c' is some positive absolute constant. Then, for any $t_0, k \in \mathbb{N}$,

$$\begin{aligned} &\mathbb{P}\left(\left(\frac{1}{m}\sum_{i=1}^{m} Z_{t,i}^{\top}\nabla f(x_t)\right)^2 < \frac{1}{2}\|\nabla f(x_t)\|^2 \text{ for every } t\in[t_0;t_0+k)\right) \\ &= \mathbb{E}\left[\prod_{t=t_0}^{t_0+k-1} 1_{E_t}\right] = \mathbb{E}\left[\prod_{t=t_0}^{t_0+k-2} 1_{E_t}\,\mathbb{E}_{\mathcal{F}_{t_0+k-2}} 1_{E_{t_0+k-1}}\right] \\ &\le \exp(-c'm)\,\mathbb{E}\left[\prod_{t=t_0}^{t_0+k-2} 1_{E_t}\right] \le \cdots \le \exp(-c'mk). \end{aligned}$$

Therefore, by letting $c_2 = \max\{1, 1/c'\}$ and

$$k \ge t_f(\delta) = \frac{c_2}{m}\log\frac{T}{\delta};$$

we get

$$\mathbb{P}\left(\left(\frac{1}{m}\sum_{i=1}^{m} Z_{t,i}^{\top}\nabla f(x_t)\right)^2 < \frac{1}{2}\|\nabla f(x_t)\|^2 \text{ for every } t\in[t_0;t_0+k)\right) \le \frac{\delta}{T};$$

which completes the proof. □

The term $t_f(\delta)$ will frequently appear in the proofs to come; in the sequel we denote

$$t_f(\delta) := \frac{c_2}{m}\log\frac{T}{\delta}; \qquad \delta\in(0;1/e]; \tag{24}$$

where $c_2 \ge 1$ is the absolute constant defined in Lemma 14.

We next show that with high probability, the norm difference term $\|\nabla f(x_{t+1})\| - \|\nabla f(x_t)\|$ can be bounded in terms of $\|\nabla f(x_t)\|$ and the perturbation terms $\frac{u}{2m}\sum_{i=1}^{m} Z_{t,i}Z_{t,i}^{\top}\tilde{H}_{t,i}Z_{t,i}$ as well as $\|Y_t\|$.

Lemma 15. Define

$$A_t(\delta) := \left\{\|\nabla f(x_{t+1})\| - \|\nabla f(x_t)\| \le \frac{\|\nabla f(x_t)\|}{8t_f(\delta)} + L\eta\left(\left\|\frac{u}{2m}\sum_{i=1}^{m} Z_{t,i}Z_{t,i}^{\top}\tilde{H}_{t,i}Z_{t,i}\right\| + \|Y_t\|\right)\right\} \tag{25}$$

where $t_f(\delta)$ is defined in Eq. (24), and let $C_1 \ge 1$ be the corresponding absolute constants defined in Lemma 1. Then there exists an absolute constant $c_4 > 0$ such that, whenever η satisfies

$$L\eta\,\frac{c_4(\ln(C_1 dmT/\delta))^{3/2}\sqrt{d}}{\sqrt{m}} \le \frac{1}{8t_f(\delta)}; \tag{26}$$

we have

$$\mathbb{P}(A_t(\delta)) \ge 1 - \frac{\delta}{T}$$

for any $\delta\in(0;1/e]$ and $t\in\mathbb{Z}^{+}$.

Proof. Since ∇f is L-Lipschitz, following the zeroth-order update step, we see that

$$\begin{aligned} \|\nabla f(x_{t+1})\| - \|\nabla f(x_t)\| &\le L\|x_{t+1} - x_t\| \qquad (27)\\ &= L\eta\left\|\frac{1}{m}\sum_{i=1}^{m} Z_{t,i}Z_{t,i}^{\top}\nabla f(x_t) + \frac{u}{2m}\sum_{i=1}^{m} Z_{t,i}Z_{t,i}^{\top}\tilde{H}_{t,i}Z_{t,i} + Y_t\right\|. \qquad (28) \end{aligned}$$

Now, it follows from Eq. (18) (with a slight modification in the absolute constant terms since here the norm is not squared) that there exists some absolute constant $c_4 > 0$ such that for any $\delta \in (0, 1/e]$, we have that with probability at least $1-\delta/T$, the event

$$\left\| \frac{1}{m}\sum_{i=1}^{m} Z_{t,i} Z_{t,i}^{\top} \nabla f(x_t) \right\| \le c_4(\ln(C_1 dmT/\delta))^{3/2}\sqrt{\frac{d}{m}}\|\nabla f(x_t)\|;$$

Hence, continuing from Eq. (28), it follows that with probability at least $1-\delta/T$,

$$\begin{aligned}&\|\nabla f(x_{t+1}) - \nabla f(x_t)\| \\ &\le L\left(c_4(\ln(C_1 dmT/\delta))^{3/2}\sqrt{\frac{d}{m}}\|\nabla f(x_t)\| + \frac{u}{2m}\left\|\sum_{i=1}^{m} Z_{t,i} Z_{t,i}^{\top} \tilde{H}_{t,i} Z_{t,i}\right\| + \|Y_t\|\right);\end{aligned}$$

and by plugging in the condition Eq. (26), we see that the event

$$A_t(\delta) = \left\{\|\nabla f(x_{t+1}) - \nabla f(x_t)\| \le \frac{\|\nabla f(x_t)\|}{8t_f(\delta)} + L\left(\frac{u}{2m}\left\|\sum_{i=1}^{m} Z_{t,i} Z_{t,i}^{\top} \tilde{H}_{t,i} Z_{t,i}\right\| + \|Y_t\|\right)\right\}$$

has probability at least $1-\delta/T$. □

We show now that if the norm of the gradient dominates the norm of the perturbation terms, and we choose the step-size sufficiently small, then in a small number of iterations, the norm of the gradient does not change very much. For notational simplicity, we denote the event

$$E(t_1, t_2, \delta) := \bigcap_{t=t_1}^{t_1+t_2-1}\left\{\|\nabla f(x_t)\| > 8t_f(\delta)\, L\left(\frac{u}{2}\left\|\frac{1}{m}\sum_{i=1}^{m} Z_{t,i} Z_{t,i}^{\top} \tilde{H}_{t,i} Z_{t,i}\right\| + \|Y_t\|\right)\right\}.$$

Lemma 16. Let $\delta \in (0, 1/e]$ and $T \in \mathbb{Z}^{+}$ be such that $T > 2t_f(\delta) + 1$. Consider any positive integer $t_f' \le 2t_f(\delta)$, and any $t_0 \in \{0, \ldots, T-1-t_f'\}$. Suppose δ satisfies the condition Eq. (26). Then, on the event

$$E(t_0, t_f', \delta) \cap \left(\bigcap_{t=t_0}^{t_0+t_f'-1} A_t(\delta)\right);$$

we have

$$\frac{1}{2}\|\nabla f(x_0)\| \le \|\nabla f(x_t)\| \le 2\|\nabla f(x_0)\|$$

for all $t \in \{t_0, \ldots, t_0 + t_f' - 1\}$.

Proof. By plugging

$$\|\nabla f(x_t)\| > 8t_f(\delta)\, L\left(\frac{u}{2}\left\|\frac{1}{m}\sum_{i=1}^{m} Z_{t,i} Z_{t,i}^{\top} \tilde{H}_{t,i} Z_{t,i}\right\| + \|Y_t\|\right)$$

into the definition of $A_t(\delta)$, we see that, on the event $E(t_0, t_f', \delta) \cap \bigcap_{t=t_0}^{t_0+t_f'-1} A_t(\delta)$, we have

$$\|\nabla f(x_{t+1}) - \nabla f(x_t)\| \le \frac{\|\nabla f(x_t)\|}{4t_f(\delta)};$$

and consequently,

$$\left(1 - \frac{1}{4t_f(\delta)}\right)\|\nabla f(x_t)\| \le \|\nabla f(x_{t+1})\| \le \left(1 + \frac{1}{4t_f(\delta)}\right)\|\nabla f(x_t)\|;$$

which leads to

$$\left(1 - \frac{1}{4t_f(\delta)}\right)^{t-t_0}\|\nabla f(x_0)\| \le \|\nabla f(x_t)\| \le \left(1 + \frac{1}{4t_f(\delta)}\right)^{t-t_0}\|\nabla f(x_0)\|$$

for all $t \in \{t_0, \ldots, t_0 + t_f'\}$. Then, since $(1 + 1/(4x))^{2x} \le 2$ and $(1 - 1/(4x))^{2x} \ge 1/2$ for any $x \ge 1$, noting that $t_f' \le 2t_f(\delta)$, we get the desired result. □

Conversely, in the following result, we show that in a small number of consecutive iterations, if the gradient is smaller than the perturbation terms in any one of the iterations, then for each of the iterations in this range, the gradient will be small and be on the same scale as the size of the perturbation terms.

Lemma 17. Let $\epsilon \in (0, 1/e]$ and $T \in \mathbb{Z}^+$ be such that $T > 2t_f(\epsilon) + 1$. Consider any positive integer $t_f' \leq 2t_f(\epsilon)$, and any $t_0 \in \{0, \ldots, T-1-t_f'\}$. Suppose η satisfies the condition Eq. (26). Then, on the event

$$\mathcal{E}^c(t_0, t_f'; \epsilon) \cap \left(\bigcap_{t=t_0}^{t_0+t_f'-1} \mathcal{A}_t(\epsilon)\right) \cap \left(\bigcap_{t=t_0}^{t_0+t_f'-1} \mathcal{G}_t(\epsilon)\right),$$

we have

$$\|\nabla f(x_t)\| \leq c_5 t_f(\epsilon) L\left(u^2 d^2 \left(\log\frac{T}{\epsilon}\right)^2 + \left(1+\sqrt{\frac{\log(T/\epsilon)}{d}}\right) r\right) \quad \forall t \in \{t_0, t_0+1, \ldots, t_0+t_f'-1\},$$

where c_5 is some absolute constant.

Proof. Let t' be the first iteration in $\{t_0, t_0+1, \ldots, t_0+t_f'-1\}$ such that

$$\|\nabla f(x_{t'})\| \leq 8t_f(\epsilon) L\left(\left\|\frac{u}{2}\frac{1}{m}\sum_{i=1}^{m} Z_{t',i} Z_{t',i}^{\top} H_{t',i} Z_{t',i}\right\| + \|Y_{t'}\|\right). \tag{29}$$

Since we are working on an event which is a subset of $\mathcal{E}^c(t_0, t_f'; \epsilon)$, t' is well-defined. By $\|H_{t',i}\| \leq u\|Z_{t',i}\|$, we see that

$$\begin{aligned}\|\nabla f(x_{t'})\| &\leq 8t_f(\epsilon) L\left(\frac{u^2}{2m}\sum_{i=1}^{m}\|Z_{t',i}\|^4 + \|Y_{t'}\|\right)\\ &\leq 8t_f(\epsilon) L\left(c_3 u^2 d^2\left(\log\frac{T}{\epsilon}\right)^2 + c_3\left(1+\sqrt{\frac{\log(T/\epsilon)}{d}}\right) r\right),\end{aligned}$$

where we used the definition of $\mathcal{G}_t(\epsilon)$.

Recall that t' is the first time step such that Eq. (29) holds. By deriving similarly as in the proof of Lemma 16, we can show that for any $j \in \{t_0, t_0+1, \ldots, t'-1\}$,

$$\|\nabla f(x_j)\| \leq 2\|\nabla f(x_{t'})\| \leq 16 t_f(\epsilon) L c_3\left(u^2 d^2\left(\log\frac{T}{\epsilon}\right)^2 + \left(1+\sqrt{\frac{\log(T/\epsilon)}{d}}\right) r\right).$$

Meanwhile, for iterations $t \in [t', t_0+t_f')$, by using the definitions of $\mathcal{A}_t(\epsilon)$ and $\mathcal{G}_t(\epsilon)$, we have

$$\begin{aligned}\|\nabla f(x_{t+1})\| &\leq \left(1+\frac{1}{8t_f(\epsilon)}\right)\|\nabla f(x_t)\| + Lc_3\left(u^2 d^2\left(\log\frac{T}{\epsilon}\right)^2 + \left(1+\sqrt{\frac{\log(T/\epsilon)}{d}}\right) r\right)\\ &= \left(1+\frac{1}{8t_f(\epsilon)}\right)^{t+1-t'}\|\nabla f(x_{t'})\|\\ &\quad + \sum_{i=0}^{t-t'}\left(1+\frac{1}{8t_f(\epsilon)}\right)^{t-t'-i} Lc_3\left(u^2 d^2\left(\log\frac{T}{\epsilon}\right)^2 + \left(1+\sqrt{\frac{\log(T/\epsilon)}{d}}\right) r\right)\\ &\leq \left(1+\frac{1}{8t_f(\epsilon)}\right)^{t_f'}\|\nabla f(x_{t'})\|\\ &\quad + 8t_f(\epsilon)\left(\left(1+\frac{1}{8t_f(\epsilon)}\right)^{t_f'} - 1\right) Lc_3\left(u^2 d^2\left(\log\frac{T}{\epsilon}\right)^2 + \left(1+\sqrt{\frac{\log(T/\epsilon)}{d}}\right) r\right)\end{aligned}$$

$$e^{1=4} \quad 8t_f(\ ) Lc_3 \quad u^2d^2 \quad \log\frac{T}{\ }^{2} + \quad \sqrt{1+\frac{\log(T= \ )}{d}}\, r \; !$$

$$+ 8t_f(\ )(e^{1=4} \quad 1) \quad Lc_3 \quad u^2d^2 \quad \log\frac{T}{\ }^{2} + \quad \sqrt{1+\frac{\log(T= \ )}{d}}\, r \; !$$

$$16t_f(\ ) Lc_3 \quad u^2d^2 \quad \log\frac{T}{\ }^{2} + \quad \sqrt{1+\frac{\log(T= \ )}{d}}\, r \; ! \; ;$$

where we used $t_f^0 \quad 2t_f(\ )$ and the fact that $(1 \quad 1=(8x))^{2x} \quad e^{1=4}$ for all $x > 0$. By de ning $c_5 := 16c_3$, we complete the proof. □

We next derive a useful result showing that the function change $f(x \ ) \quad f(x_0)$ can be decomposed into one component arising from intervals when the gradient dominates noise (which improves function value) and another component arising from intervals with small gradient which may add to function value but whose contributions are bounded in terms of $\ , u$ and r. For now, we focus on the case $\quad t_f(\ )$, since it will be useful to us in proving that there cannot be more than $T=4$ iterations with large gradient.

Lemma 18 (Function change for large). Let $c_1 > 0; c_4 > 0; c_5 > 0; C_1 \quad 1$ be the absolute constants de ned in the statements of the previous lemmas. Let $2(0; 1=e]$, and let $\quad t_f(\ ))$ be arbitrary. Consider splitting $f0; 1; : : : ; \quad 1g$ into $K := b \ = t_f(\ )c$ intervals:

$$J_k = f k t_f(\ ); : : : ; (k+1)t_f(\ ) \quad 1g; \quad 0 \quad k < K \quad 1;$$
$$J_{K \quad 1} = f(K \quad 1)t_f(\ ); : : : ; \quad 1g:$$

Let I_1 denote the set of indices k such that for every time-step t in the interval J_k, the gradient dominates the noise terms as

$$kr \ f(x_t)k > 8t_f(\ ) \ L \quad \frac{u}{2} \quad \frac{1}{m}\sum_{i=1}^{m} Z_{t;i}Z_{t;i}^{>} H_{t;i}Z_{t;i} \quad + kY_tk \; ! \; : \qquad (30)$$

Suppose we choose such that

$$\frac{1}{Lt_f(\ )} \quad \min \quad \frac{\sqrt{m}}{8c_4(\ln(C_1 dmT= \ ))^{3=2}\sqrt{d}}; \frac{m}{128c_1(\ln(C_1 dmT= \ ))^3 d} \quad : \qquad (31)$$

Then, on the event

$$E(\ ) := H(\ ) \setminus \bigcap_{t=0}^{\ 1} A_t(\ ) \; ! \setminus \bigcap_{t=0}^{\ 1} G_t(\ ) \; ! \setminus \bigcap_{k=0}^{K \ 2} B_{kt_f(\ )}(\ ; t_f(\ )) \; ! \setminus B_{(K \ 1)t_f(\ )}(\ ; \quad (K \quad 1)t_f(\ ));$$

we have the following upper bound on function value change:

$$f(x \ ) \quad f(x_0) \quad \sum_{k2I_1} \frac{\ }{2}\min_{t2J_k} kr \ f(x_t)k^2 + \frac{c_5^2}{64} \ {}^3 t_f(\ )^2 L^2 \quad u^2d^2 \quad \log\frac{T}{\ }^{2} + \quad \sqrt{2\log(T= \ )}\, r \; !^{2}$$

$$+ \quad u^4 \ {}^2 \quad c_1 d^3 \quad \log\frac{T}{\ }^{3} + \ L \ {}^2 u^4 \ {}^2 \quad c_1 d^4 \quad \log\frac{T}{\ }^{4}$$

$$+ \quad c_1 r^2 (128t_f(\ ) + \ L \ ) \log\frac{T}{\ } + \quad c_1 L \ {}^2 r^2 : \qquad (32)$$

Moreover, $P(E(\ )) \quad 1 \quad \frac{(5 \ +4)}{T}$.

Proof. Without loss of generality, we may assume that is a multiple of $t_f(\ )$.[5] Then, any interval $J_k = f t_0; : : : ; t_0 + t_f(\ ) \quad 1g$ belongs to one of the following two cases:

[5] To accommodate the last interval which has length at most $2t_f(\ ) \quad 1$, we note that the results we require for the proof, namely Lemma 14, Lemma 16 and Lemma 17, all hold for any interval length $t_f^0 \quad 2t_f(\ )$.

Case 1) (Gradient dominates noise): Recall that this means that for every $t\ 2\ J_k$, we have

$$kr\ f(x_t)k > 8t_f(\ )\ L\ \ \frac{u}{2}\ \frac{1}{m}\overset{m}{\underset{i=1}{X}} Z_{t;i}\ Z_{t;i}^{>}\ HT_{t;i}\ Z_{t;i}\ \ + kY_t k\ \ ! \ :$$

By our choice of in Eq. (31), we can apply Lemma 16 to get

$$\underset{t\,2\,J_k}{\min} kr\ f(x_t)k\quad \frac{1}{4}\underset{t\,2\,J_k}{\max} kr\ f(x_t)k:$$

We now consider the two cases when J has fewer than $t_f(\ )$ iterations and when $J = J_k$ f

Note also that on the event $B_{kt_f(\ )}(\ ;t_f(\ ))$, there exists some $t\ 2\ J_k$ such that

$$\frac{1}{m}\overset{m}{\underset{i=1}{X}} Z_{t;i}^{>}\ r\ f(x_t)\ ^2\quad \frac{1}{2}kr\ f(x_t)k^2:$$

This implies then that

$$\frac{1}{4}\underset{t\,2\,J_k}{X}\ \frac{1}{m}\overset{m}{\underset{i=1}{X}} Z_{t;i}^{>}\ r\ f(x_t)\ ^2\quad \frac{1}{4}\underset{t\,2\,J_k}{\min} kr\ f(x_t)k^2\quad \frac{1}{64}\underset{t\,2\,J_k}{\max} kr\ f(x_t)k^2$$
$$\frac{1}{64t_f(\ )}\underset{t\,2\,J_k}{X} kr\ f(x_t)k^2: \qquad (33)$$

Thus by setting $= 128t_f(\ )$ in Eq. (5) and by choosing such that

$$\frac{c_1 L\ ^2\ ^3 d}{m}\quad \_ = \frac{}{128t_f(\ )}\quad () \qquad \frac{m}{128c_1 L t_f(\ ) d\ ^3};$$

it follows that

$$\frac{3}{4}\underset{t\,2\,J_k}{X}\ \frac{1}{m}\overset{m}{\underset{i=1}{X}} Z_{t;i}^{>}\ r\ f(x_t)\ ^2 + \quad \frac{}{128t_f(\ )} + \frac{c_1 L\ ^2\ ^3 d}{m}\quad \underset{t\,2\,J_k}{X} kr\ f(x_t)k^2$$
$$= \quad \frac{3}{4}\underset{t\,2\,J_k}{X}\ \frac{1}{m}\overset{m}{\underset{i=1}{X}} Z_{t;i}^{>}\ r\ f(x_t)\ ^2 + \frac{}{64t_f(\ )}\underset{t\,2\,J_k}{X} kr\ f(x_t)k^2$$
$$\frac{}{2}\underset{t\,2\,J_k}{X}\ \frac{1}{m}\overset{m}{\underset{i=1}{X}} Z_{t;i}^{>}\ r\ f(x_t)\ ^2$$
$$\frac{}{2}\underset{t\,2\,J_k}{\min} kr\ f(x_t)k^2 \qquad (34)$$

Case 2) (Gradient does not dominate noise): there exists some $t\ 2\ J_k$ such that

$$kr\ f(x_t)k\quad 8t_f(\ )\ L\ \ \frac{u}{2}\ \frac{1}{m}\overset{m}{\underset{i=1}{X}} Z_{t;i}\ Z_{t;i}^{>}\ HT_{t;i}\ Z_{t;i}\ \ + kY_t k\ \ ! \ :$$

By our choice of in Eq. (31), we can apply Lemma 17 to get

$$kr\ f(x_t)k\quad c_5 t_f(\ )\ L\quad u^2 d^2\quad \log\frac{T}{\ }\ ^2 + \ r\ \overline{1 + \frac{\log(T=\ )}{d}}\ r\ !\qquad 8t\ 2\ J_k:$$

Hence, by setting $= 128t_f(\ )$ in Eq. (5) and choosing such that

$$\frac{c_1 L\ ^2\ ^3 d}{m}\quad \_ = \frac{}{128t_f(\ )};$$

it follows that

$$\frac{}{128 t_f(\ )} + \frac{c_1 L\ ^2\ ^3 d}{m} \underset{t\,2\,J_k}{X} \mathrm{kr}\ f(x_t)k^2$$

$$\frac{}{64 t_f(\ )} \underset{t\,2\,J_k}{X} c_5 t_f(\ )\ L \quad u^2 d^2 \quad \log\frac{T}{}\ ^2 + \quad r\,\overline{1+\frac{\log(T=\ )}{d}}\,r\ !!\ ^2$$

$$\frac{c_5^2}{64} t_f(\ )^2\ ^3L^2 \quad u^2 d^2 \quad \log\frac{T}{}\ ^2 + \quad r\,\overline{1+\frac{\log(T=\ )}{d}}\,r\ !\ ^2 \tag{35}$$

Without loss of generality, we may assume that is a multiple of $t_f(\ )$.[6] Then, any interval $J_k = f\,t_0; : : : ; t_0 + t_f(\ ) \quad 1g$ belongs to one of the following two cases:

Having studied the two cases, we may now proceed to use them to complete the proof. Let I_1^c denote the complement of I_1 in $f\,0; 1; : : : ; K \quad 1g$. Then,

$$\frac{3}{4}\overset{1}{\underset{t=0}{X}} \frac{1}{m}\underset{i=1}{X^m} Z^{>}_{t;i}\ r\ f(x_t)\ ^2 + \ \frac{}{} + \frac{c_1 L\ ^2\ ^3 d}{m} \overset{1}{\underset{t=0}{X}} \mathrm{kr}\ f(x_t)k^2$$

$$= \underset{k\,2\,I_1}{X} \quad \frac{3}{4}\underset{t=\,2\,J_k}{X} \frac{1}{m}\underset{i=1}{X^m} Z^{>}_{t;i}\ r\ f(x_t)\ ^2 + \quad \frac{}{128 t_f(\ )} + \frac{c_1 L\ ^2\ ^3 d}{m} \underset{t\,2\,J_k}{X} \mathrm{kr}\ f(x_t)k^2\ !$$

$$+ \underset{k\,2\,I_1^c}{X} \quad \frac{3}{4}\underset{t\,2\,J_k}{X} \frac{1}{m}\underset{i=1}{X^m} Z^{>}_{t;i}\ r\ f(x_t)\ ^2 + \quad \frac{}{128 t_f(\ )} + \frac{c_1 L\ ^2\ ^3 d}{m} \underset{t\,2\,J_k}{X} \mathrm{kr}\ f(x_t)k^2\ !$$

$$\underset{k\,2\,I_1}{X} \frac{}{2}\underset{t\,2\,J_k}{\min} \mathrm{kr}\ f(x_t)k^2 + \underset{k\,2\,I_1^c}{X} \quad t_f(\ )\ @\overset{0}{}\frac{c_5^2}{64} t_f(\ )^2\ ^3L^2 \quad u^2 d^2 \quad \log\frac{T}{}\ ^2 + \quad r\,\overline{1+\frac{\log(T=\ )}{d}}\,r\ !\ ^2\ \overset{1}{A}$$

$$\underset{k\,2\,I_1}{X} \frac{}{2}\underset{t\,2\,J_k}{\min} \mathrm{kr}\ f(x_t)k^2 + \frac{c_5^2}{64} t_f(\ )^2\ ^3L^2 \quad u^2 d^2 \quad \log\frac{T}{}\ ^2 + \quad r\,\overline{1+\frac{\log(T=\ )}{d}}\,r\ !\ ^2 \quad : \tag{36}$$

and so by Eq. (5),

$$f(x\ ) \quad f(x_0) \quad \underset{k\,2\,I_1}{X} \frac{}{2}\underset{t\,2\,J_k}{\min} \mathrm{kr}\ f(x_t)k^2 + \frac{c_5^2}{64} t_f(\ )^2\ ^3L^2 \quad u^2 d^2 \quad \log\frac{T}{}\ ^2 + \quad r\,\overline{1+\frac{\log(T=\ )}{d}}\,r\ !\ ^2$$

$$+ \quad u\ ^4\ ^2 \quad c_1 d^3 \quad \log\frac{T}{}\ ^3 + \quad L\ ^2 u^4\ ^2 \quad c_1 d^4 \quad \log\frac{T}{}\ ^4$$

$$+ \quad c_1 r^2(\quad + \quad L\ )\log\frac{T}{} + \quad c_1 L\ ^2 r^2 :$$

Note that we choose $= 128 t_f(\ )$. In addition, observe that by our choice of (such that $\frac{1}{e}$), it follows that $q\,\overline{1+\frac{\log(T=\ )}{d}} \quad p\,\overline{2\log(T=\ )}$.

We can now complete our proof by using the union bound (suppressing the dependence of some of the events on for notational simplicity) to derive

$$P(E^c) \quad P(H^c) + \overset{1}{\underset{t=0}{X}} P(A_t^c) + \overset{1}{\underset{t=0}{X}} P(G_t^c) + \overset{K\ 1}{\underset{k=0}{X}} P(B^c_{kt_f(\ )}(\ ; t_f(\ )))$$

$$\frac{(\ +4)}{T} + \frac{}{T} + 2\frac{}{T} + \frac{K}{T} \quad \frac{(5\ +4)}{T} :$$

□

[6] To accommodate the last interval which has length at most $2t_f(\ ) \quad 1$, we note that the results we require for the proof, namely Lemma 14, Lemma 16 and Lemma 17, all hold for any interval length $t_0 \quad 2t_f(\ )$.

We are now ready to show that if sufficiently many iterations have a large gradient, then with high probability, the function value of the last iterate $f(x_T)$, will be less than $\min_x f(x)$, a contradiction. Hence this limits the number of iterations that can have a large gradient.

Proposition 4. Let $c_1 > 0; c_2 \geq 1; c_4 > 0; c_5 > 0; C_1 \geq 1$ be the absolute constants defined in the statements of the previous lemmas, and let $\delta \in (0; 1/e]$ be arbitrary. Suppose we choose u, r, η and T such that

$$u \leq \frac{\sqrt{\epsilon}}{\sqrt{d}\sqrt{\rho}\log(T/\delta)}\min\left\{\frac{1}{64c_5^2c_2}; \frac{1}{2048c_1c_2}\right\}^{1/4}; \qquad r \leq \min\left\{\frac{1}{8c_5\sqrt{2c_2}}; \frac{1}{32\sqrt{c_1}}\right\}\epsilon;$$

$$\eta \leq \frac{1}{Lt_f(\epsilon)}\min\left\{\frac{1}{\log(T/\delta)}; \frac{\sqrt{m}}{8c_4(\mathrm{lr}(C_1dmT/\delta))^{3/2}\sqrt{d}}; \frac{m}{128c_1(\mathrm{lr}(C_1dmT/\delta))^3d}\right\};$$

$$T \geq \max\left\{\frac{256t_f(\epsilon)\,(f(x_0) - f^*) + \epsilon^2/L)}{\eta\epsilon^2}; 4\right\}:$$

Then, with probability at least $1 - 6\delta$, there are at most $T/4$ iterations for which $\|\nabla f(x_t)\| \geq \epsilon$.

Proof. Without loss of generality, we assume that T is a multiple of $t_f(\epsilon)$, and we similarly split $\{0; 1; \dots; T\}$ into $K = \lfloor T/t_f(\epsilon)\rfloor$ intervals $J_0; \dots; J_{K-1}$. Let I_1 denote the set of indices k such that for every $t \in J_k$,

$$\|\nabla f(x_t)\| > 8t_f(\epsilon)L\eta\left[\left\|\frac{u}{2}\left(\frac{1}{m}\sum_{i=1}^{m} Z_{t,i}Z_{t,i}^\top \tilde{H}_{t,i}Z_{t,i}\right)\right\| + \|Y_t\|\right]: \tag{37}$$

We let I_1^c denote the complement of I_1 in $\{0; 1; \dots; K-1\}$. We denote

$$\mathcal{E}_T(\delta) := \mathcal{H}_T(\delta) \cap \left(\bigcap_{t=0}^{T-1}\mathcal{A}_t(\delta)\right) \cap \left(\bigcap_{t=0}^{T-1}\mathcal{G}_t(\delta)\right) \cap \left(\bigcap_{k=0}^{K-1}\mathcal{B}_{kt_f(\epsilon)}(\delta; t_f(\epsilon))\right):$$

In the remaining part of the proof, unless otherwise stated, we shall always assume that we are working on the event $\mathcal{E}_T(\delta)$.

By Lemma 18 with $\tau = T$ and our choices of η and u in the statement of the lemma, we have

$$\begin{aligned} f(x_T) - f(x_0) \leq{} & \sum_{k\in I_1} -\frac{\eta}{2}\min_{t\in J_k}\|\nabla f(x_t)\|^2 + T\frac{c_5^2}{64}t_f(\epsilon)^2\eta^3L^2\left(\rho u^2d^2\left(\log\frac{T}{\delta}\right)^2 + \sqrt{2\log(T/\delta)}\,r\right)^2 \\ & + T\eta u^4\rho^2c_1d^3\left(\log\frac{T}{\delta}\right)^3 + TL\eta^2u^4\rho^2c_1d^4\left(\log\frac{T}{\delta}\right)^4 \\ & + \eta c_1r^2(128t_f(\epsilon) + \eta L)\log\frac{T}{\delta} + Tc_1L\eta^2r^2: \end{aligned} \tag{38}$$

Suppose that there are at least $T/4$ iterations where $\|\nabla f(x_t)\| \geq \epsilon$. Let I_ϵ denote the set of indices k for which there exists some $t \in J_k$ with $\|\nabla f(x_t)\| \geq \epsilon$. Then, by the pigeonhole principle, the set I_ϵ has at least $\lceil T/(4t_f(\epsilon))\rceil$ members. Note that, by our choices of the parameters $\eta; u; r$, it can be shown that

$$c_5t_f(\epsilon)L\eta\left(\rho u^2d^2\left(\log\frac{T}{\delta}\right)^2 + \sqrt{1 + \frac{\log(T/\delta)}{d}}\,r\right) < \epsilon; \tag{39}$$

while by Lemma 17, if k is in I_1^c, we have

$$\|\nabla f(x_t)\| \leq c_5t_f(\epsilon)L\eta\left(\rho u^2d^2\log(T/\delta) + \sqrt{1 + \frac{\log(T/\delta)}{d}}\,r\right); \qquad \forall t \in J_k:$$

This implies that $I_\epsilon \subseteq I_1$.

Observe that by Lemma 16, for any $k \in I_1$, we have

$$\frac{1}{2}\|\nabla f(x_{kt_f(\epsilon)})\| \leq \|\nabla f(x_t)\| \leq 2\|\nabla f(x_{kt_f(\epsilon)})\|; \qquad \forall t \in J_k:$$

This implies in particular that for any $k \; 2 \; I$, we have $\min_{t\,2\,J_k} kr \; f(x_t)k^2 \quad \frac{1}{16} \; ^2$, and consequently

$$\mathop{X}_{k\,2\,I_1} \frac{}{2} \min_{t\,2\,J_k} kr \; f(x_t)k^2 \quad \mathop{X}_{k\,2\,I} \frac{}{2} \; \frac{^2}{16} \quad \frac{T}{4t_f(\;)} \; \frac{}{2} \; \frac{^2}{16} = \frac{T \; ^2}{128t_f(\;)}:$$

Hence, by Eq. (38),

$$f(x_T) \quad f(x_0) \quad \frac{T \; ^2}{128t_f(\;)} + T\frac{c_5^2}{64} t_f(\;)^2 \; ^3L^2 \quad u^2d^2 \quad \log\frac{T}{}^{\;2} + \; p \; \frac{}{2\log(T=\;)r} \; !^{\,2}$$

$$+ \; T \; u^4 \; ^2 \quad c_1d^3 \quad \log\frac{T}{}^{\;3} + T \quad (\;L\;)u^4 \; ^2 \quad c_1d^4 \quad \log\frac{T}{}^{\;4}$$

$$+ \quad c_1r^2(128t_f(\;) + \;\; L\;)\log\frac{T}{} + T \quad c_1 \; Lr \; ^2: \tag{40}$$

Now, by our choices of u, r and , we have

$$T\frac{c_5^2}{64} t_f(\;)^2 \; ^3L^2 \quad u^2d^2 \quad \log\frac{T}{}^{\;2} + \; p \; \frac{}{2\log(T=\;)r} \; !^{\,2}$$

$$T \quad \frac{c_5^2}{32} t_f(\;)^2(\;L\;)^2 \quad u^4d^4 \; ^2 \quad \log\frac{T}{}^{\;4} + 2\log(T=\;)r^2 \; !$$

$$T \quad \frac{^2}{2048c_2 \;\; \log\frac{T}{}^{\;2}} + \frac{^2}{2048c_2\log(T=\;)} \; ! \quad \frac{T \; ^2}{512t_f(\;)};$$

where we used $\log(T=\;) \quad 1$ and $2c_2\log(T=\;) \quad t_f(\;)$. We also have

$$T \; u^4 \; ^2 \quad c_1d^3 \quad \log\frac{T}{}^{\;3} + T \quad (\;L\;)u^4 \; ^2 \quad c_1d^4 \quad \log\frac{T}{}^{\;4} + Tc_1L \; ^2r^2$$

$$T \quad \frac{^2}{2048c_2d\log(T=\;)} + T \quad \frac{^2}{2048c_2t_f(\;)\log(T=\;)} + T \quad \frac{^2}{1024t_f(\;)\log(T=\;)}$$

$$\frac{T \; ^2}{512t_f(\;)};$$

where we used $c_2d\log(T=\;) \quad t_f(\;)$, $c_2 \quad 1$ and $\log(T=\;) \quad 1$. Finally,

$$c_1r^2(128t_f(\;) + \;\; L\;)\log\frac{T}{} \quad \frac{(128t_f(\;)+1) \; ^2}{1024Lt_f(\;)} < \frac{^2}{L}:$$

By plugging these bounds into Eq. (40), we get

$$f(x_T) \quad f(x_0) < \quad \frac{T \; ^2}{128t_f(\;)} + \frac{T \; ^2}{512t_f(\;)} + \frac{T \; ^2}{512t_f(\;)} + \frac{^2}{L} \quad \frac{T \; ^2}{256t_f(\;)} + \frac{^2}{L}:$$

Therefore, as long as

$$T \quad \frac{256t_f(\;) \;\; (f(x_0) \quad f \;\;) + \; ^2=L)}{^2};$$

we will get $f(x_T) < f$, which is a contradiction. Thus, we can conclude that on the event $E_T(\;)$, there are at most $T=4$ iterations for which $kr \; f(x_t)k$.

We can now complete our proof by using the union bound (suppressing the dependence of some of the events on for notational simplicity) to derive

$$P(E_T^c) \quad P(H_T^c) + \mathop{X}_{t=0}^{T\;\;1} P(A_t^c) + \mathop{X}_{t=0}^{T\;\;1} P(G_t^c) + \mathop{X}_{k=0}^{K\;\;1} P(B^c_{kt_f(\;)}(\;;t_f(\;)))$$

$$\frac{(T+4)}{T} + \quad + 2 \quad + \frac{K}{T} \quad 6: \qquad \square$$

E. Escaping saddle point

In this section, we first show that the travelling distance of the iterates can be bounded in terms of the function value improvement (Appendix E.2). Utilizing this result, as well as Proposition 2 in Appendix C.3 which provides a concentration bound on the the zeroth-order noise, we then prove that sufficient function value decrease can be made near a saddle point in Appendix E.3.

E.1. Key quantities and notation

We will use to denote $_{\min}(r\ ^2 f(x_0))$, where we know that $\quad p\ \overline{\ }\ :$

E.2. Improve or Localize

In this subsection, we aim to bound the movement of the iterates across a number of steps in terms of the function value improvement made during these number of steps.

We first state a simple result separating the norm of the difference between x_{t_0+} and x_{t_0} into a few different terms.

Lemma 19. Consider the perturbed zeroth-order update Algorithm 1. Then, for any t_0 2 N and 2 N,

$$k x_{t_0+} \quad x_{t_0} k^2 \quad V_1(t_0;\ ) + V_2(t_0;\ ) + V_3(t_0;\ ) + V_4(t_0;\ ); \tag{41}$$

where

$$\begin{aligned} V_1(t_0;\ ) &:= 8\ ^2 \textstyle P_{t=t_0}^{t_0+\ \ 1} k r\ f(x_t) k^2; \quad V_2(t_0;\ ) := 8\ ^2 \textstyle P_{t=t_0}^{t_0+\ \ 1} \frac{1}{m} P_{i=1}^{m} (Z_{t;i} Z_{t;i}^{>} \quad I) r\ f(x_t) \ ^2 \\ V_3(t_0;\ ) &:= 4\ ^2 \textstyle P_{t=t_0}^{t_0+\ \ 1} Y_t \ ^2; \qquad V_4(t_0;\ ) := 4\ ^2 \textstyle P_{t=t_0}^{t_0+\ \ 1} \frac{1}{m} P_{i=1}^{m} u Z_{t;i} Z_{t;i}^{>} \bar{H}_{t;i} Z_{t;i} \ ^2 : \end{aligned} \tag{42}$$

Proof. For notational convenience, let $t_0 := 0$. Then, applying the form of the perturbed zeroth-order update in Algorithm 1, we get

$$\begin{aligned} &k x \quad x_0 k^2 \\ &= \ X_{t=0}^{\ 1} x_{t+1} \quad x_t \ ^2 \\ &= \ ^2 \ X_{t=0}^{\ 1} \frac{1}{m} X_{i=1}^{m} Z_{t;i} Z_{t;i}^{>} r\ f(x_t) + \frac{1}{m} X_{i=1}^{m} u Z_{t;i} Z_{t;i}^{>} \bar{H}_{t;i} Z_{t;i} \ + Y_t \ ^2 \\ &\quad 4\ ^2 \ X_{t=0}^{\ 1} \frac{1}{m} X_{i=1}^{m} Z_{t;i} Z_{t;i}^{>} r\ f(x_t) \ ^2 + 4\ ^2 \ X_{t=0}^{\ 1} \frac{1}{m} X_{i=1}^{m} u Z_{t;i} Z_{t;i}^{>} \bar{H}_{t;i} Z_{t;i} \ ^2 + 4\ ^2 \ X_{t=0}^{\ 1} Y_t \ ^2 \\ &\quad 4\ ^2 \ X_{t=0}^{\ 1} \frac{1}{m} X_{i=1}^{m} (Z_{t;i} Z_{t;i}^{>} \quad I) r\ f(x_t) + X_{t=0}^{\ 1} r\ f(x_t) \ ^2 + 4\ ^2 \ X_{t=0}^{\ 1} \frac{1}{m} X_{i=1}^{m} u Z_{t;i} Z_{t;i}^{>} \bar{H}_{t;i} Z_{t;i} \ ^2 + 4\ ^2 \ X_{t=0}^{\ 1} Y_t \ ^2 \\ &\quad \underbrace{8\ ^2 \ X_{t=0}^{\ 1} k r\ f(x_t) k^2}_{V_1(0;\ )} + \underbrace{8\ ^2 \ X_{t=0}^{\ 1} \frac{1}{m} X_{i=1}^{m} (Z_{t;i} Z_{t;i}^{>} \quad I) r\ f(x_t) \ ^2}_{V_2(0;\ )} + \underbrace{4\ ^2 \ X_{t=0}^{\ 1} Y_t \ ^2}_{V_3(0;\ )} + \underbrace{4\ ^2 \ X_{t=0}^{\ 1} \frac{1}{m} X_{i=1}^{m} u Z_{t;i} Z_{t;i}^{>} \bar{H}_{t;i} Z_{t;i} \ }_{V_4(0;\ )} : \end{aligned}$$

□

We now proceed to bound the terms $V_1(t_0;\ ); V_2(t_0;\ ); V_3(t_0;\ )$ and $V_4(t_0;\ )$.

First, we have the following result bounding $V_1(t_0;\ )$.

Lemma 20. Let $c_1 > 0; c_2 \quad 1; c_4 > 0; c_5 > 0; C_1 \quad 1$ be the absolute constants defined in the statements of the previous lemmas, and let 2 $(0; 1{=}e]$ be arbitrary.

Suppose we choose η such that

$$\eta \le \frac{1}{Lt_f(\epsilon)}\min\left(\frac{\sqrt{m}}{8c_4(\ln(C_1dmT/\delta))^{3/2}\sqrt{d}}, \frac{m}{128c_1(\ln(C_1dmT/\delta))^3 d}\right).$$

There are two cases to consider.

1. The first is when $\tau \ge t_f(\epsilon)$. In this case, split $\{t_0, t_0+1, \dots, t_0+\tau-1\}$ into $K := \lfloor \tau/t_f(\epsilon)\rfloor$ intervals:

$$J_k = \{t_0 + kt_f(\epsilon), \dots, t_0 + (k+1)t_f(\epsilon) - 1\}, \quad 0 \le k < K-1;$$
$$J_{K-1} = \{t_0 + (K-1)t_f(\epsilon), \dots, t_0 + \tau - 1\}.$$

Then, on the event

$$E_{t_0,\tau}(\epsilon) := H_{t_0,\tau}(\delta) \cap \left(\bigcap_{t=t_0}^{t_0+\tau-1} A_t(\delta)\right) \cap \left(\bigcap_{t=t_0}^{t_0+\tau-1} G_t(\delta)\right) \cap \left(\bigcap_{k=0}^{K-2} B_{t_0+kt_f(\epsilon)}(\delta, t_f(\epsilon))\right) \cap B_{t_0+(K-1)t_f(\epsilon)}(\delta, \tau-(K-1)t_f(\epsilon)),$$

we have that

$$\begin{aligned} V_1(t_0,\tau) &= 8\eta^2\sum_{t=t_0}^{t_0+\tau-1}\|\nabla f(x_t)\|^2 \\ &\le 64\eta t_f(\epsilon)\left((f(x_0) - f(x_\tau)) + N_{u,r}(\tau,\delta)\right), \end{aligned}$$

where

$$\begin{aligned} N_{u,r}(\tau,\delta) := {} & \frac{c_5^2}{64}\eta^3 t_f(\epsilon)^2L^2\tau\left(u^2d^2\left(\log\frac{T}{\delta}\right)^2 + \left(\sqrt{2\log(T/\delta)}\,r\right)^2\right) \\ & + \tau\eta u^4\rho^2c_1d^3\left(\log\frac{T}{\delta}\right)^3 + \tau L\eta^2u^4\rho^2c_1d^4\left(\log\frac{T}{\delta}\right)^4 \\ & + \eta c_1r^2(128\eta t_f(\epsilon) + \tau\eta L)\log\frac{T}{\delta} + \tau c_1L\eta^2r^2 \\ & + c_5^2t_f^3(\epsilon)\eta^3L^2\left(u^2d^2\log(T/\delta) + \left(\sqrt{2\log(T/\delta)}\,r\right)^2\right). \end{aligned} \tag{43}$$

2. The second is when $\tau < t_f(\epsilon)$. Suppose we choose u and r such that

$$u \le \frac{\sqrt{\epsilon}}{d\sqrt{\log(T/\delta)}}\min\left(\frac{1}{64c_5^2c_2}, \frac{1}{2048c_1c_2}\right)^{1/4}, \qquad r \le \epsilon\min\left(\frac{1}{8c_5\sqrt{2c_2}}, \frac{1}{32\sqrt{c_1}}\right).$$

Suppose the event $\bigcap_{t=t_0}^{t_0+\tau-1}(A_t(\delta) \cap G_t(\delta))$ holds. Suppose also that $\|\nabla f(x_{t_0})\| \le \epsilon$. Then,

$$V_1(t_0,\tau) \le 32\eta^2\tau^2\epsilon^2 \le 32\eta^2(t_f(\epsilon))^2\epsilon^2.$$

Proof. 1. We first consider the case where $\tau \ge t_f(\epsilon)$. Let I_1 denote the set of indices k such that for every time-step t in the interval J_k, the gradient dominates the noise terms as

$$\|\nabla f(x_t)\| > 8t_f(\epsilon)L\eta\left(\frac{u}{2}\left\|\frac{1}{m}\sum_{i=1}^m Z_{t,i}Z_{t,i}^\top \tilde{H}_{t,i}Z_{t,i}\right\| + \|Y_t\|\right). \tag{44}$$

WLOG, we may assume that $t_0 := 0$, and denote $V_1(\tau) := V_1(0,\tau)$. WLOG, we also assume that τ is a multiple of $t_f(\epsilon)$. From Lemma 18, on the event that $E(\epsilon)$ holds and by our choice of η, we have

$$f(x_\tau) - f(x_0) \le -\sum_{k\in I_1}\frac{\eta}{2}\min_{t\in J_k}\|\nabla f(x_t)\|^2 + \frac{c_5^2}{64}\eta^3t_f(\epsilon)^2L^2\tau\left(u^2d^2\left(\log\frac{T}{\delta}\right)^2 + \left(\sqrt{2\log(T/\delta)}\,r\right)^2\right)$$

$$+ \quad u^{4\ 2}\ c_1 d^3 \ \ \log \frac{T}{}\ ^{3} \ + \ L \ ^2 u^{4\ 2}\ c_1 d^4 \ \ \log \frac{T}{}\ ^{4}$$

$$+ \ c_1 r^2 (128 t_f(\ ) + \ L) \log \frac{T}{} + \ c_1 L \ ^2 r^2 :$$

By Lemma 16 (and our choice of), it follows that for any $k\ 2\ I_1$, on the event $\ _{t 2 J_k} A_t(\ )$, we have

$$\sum_{t 2 J_k} kr\ f(x_t)k^2 \quad 4 t_f \min_{t 2 J_k} kr\ f(x_t)k^2 :$$

Thus, on the event that $E(\ )$ holds, for our choice of , we have

$$\sum_{k 2 I_1} \sum_{t 2 J_k} kr\ f(x_t)k^2 \quad 4 t_f(\ ) \sum_{k 2 I_1} \min_{t 2 J_k} kr\ f(x_t)k^2$$

$$8 t_f(\ ) \sum_{k 2 I_1} \frac{}{2} \min_{t 2 J_k} kr\ f(x_t)k^2$$

$$8 t_f(\ ) @ (f(x_0) \quad f(x\ )) + \frac{c_5^2}{64}\ ^3 t_f(\ )^2 L^2 \quad u^2 d^2 \quad \log \frac{T}{}\ ^2 + \ p\ \overline{2 \log(T = \ ) r} \ \ !\ _2 1 \ A$$

$$+ \ 8 t_f(\ ) \quad u^{4\ 2}\ c_1 d^3 \ \ \log\frac{T}{}\ ^3 \ + \ L \ ^2 u^{4\ 2}\ c_1 d^4 \ \ \log\frac{T}{}\ ^4 !$$

$$+ \ 8 t_f(\ ) \quad c_1 r^2 (128 t_f(\ ) + \ L) \log\frac{T}{} + \ c_1 L \ ^2 r^2 \quad :$$

Similarly, for any $k\ 2\ I_1^c$ (where I_1^c denotes the complement of I_1 in $f 0; 1; :::; K \quad 1 g$, i.e. intervals where the gradient is smaller than than the perturbation terms in some iteration), on the event $(\ _{t 2 J_k} A_t(\ )) \setminus (\setminus_{t 2 J_k} G_t(\ ))$, by Lemma 17 (and our choice of), we have

$$kr\ f(x_t)k \quad c_5 t_f(\ )\ L \quad u^2 d^2 \ \ \log(T = \ ) + \ p\ \overline{2 \log(T = \ ) r} \quad ; \qquad 8 t\ 2\ J_k :$$

On the event that $E(\ )$ holds, this gives us then

$$\sum_{k 2 I_1^c} \sum_{t 2 J_k} kr\ f(x_t)k^2 \qquad c_5^2 t_f^2(\ ) \ ^2 L^2 \quad u^2 d^2 \ \ \log(T = \ ) + \ p\ \overline{2 \log(T = \ ) r} \ \ ^2 \quad :$$

Hence, on the event that $E(\ )$ holds, we have that

$$\sum_{t=0}^{1} kr\ f(x_t)k^2 = \sum_{k 2 I_1} \sum_{t 2 J_k} kr\ f(x_t)k^2 + \sum_{k 2 I_1^c} \sum_{t 2 J_k} kr\ f(x_t)k^2$$

$$8 t_f(\ ) @ (f(x_0) \quad f(x\ )) + \frac{c_5^2}{64}\ ^3 t_f(\ )^2 L^2 \quad u^2 d^2 \quad \log \frac{T}{}\ ^2 + \ p\ \overline{2 \log(T = \ ) r} \ \ !\ _2 1 \ A$$

$$+ \ 8 t_f(\ ) \quad u^{4\ 2}\ c_1 d^3 \ \ \log\frac{T}{}\ ^3 \ + \ L \ ^2 u^{4\ 2}\ c_1 d^4 \ \ \log\frac{T}{}\ ^4 !$$

$$+ \ 8 t_f(\ ) \quad c_1 r^2 (128 t_f(\ ) + \ L) \log\frac{T}{} + \ c_1 L \ ^2 r^2$$

$$+ \ 8 t_f(\ ) \qquad c_5^2 t_f^2(\ ) \ ^2 L^2 \quad u^2 d^2 \ \ \log(T = \ ) + \ p\ \overline{2 \log(T = \ ) r} \ \ ^2 \quad :$$

This yields the final result for the case $\quad t_f(\ )$.

2. We next consider the case where $1 \ < t_f(\ )$. Recall the notation that

$$E(t_0; t_0 + \ ; \ ) := \bigcap_{t=t_0}^{t_0+ \ 1} \left(\| r \ f(x_t)\| > 8 t_f(\ ) L \ \frac{u}{2} \ \frac{1}{m}\sum_{i=1}^{m} Z_{t;i} Z_{t;i}^{>} H_{t;i} Z_{t;i} \ + \|Y_t\| \right)$$

There are two cases to consider.

(a) On the event $E(t_0; t_0 + \ ; \ ) \cap \ \cap_{t=t_0}^{t_0+ \ 1} A_t(\ )$; we have by Lemma 16 that $\| r \ f(x_t)\| \ 2\| r \ f(x_0)\|$ for each $t \ 2 \ f\ 0; 1; :::; \ 1g$. Then,

$$V_1(t_0; \ ) = 8 \ ^2 \sum_{t=t_0}^{t_0+ \ 1} \| r \ f(x_t)\|^2 \quad 8 \ ^2 \ ^2 \ 4\| r \ f(x_0)\|^2 \quad 32 \ ^2 \ ^2 \ ^2;$$

where the final inequality uses the assumption that $\| r \ f(x_0)\| \ \ :$

(b) Suppose the event $E^c(t_0; t_0 + \ ; \ ) \cap \ \cap_{t=t_0}^{t_0+ \ 1} A_t(\ ) \ \cap \ \cap_{t=t_0}^{t_0+ \ 1} G_t(\ )$ holds. In this case, by Lemma 17, we have that for each $t \ 2 \ f\ t_0; t_0 + 1; :::; t_0 + \ \ 1g$

$$\| r \ f(x_t)\| \quad c_5 t_f(\ ) L \quad u^2 d^2 \quad \log \frac{T}{\ } \ ^2 + \ \sqrt{1 + \frac{\log(T= \ )}{d}} \ r \ !$$

$$;$$

where the final inequality follows by our choice of u and r (cf. Eq. (39)). Hence,

$$\begin{aligned} V_1(t_0; \ ) &= 8 \ ^2 \sum_{t=t_0}^{t_0+ \ 1} \| r \ f(x_t)\|^2 \\ & \quad 8 \ ^2 \ ^2 \ \left(c_5 t_f(\ ) L \quad u^2 d^2 \quad \log \frac{T}{\ } \ ^2 + \ \sqrt{1 + \frac{\log(T= \ )}{d}} \ r \right) !! \ ^2 \\ & \quad 8 \ ^2 \ ^2 \ ^2 < 32 \ ^2 \ ^2 \ ^2 : \end{aligned}$$

The final result for the case $\ < t_f(\ )$ then follows.

□

We proceed to bound $V_2(t_0; \ )$.

Lemma 21. Let $c_1 > 0; c_2 \ 1; c_4 > 0; c_5 > 0; C_1 \ 1$ be the absolute constants defined in the statements of the previous lemmas, and let $\ 2 \ (0; 1=e]$ be arbitrary and $\ > \ 0$ be arbitrary. Suppose we choose $\ $ such that

$$\frac{1}{L t_f(\ )} \ \min \ \frac{p \overline{m}}{8 c_4 (\ln(C_1 dmT= \ ))^{3=2} p \overline{d}}; \frac{m}{128 c_1 (\ln(C_1 dmT= \ ))^3 d} \ :$$

Let T_s denote an integer such that $T_s \ \max f \ ; t_f(\ ) g$, and for any $F > 0$, define

$$B(\ ; F) := \frac{8 t_f(\ )(F + N_{u;r}(T_s; \ ))}{\ } \ T_s + \frac{d}{m} \ (\ln(CT^2= \ ))^2; \qquad b(\ ; F) := \frac{t_f(\ ) F}{\ }:$$

Let $c^0; C > 0$ denote the same constants as in the statement of Proposition 2. Denote the event that

$$\text{either} \quad \sum_{t=t_0}^{t_0+ \ 1} \frac{d}{m} (\ln(CT^2= \ ))^2 \| r \ f(x_t)\|^2 \quad B(\ ; F)$$

$$\text{or} \quad \sqrt{\frac{V_2(t_0; \ )}{8 \ ^2}} \quad c^0 \sqrt{\max \left(\sum_{t=t_0}^{t_0+ \ 1} \frac{d}{m} (\ln(CT^2= \ ))^2 \| r \ f(x_t)\|^2; b(\ ; F) \right) \ \log \frac{CT^2}{\ } + \log \ \log \frac{B(\ ; F)}{b(\ ; F)} \ + 1}$$

holds as $L_{t_0;}(\ ;F)$[7]. We show that $P(L_{t_0;}(\ ;F)) \quad 1 \quad \frac{}{T}$: Finally, denote the event $M_{t_0;T_s}(F)$ as the event that $f(x_{t_0}) \quad f(x_{t_0+T_s}) < F$.

Then, on the event $L_{t_0;}(\ ) \cap E_{t_0;T_s}(\ ) \cap M_{t_0;T_s}(F)$ (where $E_{0;T_s}(\ )$ is as defined in Lemma 20),

$$V_2(t_0;\ ) \quad 8c^{00}\ _1(\ ;F)\ t_f(\ )\max \quad \frac{8d}{m}(\mathrm{lr}(CT^2=\ ))^2(F + N_{u;r}(T_s;\ ))\ ;\ F \quad ; \tag{45}$$

where

$$_1(\ ;F) := \log \quad \frac{CT^2}{} \quad + \log \quad \log \quad \frac{B(\ ;F)}{b_1(\ ;F)} \quad +1 \quad :$$

Proof. We note that $P(L_{t_0;}(\ ;F)) \quad 1 \quad \frac{}{T}$: is a direct consequence of Proposition 2. In the rest of the proof, without loss of generality, we assume that $t_0 = 0$ for notational simplicity. On the event $L_{0;}(\ ;F) \cap E_{0;T_s}(\ ) \cap M_{t_0;T_s}(F)$, suppose that

$$\begin{aligned}
&\sum_{t=0}^{1} \frac{d}{m}(\mathrm{lr}(CT^2=\ ))^2 \mathrm{kr}\ f(x_t)\mathrm{k}^2 \quad B(\ ;F) = \frac{8t_f(\ )(F + N_{u;r}(T_s;\ ))}{} \quad T_s + \frac{d}{m} \quad (\mathrm{lr}(CT^2=\ ))^2 \\
=)\ & \sum_{t=0}^{1} \mathrm{kr}\ f(x_t)\mathrm{k}^2 \quad 8t_f(\ )(F + N_{u;r}(T_s;\ )) \\
=)\ & \sum_{t=0}^{T_s\ 1} \mathrm{kr}\ f(x_t)\mathrm{k}^2 \quad 8t_f(\ )(F + N_{u;r}(T_s;\ )) \\
=)\ & 8\ ^2T_s \sum_{t=0}^{T_s\ 1} \mathrm{kr}\ f(x_t)\mathrm{k}^2 \quad 64T_s t_f(\ )(F + N_{u;r}(T_s;\ )) \\
=)\ & 8\ ^2T_s \sum_{t=0}^{T_s\ 1} \mathrm{kr}\ f(x_t)\mathrm{k}^2 \quad 64T_s t_f(\ )(f(x_0) \quad f(x_{T_s}) + N_{u;r}(T_s;\ ));\quad \text{since} f(x_0) \quad f(x_{T_s}) \quad F \\
() \ & V_1(0;T_s) \quad 64T_s t_f(\ )(f(x_0) \quad f(x_{T_s}) + N_{u;r}(T_s;\ ));
\end{aligned}$$

where we note the last equation contradicts Lemma 20. For notational simplicity, denote

$$(\ ;F) := \log \quad \frac{CT^2}{} \quad + \log \quad \log \quad \frac{B(\ ;F)}{b(\ ;F)} \quad +1 \quad :$$

Observe that $_1$ is larger than for every 1. Since $L_{t_0;}(\ ;F)$ holds, we must have then that

$$\sqrt{\frac{V_2(0;\ )}{8\ ^2}} \quad c^{0}\sqrt{\max\left\{\sum_{t=0}^{1}\frac{d}{m}(\mathrm{lr}(CT^2=\ ))^2\mathrm{kr}\ f(x_t)\mathrm{k}^2; b(\ ;F)\right\}} \quad _1(\ ;F):$$

Now, continuing, recalling the definition of $V_1(0;T_s) = 8\ ^2T_s \sum_{t=0}^{T_s\ 1}\mathrm{kr}\ f(x_t)\mathrm{k}^2$

$$\begin{aligned}
V_2(0;\ ) \quad & c^{00}\ _1(\ ;F)\max\left\{8\ ^2\sum_{t=0}^{1}\frac{d}{m}(\mathrm{lr}(CT^2=\ ))^2\mathrm{kr}\ f(x_t)\mathrm{k}^2; 8\ ^2b(\ ;F)\right\} \\
& c^{00}\ _1(\ ;F)\max\left\{8\ ^2\sum_{t=0}^{T_s\ 1}\frac{d}{m}(\mathrm{lr}(CT^2=\ ))^2\mathrm{kr}\ f(x_t)\mathrm{k}^2; 8\ ^2b(\ ;F)\right\} \\
& c^{00}\ _1(\ ;F)\max \quad \frac{d}{m}(\mathrm{lr}(CT^2=\ ))^2\frac{V_1(0;T_s)}{T_s}; 8t_f(\ )\ F
\end{aligned}$$

[7] We note that by construction, $B(\ ;F) \quad b(\ ;F)$

$$\overset{(i)}{}c^{œ2}\ {}_1(\ ;F)\max\ \frac{d}{m}(\mathrm{lr}(\ CT^2=\ ))^2(64\ t_f(\ )(f(x_0)\quad f(x_{T_s})+N_{u;r}(T_s;\ )))\ ;8\ t_f(\ )\ F$$

$$\overset{(ii)}{}c^{œ2}\ {}_1(\ ;F)\max\ \frac{d}{m}(\mathrm{lr}(\ CT^2=\ ))^2(64\ t_f(\ )(F+N_{u;r}(T_s;\ )))\ ;8\ t_f(\ )\ F$$

$$=c^{œ2}\ {}_1(\ ;F)(8\ t_f(\ ))\max\ \frac{d}{m}(\mathrm{lr}(\ CT^2=\ ))^2(8(F+N_{u;r}(T_s;\ )))\ ;\ F\quad :$$

We note that (i) is a consequence of Lemma 20, while (ii) comes from our assumption that the event $M_{t_0;T_s}(F)$ holds, i.e. $f(x_{t_0})\quad f(x_{t_0+T_s})\quad F$.

□

We next bound $V_3(t_0;\ )$ and $V_4(t_0;\ )$.

Lemma 22. Let $c>0$ denote the same constant in Lemma 7. Consider any arbitrary $0<\quad 1=e;$ and let $\quad t_f(\ )$ be arbitrary. Let $N_{t_0;}(\ )$ denote the event that

$$V_3(t_0;\ ) := 4\ ^2 \sum_{t=t_0}^{t_0+\ \ 1} Y_t\ ^2\quad 4c_6\ ^2\ \log(2dT=\ )r^2;$$

where $c_6>0$ is an absolute constant. Then, by Lemma 7, $P(N_{t_0;}(\ ))\quad 1\quad \overline{T}$: Denote the event

$$O_t(\ ) := \left(\frac{1}{m}\sum_{i=1}^{m} kZ_{t;i}k^8\quad c_7 d^4\quad \log\ \frac{T}{—}\ ^4\right);$$

where $c_7>0$ is an absolute constant. Then, on the event $\bigcap_{t=t_0}^{t_0+\ \ 1}O_t(\ )$, we have

$$V_4(t_0;\ )\quad 4c_7\ ^2\ ^2\ ^2u^4d^4\quad \log\ \frac{T}{—}\ ^4:$$

Moreover, for each t, $P(O_t(\ ))\quad 1\quad \overline{T}$.

Proof. The proof for $V_3(t_0;\ )$ follows directly from Lemma 7, by picking c_6 to be the c that appears in the statement of Lemma 7. Meanwhile, observe that

$$V_4(t_0;\ ) = 4\ ^2\ \sum_{t=0}^{\ 1}\frac{1}{m}\sum_{i=1}^{m} uZ_{t;i}Z_{t;i}^{>}H_{t;i}\,Z_{t;i}\ ^2$$

$$4\ ^2\ @\sum_{t=t_0}^{t_0+\ \ 1}\frac{1}{m}\sum_{i=1}^{m} uZ_{t;i}Z_{t;i}^{>}H_{t;i}\,Z_{t;i}\ ^2 A$$

$$\overset{(iii)}{}4\ ^2\sum_{t=t_0}^{t_0+\ \ 1}\frac{1}{m}\sum_{i=1}^{m}\ ^2u^4kZ_{t;i}k^8$$

$$4c_7\ ^2\ ^2\ ^2u^4d^4\quad \log\ \frac{T}{—}\ ^4:$$

Above, to derive (iii), we used the bound that $H_{t;i}\quad u\ kZ_{t;i}k$. The final inequality is a consequence of our assumption that $\bigcap_{t=t_0}^{t_0+\ \ 1}O_t(\ )$ holds. Finally, the result that $P(O_t(\ ))\quad 1\quad \overline{T}$ holds due to Lemma 11, where we note that we may pick the absolute constant c_7 to be equal to $2Cc^4$, where $c;C>0$ are the absolute constants that appear in the statement of Lemma 11. □

Finally, combining the earlier results, we have the following technical result, which bounds the travelling distance of the iterates in terms of the decrease in function value decrease.

Lemma 23 (Improve or Localize) Consider the perturbed zeroth-order update Algorithm 1. Let $c^0 > 0; c_1 > 0; c_2 \quad 1; c_4 > 0; c_5 > 0; c_6 > 0; c_7 > 0; C_1 \quad 1$ be the absolute constants defined in the statements of the previous lemmas, and let $\ \in (0;1=e]$ be arbitrary. Consider any $T_s \quad t_f(\ )$. For any $F > 0$, suppose $f(x_{T_s}) \quad f(x_0) > \quad F$; i.e. $f(x_0) \quad f(x_{T_s}) < F$. Suppose that the event

$$P_{t_0;T_s}(\ ;F) := \cap_{=1}^{T_s} (L_{t_0;}(\ ;F)\cap N_{t_0;}(\ )) \cap \quad \cap_{t=t_0}^{t_0+T_s \quad 1} O_t(\ )\cap A_t(\ )\cap G_t(\ ) \quad \cap \quad \cap_{=t_f(\ )}^{T_s \quad 1} E_{t_0;}(\ )$$

holds, where the events $E_{t_0;}(\ ); L_{t_0;}(\ ); N_{t_0;}(\ ); O_t(\ )$ are as defined in Lemma 20, Lemma 21 and Lemma 22, and $G_t(\ )$ and $A_t(\ )$ are as defined in Lemma 13 and Lemma 15. Suppose we choose $u, \ $ and $\ $ such that

$$u \quad \frac{p\_}{d^{p\_}\log(T=\ )} \quad \min \quad \frac{1}{64c_5^2c_2}; \frac{1}{2048c_1c_2} \quad ^{1=4}; \qquad r \quad \min \quad \frac{1}{8c_5 \ p\overline{2c_2}}; \frac{1}{32 \ p\overline{c_1}} \quad ;$$

$$\frac{1}{Lt_f(\ )} \min \quad \frac{1}{\log(T=\ )}; \frac{p\overline{m}}{8c_4(\mathrm{lr}(C_1dmT=\ ))^{3=2}\ p\overline{d}}; \frac{m}{128c_1(\mathrm{lr}(C_1dmT=\ ))^3 d} \quad :$$

Suppose $\quad \min \ \{1; \frac{1}{t_f(\ )}; \frac{1}{t_f \ L}\}$. Suppose also we pick u and r small enough such that

$$u \quad \frac{r^{1=2}}{d\log(T=\ )^{1=2}}; \quad r^2 \quad \min \left\{ \frac{F}{T_s\log(T=\ )\ \frac{65c_5^2}{8} + 132c_1 + 1}; \frac{F}{4c_6\log(2dT=\ ) + 4c_7 \ T_s}; \right\} :$$

Then, for each $\ \in \{0;1;:::;T_s\}$, we have that

$$\|x_{t_0+} \quad x_{t_0}\|^2 \quad _{T_s}(\ ;F);$$

where

$$_{T_s}(\ ;F) \quad \max \ 128T_st_f(\ )F; 32 \ ^2(t_f(\ ))^2 \ ^2 \ + 8c^{02} \ _1(\ ;F) \ t_f(\ )\max \ \frac{16d}{m}(\mathrm{lr}(CT^2=\ ))^2F; T_sF \ + T_s \ t_f(\ )F;$$

where $_1(\ ;F)$ is defined as in Lemma 21. Moreover, $P(P_{t_0;T_s}(\ ;F)) \quad 1 \quad \frac{12T_s}{T}$.

Proof. We recall that

$$\|x_{t_0+} \quad x_{t_0}\|^2 \quad \underbrace{8 \ ^2 \quad \sum_{t=t_0}^{t_0+ \ 1} \|r \ f(x_t)\|^2}_{V_1(t_0;\ )} + \underbrace{8 \ ^2 \quad \sum_{t=t_0}^{t_0+ \ 1} \frac{1}{m}\sum_{i=1}^m (Z_{t;i}Z_{t;i}^> \quad I)r \ f(x_t) \ ^2}_{V_2(t_0;\ )}$$
$$+ \underbrace{4 \ ^2 \quad \sum_{t=t_0}^{t_0+ \ 1} Y_t \ ^2}_{V_3(t_0;\ )} + \underbrace{4 \ ^2 \quad \sum_{t=t_0}^{t_0+ \ 1} \frac{1}{m}\sum_{i=1}^m uZ_{t;i} \ Z_{t;i}^> H_{t;i} Z_{t;i} \ ^2}_{V_4(t_0;\ )}$$

By Lemma 20, Lemma 21, and Lemma 22, which bound $V_1(t_0;\ )$, $V_2(t_0;\ )$, and $V_3(t_0;\ ); V_4(t_0;\ )$ respectively, on the event $P_{t_0;T_s}(\ ;F)$, we have, for any $0 \quad \quad T_s$,

$$\begin{aligned}\|x \quad x_0\|^2 \quad & V_1(0;\ ) + V_2(0;\ ) + V_3(0;\ ) + V_4(0;\ ) \\ & \max \ 64 \ t_f(\ )(F + N_{u;r}(\ ;\ )); 32 \ ^2(t_f(\ ))^2 \ ^2 \\ & + 8c^{02} \ _1(\ ;F) \ t_f(\ )\max \ \frac{8d}{m}(\mathrm{lr}(CT^2=\ ))^2(F + N_{u;r}(T_s;\ )) ; \ F \\ & + 4c_6 \ ^2 \ \log(2dT=\ )r^2 + 4c_7 \ ^2 \ ^2 \ ^2u^4d^4(\log(T=\ ))^4 ;\end{aligned}$$

where $N_{u;r}(\ ;\ )$ is defined as in Lemma 20.

For the simplified bound (which does not contain $N_{u;r}(\ ;\ )$), it remains for us to show that our choice of u and r ensures that $N_{u;r}(T_s;\ )\quad F$ and

$$4c_6\ {}^2T_s\log(2dT=\ )r^2+4c_7\ {}^2T_s^2\ {}^2u^4d^4(\log(T=\ ))^4\qquad T_st_f(\ )F:$$

First, our choice of u ensures that

$$u^4d^4\ {}^2(\log(T=\ ))^4\qquad r^2:$$

Next, recall that

$$N_{u;r}(\ ;\ ):=\ \frac{c_5^2}{64}\ {}^3t_f(\ )^2L^2\quad u^2d^2\quad \log\frac{T}{}\ \ ^2+\ p\ \overline{2\log(T=\ )}r\ \ !\ ^2$$
$$+\quad u\ {}^4\ {}^2\ \ c_1d^3\quad\log\frac{T}{}\ \ ^3+\ \ L\ \ {}^2u^4\ {}^2\ \ c_1d^4\quad\log\frac{T}{}\ \ ^4$$
$$+\quad c_1r^2(128t_f(\ )+\ \ L\ )\log\frac{T}{}+\ \ c_1L\ \ {}^2r^2$$
$$+\ \ c_5^2t_f^3(\ )\ \ {}^3L^2\quad u^2d^2\ \ \log(T=\ )+\ p\ \overline{2\log(T=\ )}r\ \ \ ^2:$$

Recalling our choice of such that

$$\min f\,1;\frac{1}{t_f(\ )};\frac{1}{t_f(\ )L}g;$$

it follows that

$$N_{u;r}(T_s;\ )\qquad T_sr^2\quad\frac{8c_5^2}{64}\log(T=\ )+2c_1+2c_1+(128c_1+1)\log(T=\ )+\ c_1+8c_5^2\log(T=\ )$$
$$T_sr^2\log(T=\ )\quad\frac{65c_5^2}{8}+132c_1+1\qquad F;$$

where the last inequality follows choosing r such that $r^2\quad\frac{F}{T_s\log(T=\ )\ \ \frac{65c_5^2}{8}+132c_1+1}$. Similarly, we have

$$4c_6\ {}^2T_s\log(2dT=\ )r^2+4c_7\ {}^2T_s^2\ {}^2u^4d^4(\log(T=\ ))^4$$
$$T_st_f(\ )\ \ 4c_6\ \ \log(2dT=\ )r^2+4c_7\ T_s\ \ {}^2u^4d^4(\log(T=\ ))^4$$
$$T_st_f(\ )\ \ 4c_6\ \ \log(2dT=\ )r^2+4c_7\ T_sr^2$$

By choosing r such that

$$r^2\qquad\frac{F}{4c_6\log(2dT=\ )+4c_7\ T_s};$$

it follows that

$$4c_6\ {}^2T_s\log(2dT=\ )r^2+4c_7\ {}^2T_s^2\ {}^2u^4d^4(\log(T=\ ))^4\qquad T_st_f(\ )F;$$

as desired.

We next lower bound the probability of

$$P_{t_0;T_s}(\ ;F\ ):=\ \backslash\ {}^{T_s}_{=1}\ (L_{t_0;}\ (\ ;F)\ \backslash\ N_{\ t_0;}\ (\ ))\ \backslash\quad\backslash\ {}^{t_0+T_s}_{t=t_0}\ {}^1O_t(\ )\ \backslash\ A_{\ t}(\ )\ \backslash\ G_t(\ )\quad\backslash\quad\backslash\ {}^{T_s}_{=t_f(\ )}E_{t_0;}(\ )\quad:$$

Observe that

$$\backslash\ {}^{T_s}_{=t_f(\ )}E_{t_0;}(\ )$$

$$= \backslash^{T_s}_{=t_f(\,)} H_{t_0;}(\,)\backslash \backslash_{t=t_0}^{t_0+\;1} A_t(\,)\backslash G_t(\,) \;!\; \backslash \backslash_{k=0}^{K\;2} B_{t_0+kt_f(\,)}(\,;t_f(\,)) \backslash B_{t_0+(K\;1)t_f(\,)}(\,;\;(K\;1)t_f(\,)) \;!\;!$$

$$= \backslash^{T_s}_{=t_f(\,)} H_{t_0;}(\,)\backslash \backslash_{k=0}^{K\;2} B_{t_0+kt_f(\,)}(\,;t_f(\,)) \backslash B_{t_0+(K\;1)t_f(\,)}(\,;\;(K\;1)t_f(\,)) \;!\; \backslash \backslash_{t=t_0}^{T_s\;1} A_t(\,)\backslash G_t(\,) \;!\;!\;:$$

Note this implies that $\backslash^{T_s}_{=t_f(\,)} E_{t_0;}(\,)\backslash \;\backslash^{T_s\;1}_{t=t_0} A_t(\,)\backslash G_t(\,) = \backslash^{T_s}_{=t_f(\,)} E_{t_0;}(\,)$ We note that by Lemma 1,

$$P \;\backslash^{T_s}_{=t_f(\,)} H_{t_0;}(\,)^{c} \;\frac{5T_s}{T}:$$

Meanwhile, we note that

$$\backslash^{T_s\;1}_{t=t_0} B_t(\,;t_f(\,)) \;\backslash^{T_s}_{=t_f(\,)} \backslash_{k=0}^{K\;2} B_{t_0+kt_f(\,)}(\,;t_f(\,)) \;!\; \backslash B_{t_0+(K\;1)t_f(\,)}(\,;\;(K\;1)t_f(\,)) \;!\;:$$

Hence, by Lemma 14, we have that

$$P \;\backslash^{T_s}_{=t_f(\,)} \backslash_{k=0}^{K\;2} B_{t_0+kt_f(\,)}(\,;t_f(\,)) \;!\; \backslash B_{t_0+(K\;1)t_f(\,)}(\,;\;(K\;1)t_f(\,)) \;!!\;^{c}!$$

$$P \;\backslash^{T_s\;1}_{t=t_0} B_t(\,;t_f(\,))^{c} \;\frac{T_s}{T}:$$

Meanwhile, by Lemma 13 and Lemma 15, we may bound

$$P \;\backslash_{t=t_0}^{T_s\;1} A_t(\,)\backslash G_t(\,) \;!\;^{c}! \;\frac{T_s}{T} + \frac{2T_s}{T} = \frac{3T_s}{T}:$$

Hence, it follows that

$$P \;\backslash^{T_s}_{=t_f(\,)} E_{t_0;}(\,)\backslash \;\backslash^{T_s\;1}_{t=t_0} A_t(\,)\backslash G_t(\,)^{c} \;\frac{5T_s}{T} + \frac{T_s}{T} + \frac{3T_s}{T} = \frac{9T_s}{T}:$$

Meanwhile, it follows from our results in the preceding lemmas that

$$P \;\backslash^{T_s}_{=1} (L_{t_0;}(\,;F)\backslash N_{t_0;}(\,))\backslash \;\backslash^{T_s\;1}_{t=t_0} O_t(\,)^{c} \;\frac{3T_s}{T}:$$

Hence, it follows that $P(P_{t_0;T_s}(\,;F))\;\;1\;\;\frac{12T_s}{T}$.

□

E.3. Proving function value decrease near saddle point

We next build on the technical result earlier to prove that each time we are near the saddle point, there is a constant probability of making significant function value decrease. We briefly provide a high-level proof outline below. In our proof, we introduce a coupling argument connecting two closely-related sequences both starting from the saddle, differing only in the sign of their perturbative term along the minimum eigendirection of the Hessian at the saddle. Specifically, when function decrease from a saddle is not sufficiently large, due to the earlier technical result, we know that the coupled sequences will remain within a radius of the original saddle for a large number (which we will denote as T_s) of iterations. We then utilize this fact to show that the difference of the coupled sequence will (with some constant probability) grow exponentially large, eventually moving out of their specified radius within T_s iterations, leading to a contradiction.

Our first result formally introduces the coupling, setting the stage for the rest of our arguments. For notational convenience, in this section, unless otherwise specified, we will often assume that the initial iterate x_0 is an -saddle point.

Lemma 3. Suppose x_0 is an ϵ-approximate saddle point. Without loss of generality, suppose that the minimum eigendirection of $H := \nabla^2 f(x_0)$ is the e_1 direction (i.e. the first basis vector in $\mathbb{R}^d$), and let γ to denote $-\lambda_{\min}(\nabla^2 f(x_0))$ (note $\gamma \geq \sqrt{\rho\epsilon}$). Consider the following coupling mechanism, where we run the zeroth-order gradient dynamics, starting with x_0, with two isotropic noise sequences Y_t and Y_t' respectively, where $(Y_t)_1 = -(Y_t)_1'$, and $(Y_t)_j = (Y_t)_j'$ for all other $j \neq 1$. Suppose that the sequence $\{Z_{t,i}\}_{t \in T, i \in [m]}$ is the same for both sequences. Let $\{x_t\}$ denote the sequence with the $\{Y_t\}$ noise sequence, and let the $\{x_t'\}$ denote the sequence with the $\{Y_t'\}$ noise sequence, where $x_0' = x_0$; and

$$x'_{t+1} = x'_t - \eta\left(\frac{\sum_{i=1}^m Z_{t,i} Z_{t,i}^\top \nabla f(x'_t) + \frac{u}{2} Z_{t,i} Z_{t,i}^\top \tilde{H}'_{t,i} Z_{t,i}}{m} + Y'_t\right);$$

and $\tilde{H}'_{t,i} := \frac{H'_{t,i,+} - H'_{t,i,-}}{2}$, with $H'_{t,i,+} = \nabla^2 f(x'_t + \theta'_{t,i,+} u Z'_i)$ for some $\theta'_{t,i,+} \in [0,1]$, and $H'_{t,i,-} = \nabla^2 f(x_t - \theta'_{t,i,-} u Z'_i)$ for some $\theta'_{t,i,-} \in [0,1]$. Then, for any $t \geq 0$,

$$\begin{aligned}
&\hat{x}_{t+1} \\
&:= x_{t+1} - x'_{t+1} \\
&= -\underbrace{\eta \sum_{\tau=0}^{t} (I - \eta H)^{t-\tau} \hat{\zeta}_{g_0}(\tau)}_{W_{g_0}(t+1)} - \underbrace{\eta \sum_{\tau=0}^{t} (I - \eta H)^{t-\tau} (\mathcal{H}_\tau - H)\hat{x}_\tau}_{W_H(t+1)} \\
&\quad - \underbrace{\eta \sum_{\tau=0}^{t} (I - \eta H)^{t-\tau} \hat{\zeta}_u(\tau)}_{W_u(t+1)} - \underbrace{\eta \sum_{\tau=0}^{t} (I - \eta H)^{t-\tau} \hat{Y}_\tau}_{W_p(t+1)}
\end{aligned}$$

where

$$\begin{aligned}
&\zeta_{g_0}(t) = \frac{1}{m}\sum_{i=1}^m (Z_{t,i} Z_{t,i}^\top - I)\nabla f(x_t); \\
&\zeta'_{g_0}(t) = \frac{1}{m}\sum_{i=1}^m (Z_{t,i} (Z_{t,i})^\top - I)\nabla f(x'_t); \\
&\hat{\zeta}_{g_0}(t) = \zeta_{g_0}(t) - \zeta'_{g_0}(t); \quad \zeta_u(t) = \frac{1}{m}\sum_{i=1}^m \frac{u}{2} Z_{t,i} Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i}; \\
&\zeta'_u(t) = \frac{1}{m}\sum_{i=1}^m \frac{u}{2} Z_{t,i} Z_{t,i}^\top \tilde{H}'_{t,i} Z_{t,i}; \; \hat{\zeta}_u(t) = \zeta_u(t) - \zeta'_u(t); \\
&\hat{Y}_t = Y_t - Y'_t; \quad \mathcal{H}_t = \int_0^1 \nabla^2 f(a x_t + (1-a) x'_t)\,da.
\end{aligned}$$

Proof. Observe that

$$\begin{aligned}
\hat{x}_{t+1} &:= x_{t+1} - x'_{t+1} \\
&= x_t - \eta(\nabla f(x_t) + \zeta_{g_0}(t) + \zeta_u(t) Y_t) - x'_t + \eta(\nabla f(x'_t) + \zeta'_{g_0}(t) + \zeta'_u(t) + Y'_t) \\
&= \hat{x}_t - \eta\big((\nabla f(x_t) - \nabla f(x'_t)) + \zeta_{g_0}(t) - \zeta'_{g_0}(t) + (\zeta_u(t) - \zeta'_u(t)) + (Y_t - Y'_t)\big) \\
&= \hat{x}_t - \eta H \hat{x}_t - \eta(\mathcal{H}_t - H)\hat{x}_t - \eta\hat{\zeta}_{g_0}(t) - \eta\hat{\zeta}_u(t) - \eta\hat{Y}_t \\
&= -\underbrace{\eta \sum_{\tau=0}^{t} (I - \eta H)^{t-\tau} \hat{\zeta}_{g_0}(\tau)}_{W_{g_0}(t+1)} - \underbrace{\eta \sum_{\tau=0}^{t} (I - \eta H)^{t-\tau} (\mathcal{H}_\tau - H)\hat{x}_\tau}_{W_H(t+1)} - \underbrace{\eta \sum_{\tau=0}^{t} (I - \eta H)^{t-\tau} \hat{\zeta}_u(\tau)}_{W_u(t+1)} - \underbrace{\eta \sum_{\tau=0}^{t} (I - \eta H)^{t-\tau} \hat{Y}_\tau}_{W_p(t+1)}
\end{aligned}$$

where

$$\zeta_{g_0}(t) = \frac{1}{m}\sum_{i=1}^m (Z_{t,i} Z_{t,i}^\top - I)\nabla f(x_t); \quad \zeta'_{g_0}(t) = \frac{1}{m}\sum_{i=1}^m (Z_{t,i} (Z_{t,i})^\top - I)\nabla f(x'_t); \quad \hat{\zeta}_{g_0}(t) = \zeta_{g_0}(t) - \zeta'_{g_0}(t);$$

$$_u(t) = \frac{1}{m}\sum_{i=1}^{m} \frac{u}{2} Z_{t;i} Z_{t;i}^\top H_{t;i} Z_{t;i}; \quad {}^0_u(t) = \frac{1}{m}\sum_{i=1}^{m} \frac{u}{2} Z_{t;i} Z_{t;i}^\top H^0_{t;i} Z_{t;i}; \quad \hat{}_u(t) = {}_u(t) \quad {}^0_u(t);$$

$$\hat{Y}_t = Y_t \quad Y^0_t; \quad H_t = \int_0^1 \nabla^2 f(a x_t + (1 \quad a) x^0_t)\, da:$$

To derive the final equality, we utilized the fact that $x^0_0 = x_0$. This completes our proof. □

Suppose x_0 is an -saddle point. Recall that > 0 denotes $_{\min}(\nabla^2 f(x_0))$, where we know that $\overline{\sqrt{\ }}$.

$$:= \begin{cases} \min\{\ ;\ 1; L\} & \text{if } f(\ ) \text{ is } (\ ;\ ;\ \sqrt{\ })\text{-strict saddle for any} > \sqrt{\ } \\ \sqrt{\ } & \text{otherwise} \end{cases}$$

In the sequel, for any $t \quad 0$, it is helpful to define the quantities

$$(t)^2 := \frac{(1+\ )^{2t}}{(\ )^2+2}; \qquad (t)^2 := \frac{(1+\ )^{2t} \quad 1}{(\ )^2+2}: \tag{46}$$

We next introduce some probabilistic events (and their implications) which, if true, can be used to bound the sizes of $\|W_{g_0}(t+1)\|; \|W_u(t+1)\|, \|W_u(t+1)\|$ (and as we will see in the next result, indirectly bound $\|W_H(t+1)\|$. These bounds will be useful in the final proof of making function value progress near a saddle point.

Lemma 24. We assume $2\ (0; 1=e]$ throughout the lemma. Suppose that we pick u; and as specified in Lemma 23. Suppose $T_s \quad t_f(\ )$. Suppose also that

$$f(x_{T_s}) \quad f(x_0) > \quad F; \quad f(x^0_{T_s}) \quad f(x_0) > \quad F:$$

Then, we have the following results.

1. Let $S(\ )$ denote the event

$$S(\ ) := \left\{ \max\{\|x_t \quad x_0\|^2; \|x^0_t \quad x_0\|^2\} \quad _{T_s}(\ ;F); \quad \forall 0 \quad t \quad T_s \right\}:$$

   In addition, let $S_u(\ )$ denote the event

$$S_u(\ ) := \left(\|W_u(t+1)\| \quad (t+1)\frac{\sqrt{3}}{\sqrt{\ }}\ 2c_3\, d\, {}^2(\log(T=\ ))^2 \quad u^2; \quad \forall 0 \quad t \quad T_s \quad 1 \right);$$

   where c_3 is the same absolute constant as c_3 in the preceding lemmas. Then,

$$P(S(\ ) \cap S_u(\ )) \quad 1 \quad \frac{24T_s}{T}:$$

2. Consider defining the event $R_t(\ )$, which is the event where

   either $\sum_{=0}^{t}(1+\ )^{2(t\ )}\frac{dL^2}{m}\|x \quad x^0\|^2(\log(CT^2=\ ))^2 \quad G_{T_s}(\ ;F)$; or

   $$\|W_{g_0}(t+1)\| \quad c^0\sqrt{\max\left(\log\frac{CT^2}{}\ {}^2\sum_{=0}^{t}\frac{dL^2}{m}(1+\ )^{2(t\ )}\|x \quad x^0\|^2; g(t+1)\right)\ \log\frac{CdT^2}{} + \log\ \log\frac{G_{T_s}(\ ;F)}{g(t+1)} + 1}$$

   normalsize holds. Above, ^{0}; C refer to the same constants as in Proposition 2, and

$$G_{T_s}(\ ;F) := 8\sum_{=0}^{T_s\ 1}(1+\ )^2\frac{dL^2}{m}(\log(CT^2=\ ))^2\ _{T_s}(\ ;F) + \left(\frac{(T_s)\ r}{60\sqrt{d}}\right)^2; \qquad g(t+1) := \left(\frac{(t+1)\ r}{60\sqrt{d}}\right)^2:$$

Then, $P(R_t(\ )) \quad 1 \quad \frac{}{T}$. Suppose the event

$$\bigcap_{t=0}^{T_s \quad 1} R_t(\ ) \quad \bigcap S \quad (\ )$$

holds. Then, the event $S_{g_0}(\ )$ holds, where

$$S_{g_0}(\ ) := \bigcap_{t=0}^{T_s \quad 1} S_{g_0;t}(\ );$$

and $S_{g_0;t}(\ )$ is defined as

$$S_{g_0;t}(\ ) := \left\{ \|W_{g_0}(t+1)\| \quad {}_1(\ ; F\ ) c^0 \sqrt{t \quad \max \quad lr \left(\frac{CT^2}{} \quad {}^2\sum_{=0}^{t} \frac{dL^2}{m}(1 + \quad)^{2(t \quad)} \|x \quad x^0\|^2 ; g(t+1) \right)} \right\};$$

where

$$_1(\ ; F\ ) := \quad \log \quad \frac{CdT^2}{} \quad + \log \quad \log \quad \frac{G_{T_s}(\ ; F\ )}{g(1)} \quad + 1 \quad :$$

3. In addition, let $S_p(\ )$ denote the event

$$S_p(\ ) := \quad \left(\|W_p(t+1)\| \quad \frac{2\sqrt{2\log(T/\ )} \quad (t+1) \quad r}{\sqrt{d}} \quad 80 \quad t \quad T_s \quad 1 \right) :$$

Then, $P(S_p(\ )) \quad 1 \quad \frac{T_s}{T}$.

Proof. We consider the three claims separately.

1. Note that our assumptions satisfy the conditions required in Lemma 23. Hence, by Lemma 23, on the event $P_{0;T_s}(\ ; F\ )$, we have that $\|x \quad x_0\|^2 \quad {}_{T_s}(\ ; F\ )$. Simultaneously, on the event $P_{0;T_s}(\ ; F\ )$, we know that $\bigcap_{t=0}^{T_s \quad 1} G_t(\ )$ holds, i.e.

$$\frac{1}{m}\sum_{i=1}^{m} \|Z_{t;i}\|^4 \quad 2c_3 d^2 (\log(T/\ ))^2; \quad 80 \quad t \quad T_s \quad 1: \tag{47}$$

Thus, for $W_u(t+1)$, we have that

$$\begin{aligned}
\|W_u(t+1)\| = {} & \sum_{=0}^{t} (I \quad H\ )^t \quad \hat{}_u(\ ) \\
& \sum_{=0}^{t} (I \quad H\ )^t \quad {}_u(\ ) \quad + \quad \sum_{=0}^{t} (I \quad H\ )^t \quad \hat{}^0_u(\ ) \\
& \sum_{=0}^{t} (1 + \quad)^t \quad \left(\frac{1}{m}\sum_{i=1}^{m} \frac{u}{2} Z_{t;i} \ Z_{t;i} \ \tilde{H}_{t;i} \ Z_{t;i} \quad + \quad \frac{1}{m}\sum_{i=1}^{m} \frac{u}{2} Z_{t;i} \ Z_{t;i} \ \tilde{H}^0_{t;i} \ Z_{t;i} \right) \\
& \sum_{=0}^{t} (1 + \quad)^t \quad \frac{}{m} \sum_{i=1}^{m} \|Z_{t;i}\|^4 u^2 \\
\overset{\text{(iv)}}{} & \sum_{=0}^{t} (1 + \quad)^t \quad (2c_3) d^2 (\log(T2/\ ))^2 u^2 \\
& \frac{(1 + \quad)^{t+1}}{} \quad 2c_3 \ Cd^2 (\log(T/\ ))^2 \quad u^2 \\
\overset{\text{(v)}}{=} {} & (t+1) \frac{\sqrt{(\quad)^2 + 2}}{} \quad 2c_3 \ d^2 (\log(T/\ ))^2 \quad u^2 \\
& (t+1) \frac{\sqrt{3}}{\sqrt{}} \quad 2c_3 \ d^2 (\log(T/\ ))^2 \quad u^2
\end{aligned}$$

$$\overset{(vi)}{}\ (t+1)\,\frac{\sqrt{3}}{\sqrt{}}\ \ 2c_3\, d^{2}(\log(T=))^{2}\ \ u^{2}$$

where the inequality in (iv) holds due to Eq. (47), the equality in (v) holds due to the definition of $(t+1)$, and the inequality in (vi) used the fact that .

Hence the event

$$\backslash_{t=0}^{T_s}\left\{\ \|x_t\ \ x_0\|^2\ \ \ _{T_s}(\ ;F\ )\right\}\ \text{and}\ \ \backslash\, S_u(\ )$$

holds with probability at least $1\ \ \frac{12T_s}{T}$.

Note that by the coupling, the distribution of x^0 is the same as that of . Thus, by the assumption $f(x^0_{T_s})\ \ f(x_0) >\ \ F$, it follows by a similar argument that the bound $\|x^0\ \ x_0\|^2\ \ \ _{T_s}(\ ;F\ )$ also holds with probability at least $1\ \ \frac{12T_s}{T}$. The claim then follows by an application of the union bound.

2. For the second claim, observe first that the claim $P(R_t(\ ))\ \ 1\ \ \frac{}{T}$ is a consequence of Proposition 2. Suppose next that $f(x_{T_s})\ \ f(x_0) >\ \ F$. Then, by definition of the event $S(\ )$, we know that

$$\|x\ \ x_0\|^2\ \ \ _{T_s}(\ ;F\ );\qquad \|x^0\ \ x_0\|^2\ \ \ _{T_s}(\ ;F\ )$$

where $_{T_s}(\ ;F\ )$ is as defined in Lemma 23.

Suppose now that $R_t(\ )$ holds true, and suppose for contradiction that

$$\sum_{=0}^{t}(1+\ \ )^{2(t\ \ )}\frac{dL^2}{m}\|x\ \ x^0\|^2(lr(\ CT^2=))^2$$
$$G_{T_s}(\ ;F\ )$$
$$=8\sum_{=0}^{T_s\ 1}(1+\ \ )^2\ \frac{dL^2}{m}(lr(\ CT^2=))^2\ \ _{T_s}(\ ;F\ )+\ \ \left(\frac{(T_s)\ r}{60\sqrt{d}}\right)^2:$$

This implies that there exists some $0\ \ \ t\ \ \ T_s$ such that $\|x\ \ x^0\|^2\ \ 8\ \ _{T_s}(\ ;F\ )$. However, we also know that on the event $S(\ )$,

$$\|x\ \ x^0\|^2\ \ 2\|x\ \ x_0\|^2+2\|x^0\ \ x_0\|^2\ \ 4\ \ _{T_s}(\ ;F\ ):$$

This leads to a contradiction. We must then have that

$$\|W_{g_0}(t+1)\|\ \ \ _1(\ ;F\ )c^0\sqrt{t}\ \max\left\{\ lr\ \ \frac{CT^2}{}\ \ ^2\sum_{=0}^{t}\frac{dL^2}{m}(1+\ \ )^{2(t\ \ )}\|x\ \ x^0\|^2;g(t+1)\right\};$$

where

$$_1(\ ;F\ ):=\sqrt{\log\ \ \frac{CdT^2}{}\ \ +\log\ \ \log\ \ \frac{G(\ ;F\ )}{g(1)}\ \ +1}$$

3. Observe that

$$W_p(t+1)=\sum_{=0}^{t}(I\ \ \ H\ )^t\ \ \hat{Y}\ =\sum_{=0}^{t}(1+\ \ )^t\ \ (2(Y\ )_1);$$

which means that $W_p(t+1)$ is a 1-dimensional Gaussian with variance

$$^2\sum_{=0}^{t}(1+\ \ )^{2(t\ \ )}\frac{4r^2}{d}=\frac{4\ ^2r^2}{d}\frac{(1+\ \ )^{2(t+1)}\ \ 1}{2\ \ +(\ \ )^2}=\frac{4\ ^2r^2\ \ (t+1)^2}{d}:\qquad(48)$$

Since $(t+1)$ $(t+1)$, using the subGaussianity of a Gaussian distribution, it follows that for any t, with probability at least 1 $=T$,

$$k W_p(t+1)\,k \quad \frac{2\sqrt{2\log(T=\ )}\ (t+1)\ r}{\sqrt{d}}:$$

□

For any $F>0$, we are now ready to show that the algorithm makes a function decrease of F with (1) probability near an -saddle point.

Proposition 5. Suppose that x_{t_0} is an -approximate saddle point. Let ${}^0>0; c_1>0; c_2\ \ 1; c_4>0; c_5>0; c_6>0; c_7>0; C_1\ \ 1$ be the absolute constants defined in the statements of the previous lemmas, and let $2(0;1=e]$ be arbitrary. Consider any $F>0$. As in the statement of Lemma 23, suppose we choose u, r and such that

$$u\quad \frac{\sqrt{\ }}{d\sqrt{\ }\log(T=\ )}\ \min\ \frac{1}{64c_5^2c_2};\frac{1}{2048c_1c_2}\ {}^{1=4};\qquad r\qquad \min\ \frac{1}{8c_5\sqrt{2c_2}};\frac{1}{32\sqrt{c_1}};$$

$$\frac{1}{Lt_f(\ )}\ \min\ \frac{1}{\log(T=\ )};\frac{\sqrt{m}}{8c_4(\mathrm{lr}(\ C_1dmT=\ ))^{3=2}\sqrt{d}};\frac{m}{128c_1(\mathrm{lr}(\ C_1dmT=\ ))^3d}:$$

Suppose we pick

$$T_s=\max\ \ d\frac{\ }{\ }e; t_f(\ ); 4\ \ ; \tag{49}$$

where

$$=\max\ \left\{\log\ 2\sqrt{\ \ \ _{T_s}(\ ;F\ )}\ \frac{20\sqrt{\ }\sqrt{d}\sqrt{\ ^2\ ^2+2}\,!}{r};1\right\};$$

$$:=\begin{cases}\min f\ ;\ 1;Lg & \text{if } f(\ ) \text{ is } (\ ;\ ;\ \sqrt{\ })\text{-strict saddle for any } >\ \sqrt{\ }\\ \sqrt{\ } & \text{otherwise}\end{cases}$$

Suppose in addition that u; also satisfy the conditions

$$u\qquad \sqrt{\frac{\sqrt{\ }}{120\sqrt{3}c_3\sqrt{d}\,d^2(\log(T=\ ))^2}}\ r;\qquad \max\ \left\{\frac{1}{c^0c_9\ \ _1(\ ;F\ )};\frac{m}{360\,(c^0)^2c_9^2dL^2\ \ \mathrm{lr}\ \ \frac{CT^2}{\ }\ \ ^2\ \ _1(\ ;F\ )^2};\frac{1}{2}\right\};$$

where $_1(\ ;F\ )$ is as defined in Lemma 23, $c^0; c_3; C>0$ are the same constants as in the previous results, and $c_9=2\sqrt{2}+\frac{1}{20}$. Suppose also that $_{T_s}(\ ;F\ )$ satisfies the bound

$$_{T_s}(\ ;F\ )\qquad \frac{\ }{60c_9\ \ \log(T=\ )}^2: \tag{50}$$

Then, with probability at least $1-\frac{1}{3}\ \ \frac{13T_s}{T}$, $f(x_{t_0+T_s})\ \ f(x_{t_0})\ \ \ F$.

Proof of Proposition 5. Without loss of generality, we assume that $t_0=0$. By Lemma 3, we have

$$\begin{aligned}\hat{x}_{t+1}&\\ :=\ x_{t+1}\ \ x^0_{t+1}&\\ =\ \ &\underbrace{\sum^t_{=t_0}(I\ \ H\ )^t\ \ \hat{\ }_{g_0}(\ )}_{W_{g_0}(t+1)}\ \ \underbrace{\sum^t_{=t_0}(I\ \ H\ )^t\ (H\ \ H\ )\hat{x}}_{W_H(t+1)}\ \ \underbrace{\sum^t_{=t_0}(I\ \ H\ )^t\ \ \hat{\ }_u(\ )}_{W_u(t+1)}\ \ \underbrace{\sum^t_{=t_0}(I\ \ H\ )^t\ \ \hat{Y}}_{W_p(t+1)}\end{aligned}$$

where

$$_{g_0}(t) = \frac{1}{m}\overset{m}{\underset{i=1}{X}} (Z_{t;i}\, Z_{t;i}^{>} \quad I)r\, f(x_t); \quad {}^0_{g_0}(t) = \frac{1}{m}\overset{m}{\underset{i=1}{X}} (Z_{t;i}\,(Z_{t;i})^{>} \quad I)r\, f(x_t^0); \quad \hat{}_{g_0}(t) = {}_{g_0}(t) \quad {}^0_{g_0}(t);$$

$$_{u}(t) = \frac{1}{m}\overset{m}{\underset{i=1}{X}} \frac{u}{2} Z_{t;i}\, Z_{t;i}\, H^{>}_{t;i}\, Z_{t;i}; \quad {}^0_{u}(t) = \frac{1}{m}\overset{m}{\underset{i=1}{X}} \frac{u}{2} Z_{t;i}\, Z_{t;i}\, H^{>0}_{t;i}\, Z_{t;i}; \quad \hat{}_{u}(t) = {}_{u}(t) \quad {}^0_{u}(t);$$

$$\hat{Y}_t = Y_t \quad Y_t^0; \quad H_t = \overset{1}{\underset{0}{Z}} r^2 f(ax_t + (1 \quad a)x_t^0)da:$$

Recall that we define for $t \quad 0$,

$$(t)^2 := \frac{(1 + \quad)^{2t}}{(\quad)^2 + 2}; \qquad (t)^2 := \frac{(1 + \quad)^{2t} \quad 1}{(\quad)^2 + 2}:$$

Throughout the proof, we suppose for contradiction that

$$f(x_{T_s}) \quad f(x_0) > \quad F; \quad f(x^0_{T_s}) \quad f(x_0) > \quad F;$$

and assume the event

$$\backslash^{T_s \quad 1}_{t=0} R_t(\ ) \quad \backslash S \quad (\ ) \backslash S_u(\ ) \backslash S_p(\ )$$

holds, where the events intersected are defined in Lemma 24. Then, by Lemma 24, the event $S_{g_0}(\ )$ (also defined in Lemma 24) holds[8].

Consider the following induction argument, where we seek to show that there exists an absolute constant $c_9 > 0$ such that for every $t\ 2\ f\, 0; 1; : : : ; T_s g$,

$$kx_t \quad x_t^0 k \quad c_9 \log(T = \ ) \frac{(t)\ r}{p \overline{d}}; \text{ and } \max fk\, W_{g_0}(t)k; kW_H(t)k; kW_u(t)kg \quad \frac{(t+1)\ r}{p \overline{d}} \tag{51}$$

Combined with a lower bound on $kW_p(t+1)k$ (which makes use of the property that $W_p(t+1)$ is a 1-dimensional Gaussian), we will then use the inductive claim in Eq. (51) to show that

$$kW_p(T_s)k \quad 2 \quad W_{g_0(T_s)} \quad + kW_H(T_s)k + kW_u(T_s)k \ :$$

Since $W_p(t+1)$ is a 1-dimensional Gaussian random variable with a standard deviation that grows exponentially with t, by our choice of T_s, we will see that $x_{T_s} \quad x^0_{T_s}$ is larger than what expect (since our assumptions imply that $\max \overset{n}{\ } kx_{T_s} \quad x_0 k^2; \ x^0_{T_s} \quad x_0 \ ^2 \overset{o}{\ } \quad {}_{T_s}(\ ; F)$, i.e. x_{T_s} and $x^0_{T_s}$ both remain close to x_0 and hence close to each other). This yields a contradiction, implying that on the event we assumed to hold, i.e.

$$\backslash^{T_s \quad 1}_{t=0} R_t(\ ) \quad \backslash S \quad (\ ) \backslash S_p(\ )$$

the assumption

$$f(x_{T_s}) \quad f(x_0) > \quad F; \quad \text{and} f(x^0_{T_s}) \quad f(x_0) > \quad F$$

is not true, i.e. one of the sequences must have made function value progress of at least F:

We proceed to prove Eq. (51). Observe that the claim holds for the base case $t = 0$; this is true since $x_0 = x^0_0$. Now suppose that this holds for all $\quad t$. We will seek to show that Eq. (51) holds for $+ 1$ as well. We do so by bounding the norms of $W_{g_0}(t+1); W_H(t+1); W_u(t+1)$ and $W_p(t+1)$ respectively.

[8] We may also directly assume that $S_{g_0}(\ )$ also holds, but our way of reasoning prevents double counting of probabilities.

1. (Bounding $\|W_{g_0}(t+1)\|$) Since the event $S_{g_0}(\ )$ holds, it follows that for each $0 \quad t \quad T_s \quad 1$, we have that

$$\|W_{g_0}(t+1)\| \quad {}_1(\ ;F\ )c^0 \sqrt{t \ \max\left\{ lr \ \frac{CT^2}{} \ {}^2\sum_{=0}^{t} \frac{dL^2}{m}(1+ \quad)^{2(t \quad)}\|x \quad x^0\|^2; g(t+1)\right\}}$$

where

$$_1(\ ;F\ ) := \quad \log \ \frac{CdT^2}{} \ +\log \ \log \ \frac{G_{T_s}(\ ;F\ )}{g(1)} \quad +1 \quad ;$$

and the terms $G_{T_s}(\ ;F\ )$ and $g(1)$ are defined as in Lemma 24. Recall by the inductive claim in Eq. (51) that there exists $c_9 > 0$ such that

$$\|x \quad x^0\| \quad c_9 \log(T= \ )\frac{(t)\ r}{\sqrt{d}} \quad 8\ 0 \qquad t:$$

Hence, it follows that

$$\|W_{g_0}(t+1)\| \quad c^0 \ {}_1(\ ;F\ ) \ \max\left\{ \sqrt{t+1} \ \ lr \ \frac{CT^2}{} \quad \frac{c_9 \sqrt{\ dL}}{\sqrt{\ m}} \frac{(t)\ r}{\sqrt{d}};\ \frac{(t+1)\ r}{60\sqrt{d}}\right\} \ :$$

Hence, noting the choice of T_s in Eq. (49), by choosing such that

$$c^0 c_9 \ {}_1(\ ;F\ ) \ \sqrt{T_s} \ \ lr \ \frac{CT^2}{} \quad \frac{\sqrt{\ dL}}{\sqrt{\ m}} \quad \frac{1}{60} \ () \qquad \frac{m}{360\,(c^0)^2 c_9^2 dL^2 \ \ lr \ \frac{CT^2}{} \ {}^2 \ {}_1(\ ;F\ )^2}; \text{ and} \tag{52}$$

$$c^0 c_9 \ {}_1(\ ;F\ ) \qquad 1:$$

it follows that

$$\|W_{g_0}(t+1)\| \quad \frac{(t+1)\ r}{60\sqrt{d}}:$$

2. Meanwhile, the term $W_H(t+1)$ can be bounded as follows. By the inductive assumption in Eq. (51), we have that

$$\|\hat{x}\| = \|x \quad x^0\| \quad c_9 \log(T= \ )\frac{(\ )\ r}{\sqrt{d}} \quad 8\ 0 \qquad t:$$

Moreover, on the event our proof assumes, we know that

$$\max\left\{\|x \quad x_0\|^2; \|x^0 \quad x_0\|^2\right\} \qquad {}_{T_s}(\ ;F\ ):$$

Thus, using the -Hessian Lipschitz property, we have

$$\begin{aligned}
\|W_H(t+1)\| &= \sum_{=0}^{t} (I \quad H\ )^{t} \quad (H \quad H\ )\hat{x} \\
&\quad \sum_{=0}^{t} (1+ \quad)^{t} \quad \sqrt{{}_{T_s}(\ ;F\ )}\,\frac{c_9 \log(T= \ ) \ (\ )\ r}{\sqrt{d}} \\
&\quad c_9(t+1)\log(\ T= \ ) \quad \sqrt{{}_{T_s}(\ ;F\ )}\,\frac{(t)\ r}{\sqrt{d}} \\
&\quad c_9 T_s \log(T= \ ) \quad \sqrt{{}_{T_s}(\ ;F\ )}\,\frac{(t)\ r}{\sqrt{d}}:
\end{aligned}$$

Given our choice of T_s in Eq. (49), if

$$c_9 T_s \log(T= \ ) \quad \sqrt{{}_{T_s}(\ ;F\ )} \quad \frac{1}{60} \ () \qquad {}_{T_s}(\ ;F\ ) \qquad \frac{}{60 c_9 \quad \log(T= \ )}^{2}$$

it follows that

$$\|W_H(t+1)\| \quad \frac{(t+1)\ r}{60\sqrt{d}}:$$

3. Meanwhile, for $W_u(t+1)$, since the event $S_u(\ )$ holds, we have that

$$kW_u(t+1)k \qquad (t+1)\,\frac{p\,\overline{3}}{p\,\overline{\quad}}\; 2c_3\, d^{\,2}(\log(T=\ ))^{2}\;\; u^{2}:$$

Now, by picking

$$(t+1)\,\frac{p\,\overline{3}}{p\,\overline{\quad}}\; 2c_3\, d^{\,2}(\log(T=\ ))^{2}\;\; u^{2} \qquad \frac{(t+1)\;\; r}{60\;\; p\,\overline{d}} \quad () \qquad u \qquad s\,\overline{\frac{p\,\overline{\quad}\qquad r}{120\;\; p\,\overline{3}c_3\;\; p\,\overline{d}\; d^{\,2}(\log(T=\ ))^{2}}};$$

it follows that with probability $1 \quad =T$, $kW_u(t+1)k \quad \frac{(t+1)\;\; r}{60\; p\,\overline{d}}$.

4. Meanwhile, observe that since $S_p(\ )$ holds, it follows that

$$W_p(t+1) \qquad \frac{2\;p\,\overline{2\log(T=\ )}\;\; (t+1)\;\; r}{p\,\overline{d}}:$$

Combining the bounds for $W_{g_0}; W_p; W_H$ and W_u, it follows that

$$\begin{aligned} k\hat{x}_{t+1}k \quad & k\;\; W_{g_0}(t+1)k + kW_p(t+1)k + kW_H(t+1)k + kW_u(t+1)k \\ & \frac{(t+1)\;\; r}{p\,\overline{d}} \quad \frac{1}{60} + \frac{1}{60} + \frac{1}{60} + 2\;p\,\overline{2\log(T=\ )} \\ & \frac{(t+1)\;\; r}{p\,\overline{d}} \quad \frac{1}{20} + 2\;p\,\overline{2}\quad \log(T=\ ); \end{aligned}$$

where the final inequality uses the fact that $0 < \quad 1=e$ (which implies $\log(T=\ ) \quad 1$). Hence, we see that the first part of the inductive claim of Eq. (51) holds with the constant $a_{gt} := \frac{1}{20} + 2\;p\,\overline{2}$, and the second part follows naturally as a consequence of our argument above.

Meanwhile, observe that for any such that $\quad \frac{1}{2}$, we have that $(1+\quad)^{\frac{1}{\ }} \quad 2$. Thus, by choosing such that $\quad \frac{1}{2}$, we have that for any $t \quad \frac{1}{\ }$,

$$(t+1)^{\,2} \quad \frac{1}{2}\;\; (t+1)^{\,2}:$$

Hence, following Eq. (48), by choosing $T_s \quad \frac{1}{\ }$, $W_p(T_s)$ is a 1-dimensional Gaussian with variance at least $\frac{2\;\;^2 r^2\;\; (T_s)}{d}$, such that with probability at least 2/3,

$$kW_p(T_s)k \quad \frac{(T_s)\;\; r}{10\;\; p\,\overline{d}}:$$

Simultaneously, we know that on the event

$$\backslash_{t=0}^{T_s\;\;1} R_t(\ )\;\; \backslash S\;\;(\ ) \backslash S_u(\ ) \backslash S_p(\ );$$

we have

$$kW_{g_0}(T_s)k + kW_H(T_s)k + kW_u(T_s)k \quad \frac{3\;\;(T_s)\;\; r}{60\;\; p\,\overline{d}} = \frac{(T_s)\;\; r}{20\;\; p\,\overline{d}}:$$

We note that by Lemma 24, we have

$$P\quad \backslash_{t=0}^{T_s\;\;1} R_t(\ )\;\; \backslash S\;\;(\ ) \backslash S_u(\ ) \backslash S_p(\ ) \qquad 1 \qquad \frac{24T_s}{T} + \frac{T_s}{T} + \frac{T_s}{T} \quad = 1 \qquad \frac{26T_s}{T}:$$

Thus, with probability at least 2=3 $\quad \frac{26T_s}{T}$, we have

$$k\hat{x}_{T_s}k \quad \frac{1}{2}kW_p(T_s)k \quad \frac{(T_s)\ r}{20\sqrt{d}}$$

Thus, choosing $T_s \quad -$, where

$$= \max\left(\log_2\left(\frac{\sqrt{\quad}\ \overline{\ _{T_s}(\ ;F\ )}\ 20\sqrt{\ }\sqrt{d}\sqrt{\ \overline{2\ \ 2+2}}}{r}\right);1\ \right);$$

noting that if $\quad 1=2$, then $(1+\quad)^{\frac{1}{}} \quad (1+\quad)^{\frac{1}{}} \quad 2$, we have that with probability at least 2=3 $\quad \frac{26T_s}{T}$,

$$\begin{aligned} k\hat{x}_{T_s}k \quad & \frac{(T_s)\ r}{20\sqrt{d}} = \frac{r}{20\sqrt{d}}\frac{\sqrt{(1+\quad)^{T_s}}}{\sqrt{2\quad+(\quad)^2}} \\ & \frac{r}{20\sqrt{d}}\frac{\sqrt{(1+\quad)^{\log_2\left(\frac{\sqrt{\ _{T_s}(\ ;F\ )}\ 20\sqrt{\ }\sqrt{d}\sqrt{2\ \ 2+2}}{r}\right)}}}{\sqrt{2\quad+(\quad)^2}} \\ & \frac{r}{20\sqrt{\ }\sqrt{d}\sqrt{2\quad+(\quad)^2}}2^{\log_2\left(\frac{\sqrt{\ _{T_s}(\ ;F\ )}\ 20\sqrt{\ }\sqrt{d}\sqrt{2\ \ 2+2}}{r}\right)} > 2\sqrt{\ _{T_s}(\ ;F\ )} > 2\sqrt{(T_s;\ )}: \end{aligned}$$

Thus, at least one of $kx_{T_s} \quad x_0k$ and $x^0_{T_s} \quad x_0$ is larger than $\sqrt{(T_s;\ )}$, a contradiction. Since the two sequences have the same distribution, it follows that with probability at least 1=3 $\quad \frac{13T_s}{T}$, $f(x_{T_s}) \quad f(x_0) \quad F$. □

In the result above, we require an upper bound on the norm of $_{T_s}(\ ;F\ )$ to hold (i.e. equation 50), which in turn necessitates an upper bound on F, the function value improvement we can expect to make. Below, we show how to choose F to be as large as possible (up to constants and logarithmic factors) whilst still satisfying equation 50, assuming that u, r and are chosen appropriately small such that the dominant term of $k\ _{T_s}(\ ;F\ )k$ scales with F.

Lemma 25. Consider choosing F such that

$$F = \frac{1}{2}\left(\frac{}{60c_9\quad\log(T=\ )}\right)^2\frac{1}{T_s t_f(\ )(129+8c_2^0\ \ _1(\ ;F)(16(\mathrm{lr}(\ CT^2=\ ))^2+1))}:$$

Suppose $\quad \min\left\{1;\frac{1}{t_f(\ )};\frac{1}{t_f\ L}\right\}$. Suppose we pick u and r small enough such that

$$u \quad \frac{r^{1=2}}{d\log(T=\ )^{1=2}};\quad r^2 \quad \min\left\{\frac{F}{2\ \log(T=\ )\ \frac{65c_5^2}{8}+6c_1+1};\frac{F}{4c_6\log(2dT=\ )+\frac{8c_7}{}};\right\}:$$

Then, $N_{u;r}(T_s;\ ) \quad F$, and that

$$4c_6\ {}^2T_s\log(2dT=\ )r^2+4c_7\ {}^2T_s^2\ {}^2u^4d^4(\log(T=\ ))^4 \quad T_s t_f(\ )F:$$

Suppose in addition is small enough so that

$$32\ {}^2(t_f(\ ))^2\ {}^2 \quad \frac{1}{2}\left(\frac{}{60c_9\quad\log(T=\ )}\right)^2:$$

Suppose also that $\quad 1$[9] and $\quad \frac{m}{d}$, so that $T_s \quad — \quad \frac{d}{m}$. Then, the condition in Eq. (50) will be satisfied.

[9] Without loss of generality, we may set $= 1$ if $f(\ )$ is $(\ ;\ ;\ \sqrt{—})$-strict saddle for any $> \quad 1$.

Proof. We note that since $\frac{\iota}{\bar{\psi}} \leq T_s \leq \frac{2\iota}{\bar{\psi}}$, it follows by our choice of r that r also satisfies the condition

$$r^2 \leq \min\left\{\frac{F}{\eta T_s \log(T/\delta)\left(\frac{65c_5^2}{8} + 6c_1 + 1\right)}, \frac{F}{4c_6\log(2dT/\delta) + 4c_7\eta T_s}\right\}.$$

Hence, our choice of η, u and r satisfies the conditions in Lemma 23, and it follows then that

$$\phi_{T_s}(\delta, F) \leq \max\left(128\eta T_s t_f(\delta)F, 32\eta^2(t_f(\delta))^2\epsilon^2\right) + 8c'^2\beta_1(\delta; F)\eta t_f(\delta)\max\left(\frac{16d}{m}(\mathrm{lr}(CT^2/\delta))^2F, T_sF\right) + T_s\eta t_f(\delta)F,$$

where $\beta_1(\delta; F)$ is as defined in Lemma 21.

The condition in Eq. (50) requires that

$$\phi_{T_s}(\delta, F) \leq \left(\frac{\bar{\psi}}{60c_9\iota\rho\log(T/\delta)}\right)^2.$$

By our choice of η such that

$$32\eta^2(t_f(\delta))^2\epsilon^2 \leq \frac{1}{2}\left(\frac{\bar{\psi}}{60c_9\iota\rho\log(T/\delta)}\right)^2,$$

it suffices for us to show that

$$\begin{aligned}\frac{1}{2}\left(\frac{\bar{\psi}}{60c_9\iota\rho\log(T/\delta)}\right)^2 &\geq 128\eta T_s t_f(\delta)F + 8c'^2\beta_1(\delta; F)\eta t_f(\delta)\max\left(\frac{16d}{m}(\mathrm{lr}(CT^2/\delta))^2F, T_sF\right) + \eta T_s t_f(\delta)F \\ &= 129\eta T_s t_f(\delta)F + 8c'^2\beta_1(\delta; F)\eta t_f(\delta)\max\left(\frac{16d}{m}(\mathrm{lr}(CT^2/\delta))^2F, T_sF\right).\end{aligned}$$

By our assumption, we know that $T_s \geq \frac{d}{m}$. Thus, further simplifying indicates that it suffices for us to show

$$\frac{1}{2}\left(\frac{\bar{\psi}}{60c_9\iota\rho\log(T/\delta)}\right)^2 \geq 129\eta T_s t_f(\delta)F + 8c'^2\beta_1(\delta; F)\eta t_f(\delta)\max\left(16T_s(\mathrm{lr}(CT^2/\delta))^2F, T_sF\right). \tag{53}$$

By choosing F such that

$$F \leq \frac{1}{2}\left(\frac{\bar{\psi}}{60c_9\iota\rho\log(T/\delta)}\right)^2\frac{1}{\eta T_s t_f(\delta)\left(129 + 8c'^2\beta_1(\delta; F)\left(16(\mathrm{lr}(CT^2/\delta))^2 + 1\right)\right)},$$

we see that Eq. (53) is satisfied.

□

Remark 3. Suppose without loss of generality that $T_s = \frac{\iota}{\eta\bar{\psi}}$. Then, as a consequence of Lemma 25, we note that the amortized function value progress of decreasing function value by F over T_s iterations is

$$\begin{aligned}\frac{F}{T_s} &= \frac{1}{2}\left(\frac{\bar{\psi}}{60c_9\iota\rho\log(T/\delta)}\right)^2\frac{1}{\eta T_s^2 t_f(\delta)\left(129 + 8c'^2\beta_1(\delta; F)\left(16(\mathrm{lr}(CT^2/\delta))^2 + 1\right)\right)} \\ &= \eta\frac{\bar{\psi}^4}{\rho^2}\left(\frac{1}{2\iota^2}\frac{1}{(60c_9\iota\log(T/\delta))^2\left(t_f(\delta)\right)\left(129 + 8c'^2\beta_1(\delta; F)\left(16(\mathrm{lr}(CT^2/\delta))^2 + 1\right)\right)}\right)\end{aligned}$$

F. Proving the main result (informal statement in Theorem 1, full statement in Theorem 2)

In this section, we prove our main result. First, we need an additional result (Lemma 26) showing that with high probability, we can bound the function value increase if a saddle appears within $t_f(\delta)$ iterations immediately after we have had T_s iterations after the previous saddle. We note that such a bound is necessary because our earlier result upper bounding function increase in τ iterations (see Lemma 18) focused on the case where $\tau \geq t_f(\delta)$. Next, we state and prove Theorem 2, which is the precise version of Theorem 1 in the main text.

Lemma 26 (Function change for small τ)**.** *Let $c_1 > 0, c_4 > 0, c_5 > 0, C_1 \geq 1$ be the absolute constants defined in the statements of the previous lemmas. Let $\delta \in (0, 1/e]$, and suppose $\tau < t_f(\delta)$.*

Let J denote the interval $\{0, 1 \ldots, \tau - 1\}$ where $\tau < t_f(\delta)$.

Suppose we choose η such that

$$\eta \leq \frac{1}{Lt_f(\delta)} \cdot \min \left\{ \frac{\sqrt{m}}{8c_4(\mathrm{lr}(C_1dmT/\delta))^{3/2}\sqrt{d}}, \frac{m}{128c_1(\mathrm{lr}(C_1dmT/\delta))^3 d} \right\}. \tag{54}$$

Suppose also we pick u, r and η as prescribed in the statement of Proposition 4.

Suppose that $\min_{t \in J}\|\nabla f(x_t)\| \leq \epsilon$. Then, on the event

$$\mathcal{D}_\tau(\delta) := \mathcal{H}_{0,\tau}(\delta) \cap \left(\bigcap_{t=0}^{\tau-1} \mathcal{A}_t(\delta) \right) \cap \left(\bigcap_{t=0}^{\tau-1} \mathcal{G}_t(\delta) \right),$$

we have the following upper bound on function value change:

$$\begin{aligned} f(x_\tau) - f(x_0) \leq{} & \frac{\eta}{4}\epsilon^2 + t_f(\delta)\eta u^4\rho^2 \cdot c_1 d^3 \left(\log\frac{T}{\delta}\right)^3 + t_f(\delta)L\eta^2u^4\rho^2 \cdot c_1 d^4 \left(\log\frac{T}{\delta}\right)^4 \\ & + \eta c_1 r^2(128t_f(\delta) + \eta L)\log\frac{T}{\delta} + t_f(\delta)c_1L\eta^2r^2. \end{aligned}$$

Moreover, $\mathsf{P}(\mathcal{D}_\tau(\delta)) \geq 1 - \frac{(4t_f(\delta)+4)\delta}{T}$.

Proof. Throughout the proof, we assume that the event $\mathcal{D}_\tau(\delta)$ holds.

Let J denote $\{0, 1 \ldots, \tau - 1\}$ where $\tau < t_f(\delta)$. Then, J belongs to one of the two following cases.

Case 1) (Gradient dominates noise): Recall that this means that for every $t \in J$, we have

$$\|\nabla f(x_t)\| > 8t_f(\delta)\eta L \left(\left\| \frac{u}{2} \frac{1}{m} \sum_{i=1}^{m} Z_{t,i}Z_{t,i}^\top \tilde{H}_{t,i}Z_{t,i} \right\| + \|Y_t\| \right).$$

By our choice of η in Eq. (31), we can apply Lemma 16 to get

$$\min_{t \in J}\|\nabla f(x_t)\| \geq \frac{1}{4}\max_{t \in J}\|\nabla f(x_t)\|.$$

Thus by setting $\alpha = 128t_f(\delta)$ in Eq. (5) and by choosing η such that

$$\frac{c_1L\eta^2\chi^3 d}{m} \leq \frac{\eta}{\alpha} = \frac{\eta}{128t_f(\delta)} \quad \Longleftrightarrow \quad \eta \leq \frac{m}{128c_1Lt_f(\delta)d\chi^3},$$

it follows that

$$\begin{aligned} & -\frac{3\eta}{4}\sum_{t \in J}\left(\frac{1}{m}\sum_{i=1}^{m} Z_{t,i}^\top\nabla f(x_t)\right)^2 + \left(\frac{\eta}{128t_f(\delta)} + \frac{c_1L\eta^2\chi^3 d}{m}\right)\sum_{t \in J}\|\nabla f(x_t)\|^2 \\ ={} & -\frac{3\eta}{4}\sum_{t \in J}\frac{1}{m}\sum_{i=1}^{m}\left(Z_{t,i}^\top\nabla f(x_t)\right)^2 + \frac{\eta}{64t_f(\delta)}\sum_{t \in J}\|\nabla f(x_t)\|^2 \\ \leq{} & \frac{\eta}{64t_f(\delta)}\sum_{t \in J}\|\nabla f(x_t)\|^2 \\ \leq{} & \frac{\eta}{64t_f(\delta)}\sum_{t \in J}\max_{t \in J}\|\nabla f(x_t)\|^2 \\ \leq{} & \frac{16\eta}{64t_f(\delta)}\sum_{t \in J}\min_{t \in J}\|\nabla f(x_t)\|^2 \leq \frac{\eta}{4}\min_{t \in J}\|\nabla f(x_t)\|^2 \leq \frac{\eta}{4}\epsilon^2, \end{aligned} \tag{55}$$

where the final bound holds since we assumed $\min_{t \in J}\|\nabla f(x_t)\| \leq \epsilon$.

Case 2) (Gradient does not dominate noise): there exists some $t \in J$ such that

$$\|\nabla f(x_t)\| \le 8t_f(\delta)\eta L\left(\frac{u}{2}\left(\frac{1}{m}\sum_{i=1}^{m} Z_{t,i}Z_{t,i}^\top \tilde{H}_{t,i}Z_{t,i}\right) + \|Y_t\|\right).$$

By our choice of η in Eq. (31), we can apply Lemma 17 to get

$$\|\nabla f(x_t)\| \le c_5 t_f(\delta)\eta L\left(u^2d^2\rho\left(\log\frac{T}{\delta}\right)^2 + \left(1+\sqrt{\frac{\log(T/\delta)}{d}}\right)r\right) \qquad \forall t \in J.$$

Note that, by our choices of the parameters η, u, r, it can be shown that

$$c_5 t_f(\delta)\eta L\left(u^2d^2\rho\left(\log\frac{T}{\delta}\right)^2 + \left(1+\sqrt{\frac{\log(T/\delta)}{d}}\right)r\right) < \epsilon,$$

Hence, by setting $\alpha = 128t_f(\delta)$ in Eq. (5) and choosing η such that

$$\frac{c_1L\eta^2\chi^3d}{m} \le \frac{\eta}{\alpha} = \frac{\eta}{128t_f(\delta)},$$

it follows that

$$\begin{aligned}
&\left(\frac{\eta}{128t_f(\delta)} + \frac{c_1L\eta^2\chi^3d}{m}\right)\sum_{t\in J}\|\nabla f(x_t)\|^2 \\
&\le \frac{\eta}{64t_f(\delta)}\sum_{t\in J}\left(c_5t_f(\delta)\eta L\left(u^2d^2\rho\left(\log\frac{T}{\delta}\right)^2 + \left(1+\sqrt{\frac{\log(T/\delta)}{d}}\right)r\right)\right)^2 \\
&\le \frac{\eta}{64t_f(\delta)}\sum_{t\in J}\epsilon^2 \\
&\le \frac{\eta}{64}\epsilon^2
\end{aligned} \tag{56}$$

Combining both cases above (Eq. (55) and Eq. (56)), we see that for the choice $\alpha = 128t_f(\delta)$, the bound

$$-\frac{3\eta}{4}\sum_{t\in J}\left(\frac{1}{m}\sum_{i=1}^{m} Z_{t,i}^\top\nabla f(x_t)\right)^2 + \left(\frac{\eta}{128t_f(\delta)} + \frac{c_1L\eta^2\chi^3d}{m}\right)\sum_{t\in J}\|\nabla f(x_t)\|^2 \le \frac{\eta}{4}\epsilon^2 \tag{57}$$

always holds.

Recall by Eq. (5) that we have

$$\begin{aligned}
f(x_\tau) - f(x_0) \le\; & -\frac{3\eta}{4}\sum_{t=0}^{\tau-1}\left(\frac{1}{m}\sum_{i=1}^{m} Z_{t,i}^\top\nabla f(x_t)\right)^2 + \left(\frac{\eta}{\alpha} + \frac{c_1L\eta^2\chi^3d}{m}\right)\sum_{t=0}^{\tau-1}\|\nabla f(x_t)\|^2 \\
& + \tau\eta u^4\rho^2\cdot c_1d^3\left(\log\frac{T}{\delta}\right)^3 + \tau L\eta^2u^4\rho^2\cdot c_1d^4\left(\log\frac{T}{\delta}\right)^4 \\
& + \eta c_1r^2(\alpha+\eta L)\log\frac{T}{\delta} + \tau c_1L\eta^2r^2.
\end{aligned}$$

By plugging in Eq. (57) above, as well as the choice $\alpha = 128t_f(\delta)$, we see that

$$\begin{aligned}
f(x_\tau) - f(x_0) \le\; & \frac{\eta}{4}\epsilon^2 + t_f(\delta)\eta u^4\rho^2\cdot c_1d^3\left(\log\frac{T}{\delta}\right)^3 + t_f(\delta)L\eta^2u^4\rho^2\cdot c_1d^4\left(\log\frac{T}{\delta}\right)^4 \\
& + \eta c_1r^2(128t_f(\delta)+\eta L)\log\frac{T}{\delta} + t_f(\delta)c_1L\eta^2r^2.
\end{aligned}$$

We can now complete our proof by using the union bound (suppressing the dependence of some of the events on δ for notational simplicity) to derive

$$\begin{aligned}\mathsf{P}(\mathcal{D}_\tau^c) &\le \mathsf{P}(\mathcal{H}_\tau^c) + \sum_{t=0}^{\tau-1}\mathsf{P}(\mathcal{A}_t^c) + \sum_{t=0}^{\tau-1}\mathsf{P}(\mathcal{G}_t^c)\\ &\le \frac{(\tau+4)\delta}{T} + \frac{\tau}{T}\delta + 2\frac{\tau}{T}\delta \le \frac{(4t_f(\delta)+4)}{T}\delta\end{aligned}$$ □

Armed with Proposition 5 and Lemma 25, we are now ready to show for T sufficiently large, with high probability, there can be no more than $T/4$ ϵ-saddle points. Combined with Proposition 4, this yields the following result.

Theorem 2. *Suppose we pick u, r, η such that they satisfy the conditions in Proposition 5 and Lemma 25. Suppose F is chosen as prescribed in Lemma 25. Suppose that $\bar{\psi} \le 1$, so that $T_s \ge \frac{\iota}{\eta\bar{\psi}} \ge \frac{d}{mL}$[10]. Suppose we pick T_s as prescribed in Proposition 5. Suppose in addition we pick r such that*

$$r^2 \le \min\left\{\frac{\epsilon^2}{4(130c_1t_f(\delta) + c_1\log(T/\delta) + c_1)}, \frac{F\bar{\psi}}{80\iota\log(T/\delta)\left(\frac{65c_5^2}{8} + 132c_1 + 1\right)}\right\}.$$

Suppose also that we choose η such that

$$\eta \le \frac{0.1}{2\epsilon^2}\frac{\bar{\psi}}{2\iota}\frac{1}{2}\left(\frac{\bar{\psi}}{60c_9\iota\rho\log(T/\delta)}\right)^2\frac{1}{t_f(\delta)\left(129 + 8c'^2\beta_1(\delta;F)\right)\left(16(\mathrm{lr}(CT^2/\delta))^2 + 1\right)}$$

Suppose

$$T \ge \left(\frac{256t_f(\delta)\left(f(x_0) - f^*\right) + \epsilon^2/L)}{\eta\epsilon^2}, \frac{\varphi\rho^2\left(f(x_0) - f^*\right)}{\eta\bar{\psi}^4}, 256\lceil\frac{\iota}{\eta\bar{\psi}}\rceil, 256t_f(\delta), 1024\right), \tag{58}$$

where

$$\varphi := 20\left(2\iota^2(60c_9\iota\log(T/\delta))^2\left(t_f(\delta)\right)\left(129 + 8c'^2\beta_1(\delta;F)\right)\left(16(\mathrm{lr}(CT^2/\delta))^2 + 1\right)\right).$$

Then, with probability at least $1 - 22\delta$, there are at least $T/2$ ϵ-approximate second order stationary points.

Proof. Consider defining the following sequence of stopping times:

$$\begin{aligned}\tau_1 &= \inf_t\{t \le T : \|\nabla f(x_t)\| < \epsilon, \lambda_{\min}(\nabla^2 f(x_t)) \le -\sqrt{\rho\epsilon}\},\\ \tau_{i+1} &= \inf_t\{t \le T : t > \tau_i + T_s, \|\nabla f(x_t)\| < \epsilon, \lambda_{\min}(\nabla^2 f(x_t)) \le -\sqrt{\rho\epsilon}\}, \quad \forall 1 \le i \le \lfloor T/T_s\rfloor.\end{aligned}$$

□

We note that if $\tau_i = T$, then $\tau_j = T$ for any $j > i$. Let N_s denote the (random) number of saddle points encountered in T iterations.

We observe that we can decompose the function change as

$$\begin{aligned}&f(x_T) - f(x_0)\\ &= (f(x_{\tau_{N_s}}) - f(x_0)) + (f(x_T) - f(x_{\tau_{N_s}}))\end{aligned}$$

[10] Recall we focus on the case L, since otherwise, by the L-Lipschitz assumption, ${}_{\min}(r^2 f(x))$ L for all x 2 R^d, i.e. -first order stationary points are also -second order stationary points.

$$= (f(x_{\tau_1}) - f(x_0)) + \sum_{i=1}^{N_s} (f(x_{\tau_i+T_s}) - f(x_{\tau_i})) + \sum_{i=1}^{N_s-1} \left(f(x_{\tau_{i+1}}) - f(x_{\tau_i+T_s}) \right) + (f(x_T) - f(x_{\tau_{N_s}}))$$

$$= \underbrace{\sum_{i=1}^{N_s} (f(x_{\tau_i+T_s}) - f(x_{\tau_i}))}_{U_1} + \underbrace{(f(x_{\tau_1}) - f(x_0)) + \sum_{i=1}^{N_s-1} \left(f(x_{\tau_{i+1}}) - f(x_{\tau_i+T_s}) \right) + (f(x_T) - f(x_{\tau_{N_s}}))}_{U_2}.$$

We first consider U_1. Letting $x_j := x_T$ for any $j \geq T$, we have that

$$\sum_{i=1}^{N_s} f(x_{\tau_i+T_s}) - f(x_{\tau_i}) = \sum_{i=1}^{\lfloor T/T_s \rfloor} (f(x_{\tau_i+T_s}) - f(x_{\tau_i})) 1_{\tau_i<T}$$

Now, by Eq. (32), observe that with probability at least $1 - \frac{(5T_s+4)\delta}{T} \geq 1 - \frac{6T_s\delta}{T}$ (note $T_s \geq 4$), for any $1 \leq i \leq T/T_s$, we have that

$$\begin{aligned}
(f(x_{\tau_i+T_s}) - f(x_{\tau_i})) 1_{\tau_i<T} \leq{}& \tau \frac{c_5^2}{64} \eta^3 t_f(\delta)^2 L^2 \left(u^2 d^2 \rho \left(\log \frac{T}{\delta} \right)^2 + \sqrt{2 \log(T/\delta)}\, r \right)^2 \\
&+ \tau \eta u^4 \rho^2 \cdot c_1 d^3 \left(\log \frac{T}{\delta} \right)^3 + \tau L \eta^2 u^4 \rho^2 \cdot c_1 d^4 \left(\log \frac{T}{\delta} \right)^4 \\
&+ \eta c_1 r^2 (128 t_f(\delta) + \eta L) \log \frac{T}{\delta} + \tau c_1 L \eta^2 r^2 \\
&:= M_{u,r,T_s}.
\end{aligned}$$

Suppose we pick u, r such that $M_{u,r,T_s} \leq 0.1F$. Recall from Proposition 5 that with probability at least $1/3 - \frac{13T_s\delta}{T}$, $(f(x_{\tau_i+T_s}) - f(x_{\tau_i})) 1_{\tau_i<T} \leq -F$. Choosing δ such that $1/3 - \frac{13T_s\delta}{T} \geq 0.3$, and letting $\mu = 0.1F$, we note that $|-F + \mu| = 0.9F \geq \frac{0.7}{0.3} 0.2F \geq \frac{0.7}{0.3}(M_{u,r,T_s} + \mu)$.

Now, let $\mathcal{E}_{\tau_i}$ denote the bad event on which

$$\text{neither } (f(x_{\tau_i+T_s}) - f(x_{\tau_i})) 1_{\tau_i<T} \leq -F, \quad \text{nor } (f(x_{\tau_i+T_s}) - f(x_{\tau_i})) 1_{\tau_i<T} \leq M_{u,r,T_s} \leq 0.1F.$$

We know that $\mathcal{E}_{\tau_i}$ has probability at most $\frac{6T_s\delta}{T}$. Let $\mathcal{E}_\tau := \cup_{i=1}^{\lfloor T/T_s \rfloor} \mathcal{E}_{\tau_i}$, such that $\mathsf{P}(\mathcal{E}_\tau) \leq 6\delta$. Then, by applying the weakened supermartingale inequality in Proposition 3, we have

$$\mathsf{P}\left(\sum_{i=1}^{T/T_s} (f(x_{\tau_i+T_s}) - f(x_{\tau_i})) 1_{\tau_i<T} \geq -N_s 0.9F + s \right) \leq \mathsf{E}\left[\exp\left(-\frac{s^2}{4N_sF^2} \right) \right] + \mathsf{P}(\mathcal{E}_\tau) \leq \exp\left(-\frac{s^2}{4(T/T_s)F^2} \right) + 6\delta.$$

Now, pick $s = 2F \sqrt{\sqrt{\log(1/\delta)}\, T/T_s}$, then

$$\mathsf{P}\left(\sum_{i=1}^{T/T_s} (f(x_{\tau_i+T_s}) - f(x_{\tau_i})) 1_{\tau_i<T} \geq -N_s 0.9F + 2F \sqrt{\sqrt{\log(1/\delta)}\, T/T_s} \right) \leq 7\delta.$$

Note that supposing for contradiction that there are at least $T/4$ saddles, we must then have that $N_s \geq T/(4T_s)$, such that

$$-N_s 0.9F + 2F \sqrt{\sqrt{\log(1/\delta)}\, T/T_s)} \leq F(-0.9T/(4T_s) + (2\sqrt{\sqrt{\log(1/\delta)}\, T/T_s})) \leq F(-0.1T/T_s),$$

where we may ensure the last inequality by picking T/T_s such that

$$T/T_s \geq \left(\frac{2}{0.125} \right)^2 \sqrt{\log(1/\delta)} = 256 \sqrt{\log(1/\delta)}.$$

Note that our choice of T ensures this.

Thus, with probability at least $1 - 7\delta$,

$$U_1 = \sum_{i=1}^{T/T_s} \left(f(x_{\tau_i+T_s}) - f(x_{\tau_i})\right) 1_{\tau_i<T} \leq -(0.1T/T_s)F.$$

Next, we bound the summand U_2. Recall that

$$U_2 = \left(f(x_{\tau_1}) - f(x_0)\right) + \sum_{i=1}^{N_s-1} \left(f(x_{\tau_{i+1}}) - f(x_{\tau_i+T_s}) \right).$$

Without loss of generality, we may analyze each of the summands $f(x_{\tau_{i+1}}) - f(x_{\tau_i+T_s})$ in the same way as we treat $(f(x_{\tau_1}) - f(x_0))$. Let us then consider the summand $f(x_{\tau_1}) - f(x_0)$. There are two cases to consider.

1. The first is when $\tau_1 < t_f(\delta)$. In this case, since we know that $\|\nabla f(x_{\tau_1})\| \leq \epsilon$ (as x_{τ_1} is an ϵ-saddle point), it follows by Lemma 26 that

$$\begin{aligned} f(x_{\tau_1}) - f(x_0) \leq \frac{\eta}{4}\epsilon^2 &+ t_f(\delta)\eta u^4\rho^2 \cdot c_1 d^3 \left(\log\frac{T}{\delta}\right)^3 + t_f(\delta)L\eta^2 u^4\rho^2 \cdot c_1 d^4 \left(\log\frac{T}{\delta}\right)^4 \\ &+ \eta c_1 r^2(128 t_f(\delta) + \eta L)\log\frac{T}{\delta} + t_f(\delta)c_1 L\eta^2 r^2 \end{aligned}$$

   with probability at least $1 - \frac{(4t_f(\delta)+4)\delta}{T}$.

2. The second case is when $\tau_1 \geq t_f(\delta)$. In this case, by Lemma 18, we have that

$$\begin{aligned} f(x_{\tau_1}) - f(x_0) \leq \tau_1 \frac{c_5^2}{64}\eta^3 t_f(\delta)^2 L^2 &\left(u^2 d^2 \rho \left(\log\frac{T}{\delta}\right)^2 + \sqrt{2\log(T/\delta)}\,r \right)^2 \\ &+ \tau_1 \eta u^4\rho^2 \cdot c_1 d^3 \left(\log\frac{T}{\delta}\right)^3 + \tau_1 L\eta^2 u^4\rho^2 \cdot c_1 d^4 \left(\log\frac{T}{\delta}\right)^4 \\ &+ \eta c_1 r^2(128 t_f(\delta) + \eta L)\log\frac{T}{\delta} + \tau_1 c_1 L\eta^2 r^2. \end{aligned}$$

   with probability at least $1 - \frac{(5\tau_1+4)\delta}{T}$.

By our choice of u, we know that

$$\begin{aligned} &t_f(\delta)\eta u^4\rho^2 \cdot c_1 d^3 \left(\log\frac{T}{\delta}\right)^3 + t_f(\delta)L\eta^2 u^4\rho^2 \cdot c_1 d^4 \left(\log\frac{T}{\delta}\right)^4 + \eta c_1 r^2(128 t_f(\delta) + \eta L)\log\frac{T}{\delta} + t_f(\delta)c_1 L\eta^2 r^2 \\ &\leq t_f(\delta) r^2 c_1 + t_f(\delta) r^2 c_1 + c_1 r^2 (128 t_f(\delta) + 1)\log(T/\delta) + c_1 r^2 \\ &= r^2(130 c_1 t_f(\delta) + c_1 \log(T/\delta) + c_1). \end{aligned}$$

Hence, by picking r such that

$$r \leq \frac{\epsilon^2}{4(130 c_1 t_f(\delta) + c_1\log(T/\delta) + c_1)},$$

it follows that

$$\begin{aligned} \frac{\eta\epsilon^2}{4} \geq t_f(\delta)\eta u^4\rho^2 \cdot c_1 d^3 \left(\log\frac{T}{\delta}\right)^3 &+ t_f(\delta)L\eta^2 u^4\rho^2 \cdot c_1 d^4 \left(\log\frac{T}{\delta}\right)^4 \\ &+ \eta c_1 r^2(128 t_f(\delta) + \eta L)\log\frac{T}{\delta} + t_f(\delta)c_1 L\eta^2 r^2. \end{aligned}$$

Then, if $\tau_1 < t_f(\delta)$, with probability at least $1 - \frac{(5t_f(\delta)+4)}{\delta}$,

$$f(x_{\tau_1}) - f(x_0) \leq \frac{\eta\epsilon^2}{2}.$$

Suppose also that we pick r such that

$$r^2 \leq \frac{F\sqrt{\rho\epsilon}}{80\iota\log(T/\delta)\left(\frac{65c_5^2}{8} + 132c_1 + 1\right)} \leq \frac{F}{40\eta T_s\log(T/\delta)\left(\frac{65c_5^2}{8} + 132c_1 + 1\right)}.$$

Then, it can be verified that

$$\begin{aligned}\frac{F}{40}\frac{T}{T_s} \geq\ & T\frac{c_5^2}{64}\eta^3 t_f(\delta)^2 L^2\left(u^2 d^2\rho\left(\log\frac{T}{\delta}\right)^2 + \sqrt{2\log(T/\delta)}r\right)^2 \\ & + T\eta u^4\rho^2\cdot c_1 d^3\left(\log\frac{T}{\delta}\right)^3 + TL\eta^2 u^4\rho^2\cdot c_1 d^4\left(\log\frac{T}{\delta}\right)^4 \\ & + \frac{T}{T_s}\eta c_1 r^2(128t_f(\delta) + \eta L)\log\frac{T}{\delta} + Tc_1 L\eta^2 r^2.\end{aligned}$$

Then, by a union bound, it follows that with probability at least $1 - 9\delta$,

$$\begin{aligned}U_2 &= (f(x_{\tau_1}) - f(x_0)) + \sum_{i=1}^{N_s - 1}\left(f(x_{\tau_{i+1}}) - f(x_{\tau_i + T_s})\right) \\ &\leq \frac{T}{T_s}\frac{\eta\epsilon^2}{2} + \frac{F}{40}\frac{T}{T_s}\end{aligned}$$

Therefore, by the union bound, with probability at least $1 - 16\delta$,

$$f(x_{\tau_{N_s}}) - f(x_0) = U_1 + U_2 \leq \frac{T}{T_s}\left(-0.1F + \eta\epsilon^2/2 + \frac{F}{40}\right)$$

By recalling our choice of F in Lemma 25, by choosing η such that

$$\begin{aligned}\eta &\leq \frac{0.1}{2\epsilon^2}\frac{\bar{\psi}}{2\iota}\frac{1}{2}\left(\frac{\bar{\psi}}{60c_9\iota\rho\log(T/\delta)}\right)^2\frac{1}{t_f(\delta)\left(129 + 8c'^2\beta_1(\delta; F)\left(16(\mathrm{lr}(CT^2/\delta))^2 + 1\right)\right)} \\ &\leq \frac{0.1}{2\epsilon^2}\frac{1}{2}\left(\frac{\sqrt{\rho\epsilon}}{60c_9\iota\rho\log(T/\delta)}\right)^2\frac{1}{\eta T_s t_f(\delta)\left(129 + 8c'^2\beta_1(\delta; F)\left(16(\mathrm{lr}(CT^2/\delta))^2 + 1\right)\right)} = \frac{0.1F}{2\epsilon^2},\end{aligned}$$

it follows that with probability at least $1 - 16\delta$,

$$\begin{aligned}f(x_{\tau_{N_s}}) - f(x_0) &= U_1 + U_2 \\ &\leq \frac{T}{T_s}\left(-0.1F + \eta\epsilon^2/2 + \frac{F}{40}\right) \\ &\leq \frac{T}{T_s}(-0.1F + 0.1F/4 + 0.1F/4) = \frac{T}{T_s}(-0.05F).\end{aligned}$$

Choose T such that

$$-(0.05T/T_s)F \leq -(f(x_0) - f^*) \iff T \geq \frac{20T_s(f(x_0) - f^*)}{F} \geq \frac{\varphi\rho^2\left(f(x_0) - f^*\right)}{\eta\bar{\psi}^4}$$

yields a contradiction, where

$$\varphi := 20\left(2\iota^2(60c_9\iota\log(T/\delta))^2\left(t_f(\delta)\right)\left(129 + 8c'^2\beta_1(\delta; F)\left(16(\mathrm{lr}(CT^2/\delta))^2 + 1\right)\right)\right)$$

Hence, with probability at least $1 - 16\delta$, there cannot be more than $T/4$ saddle points. In addition, with probability at least $1 - 6\delta$, by Proposition 4, there cannot be more than $T/4$ iterates with $\|\nabla f(x_t)\| \geq \epsilon$. Hence, with probability at least $1 - 22\delta$, there are at least $T/2$ ϵ-approximate second order stationary points.

G. More complete discussion of simulations

We test the performance of our proposed algorithm with two-point estimators (ZOPGD-2pt) against existing zeroth-order benchmarks using the *octopus function* (proposed in (Du et al., 2017)) of varying dimensions.[11] It is known that the octopus function defined on R^d, which chains d saddle points sequentially, takes exponential (in d) time for exact gradient descent to escape; it has thus emerged as a popular benchmark to evaluate and compare the performance of algorithms that seek to escape saddle points. In our experiments, we compare the performance of our two-point estimator algorithm (ZOPGD-2pt) with PAGD (Algorithm 1 in (Vlatakis-Gkaragkounis et al., 2019)) and ZO-GD-NCF (see (Zhang et al., 2022)), which are the only two existing zeroth-order algorithms that have (a) a $\tilde{O}(d/\epsilon^2)$ sample complexity for escaping saddle points (with the latter algorithm yielding the tightest bounds), and (b) performed the best empirically on escaping saddle points (see the simulation results in (Zhang et al., 2022)). We note that both PAGD and ZO-GD-NCF have to use $2d$ function evaluations per iteration to estimate the gradient while our algorithm only needs to use 2 function evaluations. In our plots, we plot the function value against the number of function evaluations. For completeness, we also plot the performance of exact gradient descent (normalized such that its x-axis is also the number of function queries).

We tested the algorithms for $d = 10$ and $d = 30$. To account for the stochasticity in the algorithms, for each algorithm, we computed the average and standard deviation over 30 trials, and plotted the mean trajectory with an additional band that represents 1.5 times the standard deviation. For our algorithmś hyperparameters, we picked

$$\eta = \frac{1}{4dL}, u = 10 \quad ^{2}, r = 0.05, m = 1(\text{ i.e. two-point estimator}) \tag{59}$$

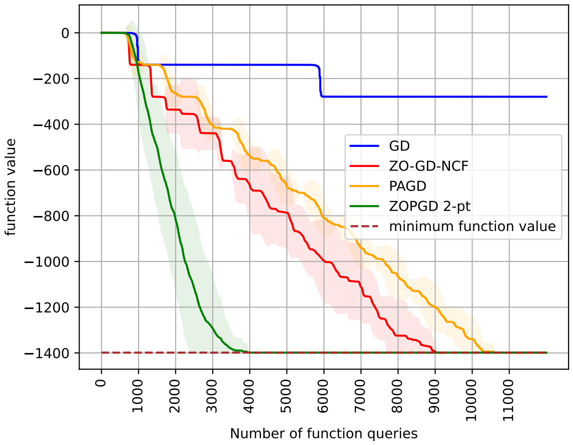
For PAGD, we used the hyperparameters listed in their paper, and for ZO-GD-NCF, we used the code from their Neurips submission. We note in particular that both methods used the step-size $\frac{1}{4L}$. For initialization, we chose a random x_0 near the saddle point at the origin, drawn from $N(0, 10 \quad ^3 I_d \quad _d)$[12] (fixed for all trials and all algorithms).

As we can see in Fig. 2, in both cases, our algorithm reaches the global minimum of the octopus function in significantly fewer function evaluations than PAGD and ZO-GD-NCF (approximately 2.5 times faster than ZO-GD-NCF, and approximately 3 times faster than PAGD), despite our algorithm only using 2 function evaluations per iteration compared to $2d$ function evaluations per iteration for both PAGD and ZO-GD-NCF. As a sanity check, we note that the number of function evaluations required for PAGD and ZO-GD-NCF to reach the global minimum approximately matches that in Figure 1 of (Zhang et al., 2022); here the correspondence is only approximate since (Zhang et al., 2022) only plots one trial while we compute the mean and standard deviation of 30 trials.

This result suggests that in addition to the theoretical convergence guarantees, there might also be empirical benefits to using two-point estimators versus existing $2d$-point estimators in the zeroth-order escaping saddle point literature.

[11] Our code can be found at `https://github.com/rafflesintown/escape-saddle-points-2pt`

[12] Using the random seed in our code, we note that $k\ r\ f(x_0)k = 0{:}011$ for $d = 10$ and $k\ r\ f(x_0)k = 0{:}030$ for $d = 30$.



(a) d = 10

(b) d = 30

Figure 2: Performance on toy octopus function, with $\tau = e, L = e, \gamma = 1$ (Here, τ, L, γ are parameters determining the properties of f. Our parameter choice is consistent with that in (Zhang et al., 2022). See (Du et al., 2017) for details about the definitions of τ, L and γ.).