ESCAPING SADDLE POINTS IN ZEROTH-ORDER OPTIMIZATION: TWO FUNCTION EVALUATIONS SUFFICE

Anonymous authors
Paper under double-blind review

ABSTRACT

Zeroth-order methods are useful in solving black-box optimization and reinforcement learning problems in unknown environments. It uses function values to estimate the gradient. As optimization problems are often nonconvex, it is a natural question to understand how zeroth-order methods escape saddle points. In this paper, we consider zeroth-order methods, that at each iteration, may freely choose $2m$ function evaluations where m ranges from 1 to d, with d denoting the problem dimension. We show that by adding an appropriate isotropic perturbation at each iteration, a zeroth-order algorithm based on $2m$ function evaluations per iteration can not only find ϵ-second order stationary points polynomially fast, but do so using only $\tilde{O}(d/\epsilon^{2.5})$ function evaluations.

1 INTRODUCTION

Zeroth-order methods, which refer to algorithms that optimize a function using only function evaluations, are an important tool in solving black-box optimization tasks and reinforcement learning problems in unknown environments. While zeroth-order optimization takes many forms (Torczon, 1997; Agarwal et al., 2013), a common approach is to utilize function values to build stochastic estimates of the gradients (Duchi et al., 2015; Flaxman et al., 2005; Nesterov and Spokoiny, 2017; Shamir, 2017), which is also the approach we will study. In many applications of zeroth-order optimization, the underlying function may be nonconvex. For nonconvex functions, under various smoothness and Lipschitz assumptions on the function, existing works have studied the convergence of zeroth-order methods to first-order stationary points (Nesterov and Spokoiny, 2017). Since finding globally optimal solutions in non-convex optimization is NP-hard, seeking first-order stationary points is a reasonable starting point for theoretical analysis. However, the presence of possible saddle points for nonconvex functions means that merely converging to first-order stationary points is insufficient. Ideally, we would like that zeroth-order methods can evade such saddle points, and instead find second-order stationary points.

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; 2018). 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; thus, to reliably evade saddle points using stochastic gradients with high probability, an additional isotropic perturbation at each iteration is often still required (Jin et al., 2019a). Even when additional perturbation is not required, 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 and Sayed, 2021b), creating an analytical issue when such idealized assumptions fail to hold. Indeed, common zeroth order estimators, such as two-point estimators (Nesterov and Spokoiny, 2017), are not subGaussian, and can have unbounded variance. For instance, for two-point zeroth-order finite-difference gradient estimators, it can be shown that the variance of the estimator is on the order of $\Omega(d\|\nabla f(x)\|^2)$ (Nesterov and Spokoiny, 2017), with both a

		Iteration Complexity	Fun. evaluations per iter.
First-order	Jin et al. (2017) (deterministic)	$\tilde{O}\left(\frac{1}{\epsilon^2}\right)$	—
	Fang et al. (2019) (SGD)	$\tilde{O}\left(\frac{1}{\epsilon^{3.5}}\right)$	—
Zeroth-order	Bai et al. (2020)	$\tilde{O}\left(\frac{1}{\epsilon^2}\right)$	$\tilde{\Omega}\left(\frac{d^2}{\epsilon^8}\right)$
	Flokas et al. (2019)	$\tilde{O}\left(\frac{1}{\epsilon^2}\right)$	$\tilde{\Omega}(d)$
	Balasubramanian and Ghadimi (2022)	$\tilde{O}\left(\frac{1}{\epsilon^{3/2}}\right)$	$\tilde{O}\left(\frac{d}{\epsilon^2}+\frac{d^4}{\epsilon}\right)$
	Algorithm 1 (this paper)	$\tilde{O}\left(\frac{d}{\epsilon^{2.5}m}\right)$	$2m$

Table 1: Selected comparison of convergence results to ϵ-second order stationary points in smooth, nonconvex functions

dependence on the problem dimension d as well as on norm of the gradient, which can be unbounded. Due to these properties (non-subGaussianity and unboundedness), it is tricky to bound the effect of such zeroth-order estimators and establish tight concentration inequalities that facilitate escape near saddle points. In addition, the large variance of the zeroth-order estimator is also an issue away from saddles when the gradient is large. While this is not an issue to show function improvement in expectation (which existing zeroth order works showing first-order stationarity have focused on), 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 (Bai et al., 2020; Flokas et al., 2019) or in some cases also the Hessian (Balasubramanian and Ghadimi, 2022), reducing in a sense the zeroth-order problem back to a first-order one. However, from existing works (Nesterov and Spokoiny, 2017), for smooth nonconvex functions, we know that two-point zeroth-order estimators suffice to find first-order ϵ-stationary points at a rate of $\tilde{O}(\frac{d}{\epsilon^2})$. This raises an important question:

Can zeroth-order methods which use fewer than $\Omega(d)$ function evaluations per iteration 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 (m ranging from 1 to d) of function evaluations per iteration can not only find ϵ-second order stationary points (cf. the definition later in Definition 1) polynomially fast, but do so using only $\tilde{O}(\mathrm{polylog}(\frac{1}{\delta})d/\epsilon^{2.5})$ 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 (and adding appropriate perturbation) suffices to efficiently escape saddle points in zeroth-order optimization, with high probability. A comparison of our results with existing zeroth-order and first order methods is shown in Table 1. 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. To the best of our knowledge, such an analysis is novel, and may be of general interest to researchers working in zeroth-order optimization. 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 the effect of 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. These technical bounds may be of independent interest.

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 $f:\mathbb{R}^d\to\mathbb{R}$ which we consider.

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

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

2. *f is L-Lipschitz, i.e.* $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \ \ \forall x, y \in \mathbb{R}^d$
3. *f is ρ-Hessian Lipschitz, i.e.* $\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\| \leq \rho\|x - y\| \ \ \forall x, y \in \mathbb{R}^d$.

In our work, we concern ourselves with finding approximate second order stationary points. Following past convention (Jin et al., 2019a), we define ϵ-second order stationary points as follows.

Definition 1. A point $x \in \mathbb{R}^d$ is an ϵ-second order stationary point if

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

We define an ϵ-approximate saddle point as follows.

Definition 2. A point $x \in \mathbb{R}^d$ is an ϵ-approximate saddle point if

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

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

Definition 3 ((Batch) two-point zeroth-order estimator with perturbation)**.** We define a m-batch two-point zeroth order estimator as follows:

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

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

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

Algorithm 1: Zeroth-order perturbed gradient descent (ZOPGD)

input : x_0, horizon T, step-size η, smoothing radius u, perturbation radius r, batch size m
for *step* $t = 0, \ldots, T_1$ **do**
  Sample $Z^{(m)} = \{Z_{t,i}\}_{i=1}^m \sim N(0, I)$ to compute $g_u^{(m)}(x_t))$.
  Update $x_{t+1} = x_t - \eta\left(g_u^{(m)}(x_t) + Y_t\right)$, where $Y_t \sim N(0, \frac{r^2}{d}I)$

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]*. Suppose*

$$u = \tilde{O}\left(\frac{\min\{\sqrt{\epsilon}, \sqrt{r}\}}{\sqrt{\rho}d}\right), \qquad r = \tilde{O}(\epsilon), \qquad \eta = \tilde{O}\left(\frac{m\sqrt{\rho\epsilon}}{d\max\{L, L^2\}}\right).$$

Then, in $T = \tilde{\Omega}(\frac{d\max\{L^2, L\}(f(x_0) - f^*)}{m\sqrt{\rho}\epsilon^{2.5}})$ iterations (with each iteration using $2m$ function evaluations), with probability at least $1 - \delta$, at least half the iterates are ϵ-approximate second-order stationary points.

Comparison to gradient-based methods. For first-order escape saddle point algorithms, standard perturbation-based methods (without acceleration) can find a ϵ-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 $\tilde{O}\left(\frac{d}{\epsilon^{2.5}}\right)$. The extra dependence on d is typical for zeroth-order algorithms (see e.g. Nesterov and 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. Our dependence on ϵ sits between that of the

[1] In our paper, we focus on the case $\sqrt{\rho\epsilon} \leq L$; otherwise, by the L-Lipschitz assumption, $\lambda_{\min}(\nabla^2 f(x)) \geq -L$ for all $x \in \mathbb{R}^d$, which implies ϵ-first order stationary points are also ϵ-second order stationary points.

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.

Comparison to existing zeroth-order methods. As Table 1 suggests, our sample complexity significantly outperforms that of Bai et al. (2020) and Balasubramanian and Ghadimi (2022). But the sample complexity in Flokas et al. (2019) outperforms our method, with a function evaluation complexity of $\tilde{O}\left(\frac{d}{\epsilon^2}\right)$. However, a key limitation of their method 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. Another limitation is that the method in Flokas et al. (2019) requires checking if the gradient is small and adding a perturbation accordingly. In contrast, our method requires only the addition of an isotropic perturbative noise at each iteration, which is significantly more straightforward from an implementation perspective. Moreover, staying true to the spirit of zeroth-order algorithms, our method supports any number of function evaluations at each iteration between 1 to d.

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

Broadly speaking, our proof shows that progress can be made both in iterations when the gradient is large (which we can define to be iterations t where $\|\nabla f(x_t)\| \geq \epsilon$), and iterations when we are at an ϵ-approximate saddle point (where progress may be made along the negative eigendirection of the Hessian matrix). While the approach is similar to the first-order case (e.g. Jin et al. (2019a)), the zeroth-order setting brings forth several challenges which we explain below.

First, due to the noise in the 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 escape saddle point purposes (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 far from straightforward due to the large variance of the zeroth-order estimator (scaling with d times the squared norm of the gradient) and hence any single iteration of the zeroth-order descent method may in fact lead to a function increase rather than decrease. Our high-probability bound on the function change for our zeroth-order algorithm (see Proposition 1 and subsequent discussion in the main text for intuition, and Lemma 18 in the appendix for the concrete bound) are novel, and may be of more general interest.

Second, the noise in the zeroth-order estimator also makes the analysis around $\epsilon-$approximate saddle points challenging because the concentration properties of the noise is hard to characterize. While intuitively a more noisy estimator should facilitate escape from saddle point, without a concentration bound on the noise, the noise may behave in unpredictable ways that prevent escape from the saddle region. 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. We overcome this issue by showing that while the zeroth-order estimators considered in our paper are merely subExponential and not subGaussian, we can leverage the fact that they are the product of subGaussian vectors to derive a tight concentration bound on the vector sums of such estimators.

We detail these challenges as well as our main technical results in the rest of this section. Note that due to the space limit, the full proof is provided in the Appendix.

3.1 Showing function decrease when gradients are large

To show that the function can make progress in the large gradient iterations, we first introduce the following function decrease lemma. Note that at each iteration, we access the function oracle $2m$ times, where m is a batch-size parameter we can control. The proof can be found in Appendix D.

Lemma 1 (Function decrease for batch zero-order optimization)**.** *Suppose at each time t, the algorithm performs the update step*

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

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 $c_1 > 0, C_1 \geq 1$ such that, for any $T \in \mathbb{Z}^+$ and $T \geq \tau' \geq \tau > 0$, $\alpha > 0$ and $\delta \in (0, 1/e]$, upon defining $\mathcal{H}_{0,\tau}(\delta)$ to be the event on which the inequality

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

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

$$\mathbb{P}(\mathcal{H}_{0,\tau}(\delta)) \geq 1 - \frac{(\tau+4)\delta}{T}, \qquad \mathbb{P}(\cap_{\tau=1}^{\tau'} \mathcal{H}_{0,\tau}(\delta)) \geq 1 - \frac{5\tau'\delta}{T}$$

The next steps are 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 $\|\nabla f(x_t)\| \geq \epsilon$. Roughly speaking, based on Eq. (2), we need to prove a lower bound of the form

$$\sum_{t=0}^{T-1} \frac{1}{m} \sum_{i=1}^{m} \left\| Z_{t,i}^\top \nabla f(x_t) \right\|^2 \geq \Omega \left(\frac{1}{\alpha} + \frac{c_1 L \eta \chi^3 d}{m} \right) \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 \tag{3}$$

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, $\frac{1}{m} \sum_{i=1}^{m} \left\| Z_{t,i}^\top \nabla f(x_t) \right\|^2$ 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. (21). We do so using the following approach.

1. Intuitively, though for each individual iteration t, $\frac{1}{m} \sum_{i=1}^{m} \left\| Z_{t,i}^\top \nabla f(x_t) \right\|^2$ could be small even as $\|\nabla f(x_t)\|^2$ is large, in a small number of (consecutive) iterations $\{t_0, \ldots, t_0 + t_f\}$, with high probability, there will be at least one iteration t within $\{t_0, \ldots, t_0 + t_f - 1\}$, such that $\frac{1}{m} \sum_{i=1}^{m} \left\| Z_{t,i}^\top \nabla f(x_t) \right\|^2 = \Omega(\|\nabla f(x_t)\|^2)$. We formalize this intuition in Lemma 14 in Appendix. Thus, we consider breaking the time-steps into chunks where each chunk has t_f consecutive iterations.
2. Consider any such interval $\{t_0, \ldots, t_0 + t_f - 1\}$. 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 η, we can show that

   $$\frac{1}{2} \|\nabla f(x_{t_0})\| \leq \|\nabla f(x_t)\| \leq 2 \|\nabla f(x_{t_0})\| \qquad \forall t \in \{t_0, \ldots, t_0 + t_f - 1\}.$$

   As a result, combined with Point 1 discussed above, we see that

   $$\sum_{t=t_0}^{t_0+t_f-1} \frac{1}{m} \sum_{i=1}^{m} \left\| Z_{t,i}^\top \nabla f(x_t) \right\|^2 \geq \Omega(\|\nabla f(x_{t_0})\|^2).$$

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

   $$\begin{aligned} \sum_{t=t_0}^{t_0+t_f-1} \frac{1}{m} \sum_{i=1}^{m} \left\| Z_{t,i}^\top \nabla f(x_t) \right\|^2 \geq \Omega(\|\nabla f(x_{t_0})\|^2) \geq & \Omega \left(\frac{1}{\alpha} + \frac{c_1 L \eta \chi^3 d}{m} \right) \sum_{t=t_0}^{t_0+t_f-1} \|\nabla f(x_t)\|^2 \\ = & \Omega \left(\frac{1}{\alpha} + \frac{c_1 L \eta \chi^3 d}{m} \right) \Omega \left(t_f \|\nabla f(x_{t_0})\|^2 \right) \end{aligned}$$

   Thus, in these intervals, we obtain function improvement on the order of $\eta \Omega(\|\nabla f(x_{t_0})\|^2)$.

(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, u and η 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 $\eta\Omega(\epsilon^2)$ for such intervals. By appropriately choosing η, u and r to limit the effects of the intervals of the case (b), we can then show a contradiction of the form $f(x_T) < f^*$. We demonstrate this formally in Proposition 5.

We formalize our approach in Appendix D. 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 5)**.** *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*

$$u = \tilde{O}\left(\frac{\sqrt{\epsilon}}{\sqrt{\rho}d}\right), \qquad r = O(\epsilon), \qquad \eta = \tilde{O}\left(\frac{m}{dL}\right), \qquad T = \tilde{\Omega}\left(\frac{\left((f(x_0) - f^*) + \epsilon^2/L)\right)}{\eta\epsilon^2}\right).$$

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

3.2 Making progress near saddle points

The previous subsection analyze how the algorithm makes progresses when gradient is large. Now we focus on the case near saddle points. 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*

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

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. Building on this technical result, we next 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 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 ϕ 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 sequences will (with some constant probability) grow exponentially large, eventually moving out of their specified radius ϕ within T_s iterations, leading to a contradiction. We note that while our

proof approach is not new (see e.g. Jin et al. (2019a)), the application to the zeroth-order setting poses significant technical challenges, which we explain in the discussion following our next result, Lemma 3. 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 often 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, 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 $j \neq 2$. 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*

$$x_{t+1}' = x_t' - \eta\left(\frac{1}{m}\sum_{i=1}^{m}\left(Z_{t,i}Z_{t,i}^\top \nabla f(x_t') + \frac{u}{2}Z_{t,i}Z_{t,i}^\top \tilde{H}_{t,i}' Z_{t,i}\right) + Y_t'\right), \quad x_0' = x_0,$$

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

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

where

$$\xi_{g_0}(t) = \frac{1}{m}\sum_{i=1}^{m}(Z_{t,i}Z_{t,i}^\top - I)\nabla f(x_t),\ \xi_{g_0}'(t) = \frac{1}{m}\sum_{i=1}^{m}(Z_{t,i}(Z_{t,i})^\top - I)\nabla f(x_t'),\ \hat{\xi}_{g_0}(t) = \xi_{g_0}(t) - \xi_{g_0}'(t),$$

$$\xi_u(t) = \frac{1}{m}\sum_{i=1}^{m}\frac{u}{2}Z_{t,i}Z_{t,i}\tilde{H}_{t,i}Z_{t,i}, \quad \xi_u'(t) = \frac{1}{m}\sum_{i=1}^{m}\frac{u}{2}Z_{t,i}Z_{t,i}\tilde{H}_{t,i}'Z_{t,i}, \quad \hat{\xi}_u(t) = \xi_u(t) - \xi_u'(t),$$

$$\hat{Y}_t = Y_t - Y_t', \quad \bar{H}_t = \int_0^1 \nabla^2 f(ax_t + (1-a)x_t')da.$$

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

$$W_{g_0}(t+1) = -\eta\sum_{\tau=0}^{t}(I-\eta H)^{t-\tau}\left(\frac{1}{m}\sum_{i=1}^{m}(Z_{\tau,i}Z_{\tau,i}^\top - I)(\nabla f(x_\tau) - \nabla f(x_\tau'))\right).$$

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,

$$\|W_{g_0}(t)\| \lesssim \eta\sqrt{\sum_{\tau=0}^{t-1}(1+\eta\gamma)^{2(t-1-\tau)}\mathbb{E}\left[\|(Z_\tau Z_\tau - I)(\nabla f(x_\tau) - \nabla f(x_\tau'))\|^2 \mid \mathcal{F}_{\tau-1}\right]} \tag{4}$$

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' are both in $\mathcal{F}_{\tau-1}$, but Z_τ is not. Then, assuming that Eq. (4) holds, since

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

it follows that

$$\|W_{g_0}(t)\| \leq \eta \sqrt{O(d) \sum_{\tau=0}^{t-1} (1+\eta\gamma)^{2(t-1-\tau)} \|\nabla f(x_\tau) - \nabla f(x'_\tau)\|^2}$$

With this bound on $\|W_{g_0}(t)\|$, we eventually prove in Proposition 6 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 and Spokoiny, 2017), and has been conjectured to be the best possible dimension dependence for general smooth nonconvex zeroth-order optimization (Balasubramanian and Ghadimi, 2022).

Key technical challenge The key challenge in the above argument is to show that an equation in the form of Eq. (4) could in fact hold. At first glance, that an inequality such as Eq. (4) should hold is rather non-obvious — this is because while the variable $(Z_\tau Z_\tau - I)(\nabla f(x_\tau) - \nabla f(x'_\tau)) \mid \mathcal{F}_{\tau-1}$ 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

$$\left\| \sum_{\tau=0}^{t-1} \boldsymbol{x}_\tau \right\| \approx \tilde{O}\left(\sqrt{\sum_{\tau=0}^{t-1} \sigma_\tau^2} \right)$$

is a very delicate matter. In our case, the analogue of $\boldsymbol{x}_\tau$ is $(I - \eta H)^{t-1-\tau}(Z_\tau Z_\tau - I)(\nabla f(x_\tau) - \nabla f(x'_\tau))$, while the analogue of σ_τ^2 is $(1 + \eta\gamma)^{2(t-1-\tau)} \mathbb{E}\left[\|(Z_\tau Z_\tau - I)\left(\nabla f(x_\tau) - \nabla f(x'_\tau)\right)\|^2 \mid \mathcal{F}_{\tau-1} \right]$. 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 $\mathbb{E}[e^{\theta \boldsymbol{Y}_\tau} \mid \mathcal{F}_{\tau-1}]$ to be defined for any fixed (and non-random) $\theta > 0$, where $\boldsymbol{Y}_\tau$ takes the form

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

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), $\mathbb{E}[e^{\theta \boldsymbol{Y}_\tau} \mid \mathcal{F}_{\tau-1}]$, 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 $(I - \eta H)^{t-1-\tau}(Z_\tau Z_\tau^\top - I)(\nabla f(x_\tau) - \nabla f(x'_\tau))$, which is subExponential rather than subGaussian. While it may be possible to force $(I - \eta H)^{t-1-\tau}(Z_\tau Z_\tau^\top - I)(\nabla f(x_\tau) - \nabla f(x'_\tau))$ to be sub-Gaussian, say by normalizing Z_τ to have norm $\sqrt{d}$ (note any bounded random vector is also subGaussian), such that $\left\|(Z_\tau Z_\tau^\top - I)g\right\|^2 \leq d^2 \|g\|^2$ for any vector $g \in \mathbb{R}^d$, a careful examination of the argument in Proposition 6 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 $g \in \mathbb{R}^d$, $\left|Z_\tau^\top g\right|$ is bounded within some log factor of $\|g\|$. On the event $\{\left|Z_\tau^\top g\right| = \tilde{O}(\|g\|)\}$, 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 $(Z_\tau Z_\tau^\top - I)(\nabla f(x_\tau) - \nabla f(x'_\tau)) \mid \mathcal{F}_{\tau-1}$ is subGaussian on the event that $\left|Z_\tau^\top \nabla f(x_\tau)\right|$ 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 to prove a concentration inequality of the form Eq. (4). 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

$$\mathrm{lr}(x) := \log\left(x \log(x)\right). \tag{5}$$

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

Proposition 2. *Let $\mathcal{F}_t,\ t \geq -1$ be a filtration. Let $(Z_t)_{t\geq 0}$ be a sequence of random vectors following the distribution $N(0, I)$ such that $Z_t \in \mathcal{F}_t$ and is independent of $\mathcal{F}_{t-1}$, and let $(v_t)_{t\geq 0}$ be a sequence of random vectors such that $v_t \in \mathcal{F}_{t-1}$. For each $\tau \geq 0$, let*

$$W_\tau = \sum_{t=0}^{\tau-1} M_t(Z_t Z_t^\top - I)v_t,$$

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

1. *For any $\theta > 0$, with probability at least $1 - \delta$, we have*

$$\|W_\tau\| \leq c'\theta \sum_{t=0}^{\tau-1} \|M_t\|_2^2 d(\mathrm{lr}(C\tau/\delta))^2 \|v_t\|^2 + \frac{1}{\theta}\log(Cd\tau/\delta).$$

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

$$\textit{either} \quad \sum_{t=0}^{\tau-1} \|M_t\|_2^2 d(\mathrm{lr}(C\tau/\delta))^2 \|v_t\|^2 \geq B$$

$$\textit{or} \quad \|W_\tau\| \leq c'\sqrt{\max\left\{\sum_{t=0}^{\tau-1} \|M_t\|_2^2 d(\mathrm{lr}(C\tau/\delta))^2 \|v_t\|^2, b\right\}(\log(C\tau d/\delta) + \log(\log(B/b) + 1))}$$

Moreover, as is clear from the bounds above, we may pick $C \geq 1$ such that $\log\left(\frac{C}{\delta}\right) \geq 1, \forall \delta \in (0, \frac{1}{e}]$.

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 $\Omega(1)$ probability near an ϵ-saddle point.

Proposition 3 (Saddle point improvement, informal version of Proposition 6)**.** *Suppose that x_{t_0} is an ϵ-approximate saddle point. Let $\delta \in (0, 1/e]$ be arbitrary. Consider any $F > 0$, such that $\phi_{T_s}(\delta, F) = \tilde{O}\left(\frac{\sqrt{\rho\epsilon}}{\rho}\right)^2$. Suppose we choose u, r and η such that*

$$u = \min\left\{\tilde{O}\left(\frac{\sqrt{\epsilon}}{\sqrt{\rho}d}\right), \tilde{O}\left(\sqrt{\frac{r\sqrt{\eta\rho\epsilon}}{d^{2.5}\rho}}\right)\right\},\ r = O(\epsilon),\ \eta = \min\left\{\tilde{O}\left(\frac{m\sqrt{\rho\epsilon}}{d\max\{L, L^2\}}\right), \tilde{O}\left(\frac{1}{\sqrt{\rho\epsilon}}\right)\right\}.$$

Suppose we pick $T_s = \tilde{O}\left(\frac{\iota}{\eta\sqrt{\rho\epsilon}}\right)$, where $\iota = \max\left\{\log\left(2\sqrt{\phi_{T_s}(\delta, F)}\frac{20\sqrt{d}\sqrt{\eta^2\gamma^2 + 2\eta\gamma}}{\eta r}\right), 1\right\}$, Then with probability at least $\frac{1}{3} - O\left(\frac{T_s\delta}{T}\right)$, $f(x_{t_0+T_s}) - f(x_{t_0}) \leq -F$.

Armed with the above result as well as Proposition 1, the main result in Theorem 1 then follows. The complete detailed analysis can be found in Appendix E.3

4 Conclusion

In this paper, we proved that it is possible to use fewer than $\Omega(d)$ function evaluations per iteration to escape saddle points and reach approximate second order stationary points efficiently in zeroth-order optimization; in fact, using just two function evaluations per iteration suffices. Along the way, we 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 of independent interest to researchers working in zeroth-order optimization as well as general stochastic optimization. There are a few limits of the current results which lead to several interesting future directions, such as extension to noisy function evaluations, as well as studying if some zeroth-order estimators such as asymmetric two-point estimators (Nesterov and Spokoiny, 2017) or single-point estimators (Flaxman et al., 2005) could actually escape saddle points without additional perturbation noise.

REFERENCES

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

Agarwal, A., Foster, D. P., Hsu, D., Kakade, S. M., and Rakhlin, A. (2013). Stochastic convex optimization with bandit feedback. *SIAM Journal on Optimization*, 23(1):213–240.

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

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

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

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

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

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

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

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

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

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

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

Flokas, L., Vlatakis-Gkaragkounis, E.-V., and Piliouras, G. (2019). Efficiently avoiding saddle points with zero order methods: No gradients required.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Torczon, V. (1997). On the convergence of pattern search algorithms. *SIAM Journal on Optimization*, 7(1):1–25.

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

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

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

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

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

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

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

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

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

A Related Work

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., 2018; Sun et al., 2019a; Staib et al., 2019), or gradient-based robust Hessian power/curvature exploitation methods (Zhang and 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 and Boumal, 2019; Han and 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 and 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 $\tilde{O}(1/\epsilon^{3.5})$ 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 and Yaroslavtsev, 2021), distributed optimization (Vlaski and 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; Flokas et al., 2019; Balasubramanian and 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 and 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 $\tilde{O}(d/\epsilon^2)$ function evaluations (Nesterov and 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 which we will frequently use in our proofs. We then describe in detail (and prove) the sequence of results that lead up to Proposition 5, showing that there cannot be too many iterations with large gradients. Next, we describe the saddle

point argument in detail, and prove Proposition 6. Finally, we combine these results and prove Theorem 2, our main result.

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 $d \geq 2$ and $m \leq d$.

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

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

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

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

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

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

Definition 5 (norm-subGaussian random vectors)**.** A random vector $\boldsymbol{x} \in \mathbb{R}^d$ is σ-norm-subGaussian (nSG(σ)), there exists $\sigma > 0$ such that

$$\mathbb{P}(\|\boldsymbol{x} - \mathbb{E}\boldsymbol{x}\| \geq s) \leq 2e^{-\frac{s^2}{2\sigma^2}} \quad \forall s \geq 0.$$

We recall the following result which provides several examples of nSG random vectors. In particular, it tells us a random vector $\boldsymbol{x} \in \mathbb{R}^d$ that is $(\sigma/\sqrt{d})-$subGaussian is also σ-subGaussian.

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

1. *A bounded random vector $\boldsymbol{x} \in \mathbb{R}^d$ so that $\|\boldsymbol{x}\| \leq \sigma$.*
2. *A random vector $\boldsymbol{x} \in \mathbb{R}^d$, where $\boldsymbol{x} = \xi \boldsymbol{e}_1$ and the random variable $\xi \in \mathbb{R}$ is σ-subGaussian.*
3. *A random vector $\boldsymbol{x} \in \mathbb{R}^d$ that is $(\sigma/\sqrt{d})-$subGaussian*

In addition, if $\boldsymbol{x} \in \mathbb{R}^d$ is zero-mean nSG(σ), its component along a single direction is also subGaussian.

Lemma 5. *Suppose $\boldsymbol{x} \in \mathbb{R}^d$ is zero-mean nSG(σ). Then, for any fixed vector $\boldsymbol{v} \in \mathbb{R}^d$, $\langle \boldsymbol{v}, \boldsymbol{x}\rangle$ is zero-mean $\|\boldsymbol{v}\|\sigma$-subGaussian.*

Proof. Without loss of generality, we assume that $\boldsymbol{v} \in \mathbb{S}^{d-1}$ is a unit vector. That $\langle \boldsymbol{v}, \boldsymbol{x}\rangle$ is zero-mean follows directly from $\boldsymbol{x}$ being zero-mean and $\boldsymbol{v}$ being fixed. Meanwhile, since $|\langle \boldsymbol{v}, \boldsymbol{x}\rangle| \leq \|\boldsymbol{v}\|\|\boldsymbol{x}\| = \|\boldsymbol{x}\|$, for any $s \geq 0$, it follows that

$$\mathbb{P}(|\langle \boldsymbol{v}, \boldsymbol{x}\rangle| \geq s) \leq \mathbb{P}(\|\boldsymbol{x}\| \geq s) \leq 2e^{-\frac{s^2}{2\sigma^2}},$$

where the last inequality follows from the fact that $\boldsymbol{x}$ is zero-mean and also nSG(σ). Hence $\langle \boldsymbol{v}, \boldsymbol{x}\rangle$ is zero-mean SG(σ), 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 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, and corresponding filtrations $\mathcal{F}_i$ generated by $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i)$. We assume $\boldsymbol{x}_i \mid \mathcal{F}_{i-1}$ is zero-mean, nSG(σ_i), with $\sigma_i \in \mathcal{F}_{i-1}$, i.e,*

$$\mathbb{E}\left[\boldsymbol{x}_i \mid \mathcal{F}_{i-1}\right] = 0,$$

and

$$\mathbb{P}\left(\|\boldsymbol{x}_i\| \geq s \mid \mathcal{F}_{i-1}\right) \leq 2e^{-\frac{s^2}{2\sigma_i^2}} \quad \forall s \geq 0,$$

where σ_i is a measurable function of $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1})$ for each i.

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

Lemma 6. *Suppose $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in \mathbb{R}^d$ satisfy Condition 1, i.e. each $\boldsymbol{x}_i \mid \mathcal{F}_{i-1}$ is mean-zero, nSG(σ_i) with $\sigma_i \in \mathcal{F}_{i-1}$. Let $\{\boldsymbol{u}_i\}$ denote a sequence of random vectors such that $\boldsymbol{u}_i \in \mathcal{F}_{i-1}$ for every $i \in [n]$. Then, there exists an absolute constant c, such that for any $\delta \in (0,1)$ and $\lambda > 0$, with probability at least $1-\delta$,*

$$\sum_{i=1}^n \langle \boldsymbol{u}_i, \boldsymbol{x}_i \rangle \leq c\lambda \sum_{i=1}^n \|\boldsymbol{u}_i\|^2 \sigma_i^2 + \frac{1}{\lambda}\log(1/\delta).$$

Proof. We note that if $\boldsymbol{x}_i$ is mean-zero and nSG(σ_i), then by Lemma 5, $\langle \boldsymbol{u}_i, \boldsymbol{x}_i \rangle \mid \mathcal{F}_{i-1}$ is zero-mean and $\|\boldsymbol{u}_i\|\sigma_i$-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 $\langle \boldsymbol{u}_i, \boldsymbol{x}_i \rangle$ is $\|\boldsymbol{u}_i\|\sigma_i$-subGaussian, for any $\lambda > 0$, we have that

$$\mathbb{E}\left[\exp(\lambda\langle \boldsymbol{u}_i, \boldsymbol{x}_i \rangle) \mid \mathcal{F}_{i-1}\right] \leq \exp(\lambda^2 \|\boldsymbol{u}_i\|^2 \sigma_i^2/2)$$

For any $\lambda > 0$ and $s \geq 0$, observe that

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^n \lambda\langle \boldsymbol{u}_i, \boldsymbol{x}_i \rangle - \lambda^2\|\boldsymbol{u}_i\|^2\sigma_i^2/2 \geq s\right) \\
&= \mathbb{P}\left(\exp\left(\lambda\sum_{i=1}^n \langle \boldsymbol{u}_i, \boldsymbol{x}_i \rangle - \lambda^2\|\boldsymbol{u}_i\|^2\sigma_i^2/2\right) \geq \exp(\lambda s)\right) \\
&\leq \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^n \langle \boldsymbol{u}_i, \boldsymbol{x}_i \rangle - \lambda^2\|\boldsymbol{u}_i\|^2\sigma_i^2/2\right)\right]\exp(-\lambda s) \\
&= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^n \langle \boldsymbol{u}_i, \boldsymbol{x}_i \rangle - \lambda^2\|\boldsymbol{u}_i\|^2\sigma_i^2/2\right)\Big|\mathcal{F}_{n-1}\right]\right]\exp(-\lambda s) \\
&= \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n-1} \langle \boldsymbol{u}_i, \boldsymbol{x}_i \rangle - \lambda^2\|\boldsymbol{u}_i\|^2\sigma_i^2/2\right)\mathbb{E}\left[\exp\left(\lambda\langle \boldsymbol{u}_n, \boldsymbol{x}_n \rangle - \lambda^2\|\boldsymbol{u}_n\|^2\sigma_n^2/2\right)\Big|\mathcal{F}_{n-1}\right]\right]\exp(-\lambda s) \\
&\overset{\text{(i)}}{\leq} \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n-1} \langle \boldsymbol{u}_i, \boldsymbol{x}_i \rangle - \lambda^2\|\boldsymbol{u}_i\|^2\sigma_i^2/2\right)\right]\exp(-\lambda s) \leq \cdots \leq \exp(-\lambda s)
\end{aligned}$$

Above, (i) follows from the fact that $\langle \boldsymbol{u}_i, \boldsymbol{x}_i \rangle \mid \mathcal{F}_{i-1}$ is zero-mean and $\|\boldsymbol{u}_i\|\sigma_i$-subGaussian for each $i \in [n]$. The final result then follows by picking $c = \frac{1}{2}$ and $s = \log(1/\delta)$. □

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 $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in \mathbb{R}^d$ satisfy Condition 1. Then, there exists an absolute constant c such that for any fixed $\delta \in (0,1)$, $\theta > 0$, with probability at least $1-\delta$,*

$$\left\| \sum_{i=1}^n \boldsymbol{x}_i \right\| \leq c\theta \sum_{i=1}^n \sigma_i^2 + \frac{1}{\theta} \log(2d/\delta).$$

Moreover, there are two corollaries.

1. *((Jin et al., 2019b, Corollary 7)) When $\{\sigma_i\}$ is deterministic, there exists an absolute constant c such that for any fixed $\delta \in (0,1)$, with probability at least $1-\delta$.*

$$\left\| \sum_{i=1}^n \boldsymbol{x}_i \right\| \leq c\sqrt{\log(2d/\delta) \sum_{i=1}^n \sigma_i^2}$$

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

$$\textit{either } \sum_{i=1}^n \sigma_i^2 \geq B \quad \textit{or } \left\| \sum_{i=1}^n \boldsymbol{x}_i \right\| \leq c\sqrt{\max\left\{\sum_{i=1}^n \sigma_i^2, b\right\} \cdot (\log(2d/\delta) + \log(\log(B/b)))}$$

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 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}$, with mean $\mathbb{E}[\boldsymbol{x}_i] \leq c'\sigma$ 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 $1-\delta$,*

$$\sum_{i=1}^n \boldsymbol{x}_i \leq C\sigma(n + \log(1/\delta)). \tag{6}$$

Proof. The result of Eq. (6) follows by applying Bernstein's inequality to $\sum_{i=1}^n \boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i]$ (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 $s \geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^n (\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i]) \geq s\right) \leq \exp\left(-c \min\left\{\frac{s^2}{n\sigma^2}, \frac{s}{\sigma}\right\}\right).$$

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

$$\mathbb{P}\left(\sum_{i=1}^n (\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i]) \geq s\right) \leq \exp\left(-c \min\left\{\frac{s^2}{n\sigma^2}, \frac{s}{\sigma}\right\}\right) \leq \exp\left(-c\left(n + \frac{\log(1/\delta)}{c}\right)\right) \leq \delta.$$

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

$$\sum_{i=1}^n (\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i]) \leq \sigma\left(n + \frac{\log(1/\delta)}{c}\right) \implies \sum_{i=1}^n \boldsymbol{x}_i \leq \sigma\left(n + \frac{\log(1/\delta)}{c}\right) + nc'\sigma,$$

where implication holds since by assumption, $\mathbb{E}[\boldsymbol{x}_i] \leq c'\sigma$ 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

$$W_\tau = \sum_{t=0}^{\tau-1} M_t(Z_t Z_t^\top - I)v_t, \tag{7}$$

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 $Z_t \sim N(0, I)$.

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

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

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. (7), we see that for the equation above, we can define $M_t = \eta(I - \eta H)^{\tau-1-t}$ (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. (7).

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 $\mathbb{R}$. When bounding a sum in the form of $\sum_{t=0}^{\tau-1}(Z_t Z_t^\top - I)v_t$, 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

$$\mathrm{lr}(x) \coloneqq \log\left(x \log(x)\right). \tag{8}$$

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

Proposition 2. *Let $\mathcal{F}_t$, $t \geq -1$ be a filtration. Let $(Z_t)_{t\geq 0}$ be a sequence of random vectors following the distribution $N(0, I)$ such that $Z_t \in \mathcal{F}_t$ and is independent of $\mathcal{F}_{t-1}$, and let $(v_t)_{t\geq 0}$ be a sequence of random vectors such that $v_t \in \mathcal{F}_{t-1}$. For each $\tau \geq 0$, let*

$$W_\tau = \sum_{t=0}^{\tau-1} M_t(Z_t Z_t^\top - I)v_t,$$

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

1. *For any $\theta > 0$, with probability at least $1 - \delta$, we have*

$$\|W_\tau\| \leq c'\theta \sum_{t=0}^{\tau-1} \|M_t\|_2^2 d(\mathrm{lr}(C\tau/\delta))^2 \|v_t\|^2 + \frac{1}{\theta}\log(Cd\tau/\delta).$$

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

$$\textit{either} \quad \sum_{t=0}^{\tau-1} \|M_t\|_2^2 d(\mathrm{lr}(C\tau/\delta))^2 \|v_t\|^2 \geq B$$

$$\textit{or} \quad \|W_\tau\| \leq c'\sqrt{\max\left\{\sum_{t=0}^{\tau-1} \|M_t\|_2^2 d(\mathrm{lr}(C\tau/\delta))^2 \|v_t\|^2, b\right\}\left(\log(C\tau d/\delta) + \log(\log(B/b) + 1)\right)}$$

Moreover, as is clear from the bounds above, we may pick $C \geq 1$ such that $\log\left(\frac{C}{\delta}\right) \geq 1, \forall \delta \in (0, \frac{1}{e}]$.

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 (Jin et al., 2019b, Corollary 8). For simplicity, we shall assume $v_t \neq 0$ in the intermediate steps; extension to the general case is straightforward.

First of all, for $0 \leq \alpha < 1$, let

$$g(\alpha;\delta) = \sqrt{\frac{2}{\pi}} \int_{\alpha}^{\sqrt{2\,\mathrm{lr}(1/\delta)}} (x^2-1)e^{-x^2/2}\,dx = \sqrt{\frac{2}{\pi}}\left(\alpha e^{-\alpha^2/2} - \frac{\delta\sqrt{2\,\mathrm{lr}(1/\delta)}}{\log(1/\delta)}\right).$$

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

$$\frac{\mathrm{lr}(1/\delta)}{(\log(1/\delta))^2} = \frac{\log\log(1/\delta) + \log(1/\delta)}{(\log(1/\delta))^2} = \frac{1}{\log(1/\delta)}\left(\frac{\log\log(1/\delta)}{\log(1/\delta)} + 1\right) \leq 1,$$

which leads to

$$g(2\delta;\delta) = \sqrt{\frac{2}{\pi}}\left(2\delta e^{-2\delta^2} - \frac{\delta\sqrt{2\,\mathrm{lr}(1/\delta)}}{\log(1/\delta)}\right) \geq \sqrt{\frac{2}{\pi}}\left(2e^{-2/e^2}\delta - \sqrt{2}\delta\right) > 0$$

for $\delta \in (0, 1/e]$. Furthermore, we obviously have $g(0;\delta) < 0$. Therefore $g(\alpha;\delta) = 0$ has a unique solution in $(0, 2\delta)$, which we denote by $\alpha(\delta)$.[3] These results imply that, for a random variable Z following the standard normal distribution, we have

$$\mathbb{E}\Big[(Z^2-1)\mathbb{1}_{\alpha(\delta)\leq|Z|\leq\sqrt{2\,\mathrm{lr}(1/\delta)}}\Big] = \sqrt{\frac{2}{\pi}}\int_{\alpha(\delta)}^{\sqrt{2\,\mathrm{lr}(1/\delta)}} (x^2-1)e^{-x^2/2}\,dx = g(h(\delta);\delta) = 0$$

and

$$\begin{aligned}
\mathbb{P}(\alpha(\delta) \leq |Z| \leq \sqrt{2\,\mathrm{lr}(1/\delta)}) &\geq 1 - 2\left(\frac{1}{\sqrt{2\pi}}\int_{\sqrt{2\,\mathrm{lr}(1/\delta)}}^{\infty} e^{-x^2/2}\,dx + \frac{1}{\sqrt{2\pi}}\int_0^{\alpha(\delta)} e^{-x^2/2}\,dx\right) \\
&\geq 1 - 2\left(\frac{1}{2}\exp\left(-\frac{2\,\mathrm{lr}(1/\delta)}{2}\right) + \frac{\alpha(\delta)}{\sqrt{2\pi}}\right) = 1 - 2\left(\frac{\delta}{2\log(1/\delta)} + \frac{\alpha(\delta)}{\sqrt{2\pi}}\right) \\
&\geq 1 - 2\left(\frac{\delta}{2} + \frac{2}{\sqrt{2\pi}}\delta\right) \geq 1 - C\delta
\end{aligned}$$

for any $\delta \in (0, 1/e]$, where we define the absolute constant $C \coloneqq 2(1/2 + 2/\sqrt{2\pi})$.

Now we let A_t denote the event

$$A_t = \left\{\alpha(\delta) \leq \frac{|Z_t^\top v_t|}{\|v_t\|} \leq \sqrt{2\,\mathrm{lr}(1/\delta)}\right\}.$$

Since $Z_t^\top v_t/\|v_t\|$ conditioned on $\mathcal{F}_{t-1}$ follows the standard normal distribution, we have

$$\mathbb{P}_{\mathcal{F}_{t-1}}(A_t) \geq 1 - C\delta, \tag{9}$$

and

$$\mathbb{E}_{\mathcal{F}_{t-1}}\left[v_t^\top\left(Z_tZ_t^\top - I\right)v_t\mathbb{1}_{A_t}\right] = 0.$$

Moreover, for any random vector $u \in \mathcal{F}_{t-1}$ that is orthogonal to v_t, we have

$$\mathbb{E}_{\mathcal{F}_{t-1}}\left[u^\top\left(Z_tZ_t^\top - I\right)v_t\mathbb{1}_{A_t}\right] = \mathbb{E}_{\mathcal{F}_{t-1}}\left[u^\top Z_t\right]\cdot\mathbb{E}_{\mathcal{F}_{t-1}}\left[Z_t^\top v_t\mathbb{1}_{A_t}\right] = 0,$$

where we used the fact that $Z_t^\top u$ is independent of $Z_t^\top v_t$ conditioned on $\mathcal{F}_{t-1}$. Therefore

$$\mathbb{E}_{\mathcal{F}_{t-1}}\left[(Z_tZ_t^\top - I)v_t\mathbb{1}_{A_t}\right] = 0.$$

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

$$\alpha(\delta) = \sqrt{-W_0\left(-\frac{2\delta^2\,\mathrm{lr}(1/\delta)}{(\log(1/\delta))^2}\right)}.$$

Consider defining then the random variable Q_t by

$$Q_t \coloneqq (Z_t Z_t^\top - I)v_t \cdot \mathbb{1}_{A_t}.$$

We now show that $Q_t \mid \mathcal{F}_{t-1}$ is norm-subGaussian. Let $u \in \mathbb{R}^d$ with $\|u\| = 1$ be arbitrary. We have

$$\begin{aligned}
u^\top Q_t &= u^\top (Z_t Z_t^\top - I)v_t \cdot \mathbb{1}_{A_t} \\
&= u^\top \left(\frac{v_t v_t^\top}{\|v_t\|^2} + I - \frac{v_t v_t^\top}{\|v_t\|^2} \right) (Z_t Z_t^\top - I)v_t \cdot \mathbb{1}_{A_t} \\
&= u^\top v_t \left(\frac{|Z_t^\top v_t|^2}{\|v_t\|^2} - 1 \right) \cdot \mathbb{1}_{A_t} + u^\top \left(I - \frac{v_t v_t^\top}{\|v_t\|^2} \right) (Z_t Z_t^\top - I)v_t \cdot \mathbb{1}_{A_t} \\
&= u^\top v_t \left(\frac{|Z_t^\top v_t|^2}{\|v_t\|^2} - 1 \right) \cdot \mathbb{1}_{A_t} + u_\perp^\top Z_t Z_t^\top v_t \cdot \mathbb{1}_{A_t},
\end{aligned}$$

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

$$\left| u^\top v_t \left(\frac{|Z_t^\top v_t|^2}{\|v_t\|^2} - 1 \right) \cdot \mathbb{1}_{A_t} \right| \leq |u^\top v_t| (2 \operatorname{lr}(1/\delta) - 1),$$

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

$$\mathbb{P}_{\mathcal{F}_{t-1}} \left(|u_\perp^\top Z_t Z_t^\top v_t \cdot \mathbb{1}_{A_t}| \geq s \right) \leq \mathbb{P}_{\mathcal{F}_{t-1}} \left(|Z_t^\top u_\perp| \sqrt{2 \operatorname{lr}(1/\delta)} \|v_t\| \geq s \right),$$

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

$$\begin{aligned}
&(2 \operatorname{lr}(1/\delta) - 1)|u^\top v_t| + \sqrt{2 \operatorname{lr}(1/\delta)} \|u_\perp\| \|v_t\| \\
&\leq 2 \operatorname{lr}(1/\delta)(|u^\top v_t| + \|u_\perp\| \|v_t\|) \leq 2\sqrt{2} \operatorname{lr}(1/\delta) \sqrt{|u^\top v_t|^2 + \|u_\perp\|^2 \|v_t\|^2} \\
&= 2\sqrt{2} \operatorname{lr}(1/\delta) \|v_t\|,
\end{aligned}$$

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

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

$$\left\| \sum_{t=0}^{\tau-1} M_t Q_t \right\| \leq c\theta \sum_{t=0}^{\tau-1} d(\operatorname{lr}(1/\delta))^2 \|M_t\|_2^2 \|v_t\|^2 + \frac{1}{\theta} \log(2d/\delta).$$

Now, consider denoting the event

$$A \coloneqq \bigcup_{t=0}^{\tau-1} A_t = \left\{ \left| Z_t^\top v_t \right| \in \left(\alpha(\delta) \|v_t\|, \sqrt{2 \operatorname{lr}(1/\delta)} \|v_t\| \right), \ \forall t \in \{0, \ldots, \tau - 1\} \right\}$$

By the union bound and Eq. (9), we note that

$$\mathbb{P}(A) \geq 1 - \tau C \delta.$$

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

$$\mathbb{P} \left(\left\| \sum_{t=0}^{\tau-1} M_t (Z_t Z_t^\top - I) v_t \right\| \leq c\theta \sum_{t=0}^{\tau-1} d(\operatorname{lr}(1/\delta))^2 \|M_t\|_2^2 \|v_t\|^2 + \frac{1}{\theta} \log(2d/\delta) \right)$$

$$\geq \mathbb{P}\left(\left\|\sum_{t=0}^{\tau-1} M_t Y_t\right\| \leq c\theta \sum_{t=0}^{\tau-1} d(\mathrm{lr}(1/\delta))^2 \|M_t\|_2^2 \|v_t\|^2 + \frac{1}{\theta}\log(2d/\delta), \text{ and } A \text{ happens}\right)$$
$$\geq 1 - \left(\mathbb{P}\left(\left\|\sum_{t=0}^{\tau-1} M_t Y_t\right\| \geq c\theta \sum_{t=0}^{\tau-1} d(\mathrm{lr}(1/\delta))^2 \|M_t\|_2^2 \|v_t\|^2 + \frac{1}{\theta}\log(2d/\delta)\right) + \mathbb{P}(A^c)\right)$$
$$\geq 1 - (\delta + \tau C\delta).$$

Now, by rescaling δ to $\delta/(C\tau+1)$, 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 6. We say that a random variable $X \in \mathbb{R}$ is *sub-Weibull(K, α)* for some $K, \alpha > 0$,

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

For instance, the standard normal distribution is sub-Weibull($1, \frac{1}{2}$). From the way we define the tail parameter α, the larger the α, 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, α) and sub-Weibull(K_Y, α) respectively. Then, XY is sub-Weibull($C(K_X \cdot K_Y), 2\alpha$) and $X+Y$ is sub-Weibull($C(K_X + K_Y), \alpha$) 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', α) sub-Weibull random variables. Then, for some absolute constant $c > 0$, for all $s \geq ncK'$, we have*

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

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 $1-\delta$,*

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

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

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

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 6 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 $s \geq nCc^k d^k$,

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

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 $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_{n-1} \subset \mathcal{F}_n$ and a sequence of random variables $X_1, \ldots, X_n$ such that $X_i \in \mathcal{F}_i$. Suppose that*

$$\mathbb{P}_{\mathcal{F}_{i-1}}(X_i \leq a) = 1 \qquad \text{and} \qquad \mathbb{P}_{\mathcal{F}_{i-1}}(X_i \leq -b) \geq p \tag{10}$$

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

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

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

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n}(X_i+\mu)\right)\right] &= \mathbb{E}\left[\mathbb{E}_{\mathcal{F}_{n-1}}\left[\exp\left(\lambda \sum_{i=1}^{n}(X_i+\mu)\right)\right]\right] \\ &= \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1}(X_i+\mu)\right)\mathbb{E}_{\mathcal{F}_{n-1}}[\exp(\lambda(X_n+\mu))]\right]. \end{aligned} \tag{11}$$

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,

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

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. (11), we have that

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

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

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

$$
\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq -n\mu + s\right) &\leq \frac{\mathbb{E}\left[\exp(\lambda(\sum_{i=1}^{n}(X_i+\mu)))\right]}{\exp(\lambda s)} \\
&\leq \exp(n\lambda^2(b-\mu)^2 - \lambda s)
\end{aligned}
$$

By finding the minimizing λ, we find that

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

which completes the proof. □

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

Proposition 4 (Weakened supermartingale concentration inequality)**.** *Consider a filtration of sigma-algebras $\mathcal{F}_0 \subset \mathcal{F}_1 \cdots \subset \mathcal{F}_n$ and a sequence of random variables $X_1, \dots, X_n$ such that $X_i \in \mathcal{F}_i$. Consider for each $i \in \{1, \dots, n\}$ a bad set B_i where $\mathbb{1}_{B_i} \in \mathcal{F}_{i-1}$, and suppose*

$$
\mathbb{P}_{\mathcal{F}_{i-1}}(X_i \mathbb{1}_{B_i^c} \leq a) = 1 \qquad \textit{and} \qquad \mathbb{P}_{\mathcal{F}_{i-1}}(X_i \mathbb{1}_{B_i^c} \leq -b) \geq p
$$

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

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

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

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

Since $\mathbb{P}\left(X_i \neq Q_i \text{ for some } i \in [n]\right) \leq \sum_i \mathbb{P}(X_i \in B_i)$, it follows that

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

which completes the proof. □

D FUNCTION DECREASE

To show that the function can make progress in the large gradient iterations, we recall the following function decrease lemma first introduced in the main text. Note that at each iteration, we access the function oracle $2m$ times, where m is a batch-size parameter we can control.

Lemma 1 (Function decrease for batch zero-order optimization)**.** *Suppose at each time t, the algorithm performs the update step*

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

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 $c_1 > 0, C_1 \geq 1$ such that, for any $T \in \mathbb{Z}^+$ and $T \geq \tau' \geq \tau > 0$, $\alpha > 0$ and $\delta \in (0, 1/e]$, upon defining $\mathcal{H}_{0,\tau}(\delta)$ to be the event on which the inequality

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

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

$$\mathbb{P}(\mathcal{H}_{0,\tau}(\delta)) \geq 1 - \frac{(\tau+4)\delta}{T}, \qquad \mathbb{P}(\cap_{\tau=1}^{\tau'} \mathcal{H}_{0,\tau}(\delta)) \geq 1 - \frac{5\tau'\delta}{T}$$

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

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

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 $\alpha \in [0, 1]$ such that $f(x + y) = f(x) + \langle \nabla f(x), y \rangle + \frac{1}{2} y^\top \nabla^2 f(x + \alpha y)\, y$. Therefore

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

with

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

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

$$x_{t+1} = x_t - \eta \left(\frac{1}{m} \sum_{i=1}^{m} \left(Z_{t,i} Z_{t,i}^\top \nabla f(x_t) + \frac{u}{2} Z_{t,i} Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i} \right) + Y_t \right) \tag{12}$$

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

Observe that

$$\begin{aligned} f(x_{t+1}) & \overset{\text{(i)}}{\leq} f(x_t) + \langle x_{t+1} - x_t, \nabla f(x_t) \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\ & \overset{\text{(ii)}}{=} f(x_t) - \eta \frac{1}{m} \sum_{i=1}^{m} \left| Z_{t,i}^\top \nabla f(x_t) \right|^2 - \eta \frac{1}{m} \sum_{i=1}^{m} \frac{u}{2} Z_{t,i}^\top \nabla f(x_t) \cdot Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i} - \eta \langle \nabla f(x_t), Y_t \rangle \\ & \quad + \frac{L\eta^2}{2} \left\| \frac{1}{m} \sum_{i=1}^{m} \left(Z_{t,i} Z_{t,i}^\top \nabla f(x_t) + \frac{u}{2} Z_{t,i} Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i} \right) + Y_t \right\|^2 \\ & \overset{\text{(iii)}}{\leq} f(x_t) - \frac{\eta}{m} \sum_{i=1}^{m} \left| Z_{t,i}^\top \nabla f(x_t) \right|^2 + \frac{\eta}{m} \sum_{i=1}^{m} \left(\frac{\left| Z_{t,i}^\top \nabla f(x_t) \right|^2}{4} + \frac{u^2 \left| Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i} \right|^2}{4} \right) - \eta \langle \nabla f(x_t), Y_t \rangle \\ & \quad + \frac{L\eta^2}{2} \left(2 \left\| \frac{1}{m} \sum_{i=1}^{m} Z_{t,i} Z_{t,i}^\top \nabla f(x_t) \right\|^2 + u^2 \left\| \frac{1}{m} \sum_{i=1}^{m} Z_{t,i} Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i} \right\|^2 + 4 \|Y_t\|^2 \right) \\ & \overset{\text{(iv)}}{\leq} f(x_t) - \frac{3\eta}{4m} \sum_{i=1}^{m} \left| Z_{t,i}^\top \nabla f(x_t) \right|^2 + \frac{\eta u^2}{m} \sum_{i=1}^{m} \frac{u^2 \rho^2 \|Z_{t,i}\|^6}{4} - \eta \langle \nabla f(x_t), Y_t \rangle \end{aligned}$$

$$
\begin{aligned}
&+ \frac{L\eta^2}{2}\left(2\left\|\frac{1}{m}\sum_{i=1}^{m} Z_{t,i}Z_{t,i}^\top \nabla f(x_t)\right\|^2 + \frac{u^2}{m}\sum_{i=1}^{m} u^2\rho^2\|Z_{t,i}\|^8 + 4\|Y_t\|^2\right) \\
&\leq f(x_t) - \frac{3\eta}{4m}\sum_{i=1}^{m}\left|Z_{t,i}^\top \nabla f(x_t)\right|^2 + \frac{\eta u^4\rho^2}{4m}\sum_{i=1}^{m}\|Z_{t,i}\|^6 + \frac{L\eta^2 u^4\rho^2}{2m}\sum_{i=1}^{m}\|Z_{t,i}\|^8 - \eta\langle \nabla f(x_t), Y_t\rangle \\
&\quad + \frac{L\eta^2}{2}\left(2\left\|\frac{1}{m}\sum_{i=1}^{m} Z_{t,i}Z_{t,i}^\top \nabla f(x_t)\right\|^2 + 4\|Y_t\|^2\right) \qquad (13)
\end{aligned}
$$

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. (12). To derive (iii), we used the fact that $ab \leq (a^2 + b^2)/2$ for any $a, b \in \mathbb{R}_{\geq 0}$, as well as two applications of the fact that $\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$ for any two vectors $a, b \in \mathbb{R}^d$. To derive (iv), we used the fact that $\left\|\tilde{H}_{t,i}\right\| \leq \rho u \|Z_{t,i}\|$.

To continue from Eq. (13), 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

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

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 $\mathfrak{c}_1 > 0, C_1 \geq 1$ such that for any $\delta \in (0, 1/e]$ and $\theta > 0$, with probability at least $1 - \delta$ conditioned on $\mathcal{F}_{t-1}$,

$$
\begin{aligned}
\left\|\sum_{i=1}^{m}(Z_{t,i}Z_{t,i}^\top - I)\nabla f(x_t)\right\| &\leq \mathfrak{c}_1\theta\sum_{i=1}^{m} d(\mathrm{lr}(C_1 m/\delta))^2\|\nabla f(x_t)\|^2 + \frac{1}{\theta}\log(C_1 dm/\delta) \\
&= \mathfrak{c}_1\theta md(\mathrm{lr}(C_1 m/\delta))^2\|\nabla f(x_t)\|^2 + \frac{1}{\theta}\log(C_1 dm/\delta). \qquad (14)
\end{aligned}
$$

Moreover, since $C_1 \geq 1$, $\log(C_1 dm/\delta)$ and $\mathrm{lr}(C_1 m/\delta)$ 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

$$
\theta = \frac{1}{\sqrt{\mathfrak{c}_1 md\ \mathrm{lr}(C_1 dm/\delta)}\|\nabla f(x_t)\|}
$$

which is $\mathcal{F}_{t-1}$-measurable, and plug it into Eq. (14) to find that the probability conditioned on $\mathcal{F}_{t-1}$ of the following event

$$
\left\|\sum_{i=1}^{m}(Z_{t,i}Z_{t,i}^\top - I)\nabla f(x_t)\right\| \leq 2\sqrt{\mathfrak{c}_1}(\mathrm{lr}(C_1 dm/\delta))^{3/2}\sqrt{md}\|\nabla f(x_t)\| \qquad (15)
$$

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 $1 - \delta$,

$$
\begin{aligned}
\left\|\frac{1}{m}\sum_{i=1}^{m} Z_{t,i}Z_{t,i}^\top \nabla f(x_t)\right\|^2 &\leq 2\left\|\frac{1}{m}\sum_{i=1}^{m}(Z_{t,i}Z_{t,i}^\top - I)\nabla f(x_t)\right\|^2 + 2\|\nabla f(x_t)\|^2 \\
&\leq 4\mathfrak{c}_1(\mathrm{lr}(C_1 dm/\delta))^3\frac{d}{m}\|\nabla f(x_t)\|^2 + 2\|\nabla f(x_t)\|^2 \\
&\leq \mathfrak{c}_2(\mathrm{lr}(C_1 dm/\delta))^3\frac{d}{m}\|\nabla f(x_t)\|^2, \qquad (16)
\end{aligned}
$$

where the last inequality comes from the fact that $\mathrm{lr}(C_1 dm/\delta) \geq 1$, our assumption at the outset of the appendix that $d \geq m$, and denoting $\mathfrak{c}_2 := 4\mathfrak{c}_1 + 2$.

Denote the event $\tilde{H}_{0,\tau}(\delta)$ as the event that

$$\begin{aligned} f(x_\tau) - f(x_0) \leq & -\sum_{t=0}^{\tau-1} \frac{3\eta}{4m} \sum_{i=1}^{m} \left| Z_{t,i}^\top \nabla f(x_t) \right|^2 + L\eta^2 \frac{\mathfrak{c}_2 d(\mathrm{lr}(C_1 dm/\delta))^3}{m} \sum_{t=0}^{\tau-1} \|\nabla f(x_t)\|^2 \\ & + \frac{\eta u^4 \rho^2}{4m} \sum_{t=0}^{\tau-1} \sum_{i=1}^{m} \|Z_{t,i}\|^6 + \frac{L\eta^2 u^4 \rho^2}{2m} \sum_{t=0}^{\tau-1} \sum_{i=1}^{m} \|Z_{t,i}\|^8 \\ & - \eta \sum_{t=0}^{\tau-1} \langle \nabla f(x_t), Y_t \rangle + 2L\eta^2 \sum_{t=0}^{\tau-1} \|Y_t\|^2 \end{aligned} \tag{17}$$

holds.

Now, continuing from Eq. (13), and using the bound in Eq. (16), summing over the iterations from $t = 0$ to $\tau - 1$, we find using the union bound that $\mathbb{P}(\cap_{\tau=1}^{\tau'} \tilde{H}_{0,\tau}(\delta)) \geq 1 - \tau'\delta$, $\mathbb{P}(\tilde{H}_{0,\tau}(\delta)) \geq 1 - \tau\delta$.

Now, by Lemma 6, for any $\delta \in (0,1), \alpha > 0$, with probability at least $1 - \delta$, there exists an absolute constant $\mathfrak{c}_3 > 0$ such that

$$-\eta \sum_{t=0}^{\tau-1} \langle \nabla f(x_t), Y_t \rangle \leq \eta \left(\frac{1}{\alpha} \sum_{t=0}^{\tau-1} \|\nabla f(x_t)\|^2 + \mathfrak{c}_3 \alpha r^2 \log(1/\delta) \right). \tag{18}$$

Meanwhile, since $Y_t \sim N(0, (r^2/d)I)$, $\|Y_t\|^2$ 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 $\mathfrak{c}_4 > 0$ such that

$$\sum_{t=0}^{\tau-1} \|Y_t\|^2 \leq \mathfrak{c}_4 r^2 (\tau + \log(1/\delta)) \tag{19}$$

with probability at least $1 - \delta$.

To bound $\sum_{t=0}^{\tau-1} \frac{1}{m} \sum_{i=1}^{m} \|Z_{t,i}\|^6$ and $\sum_{t=0}^{\tau-1} \frac{1}{m} \sum_{i=1}^{m} \|Z_{t,i}\|^8$, both sums of heavy tailed Gaussian moments, we use Lemma 11, which states that for any $k \in \mathbb{Z}^+$ and $\delta \in (0,1)$, with probability at least $1 - \delta$,

$$\frac{1}{m} \sum_{t=0}^{\tau-1} \sum_{i=1}^{m} \|Z_{t,i}\|^{2k} \leq \mathfrak{c}_5 \tau (\mathfrak{c}_6)^k d^k (1 + (\log(1/\delta))^k) \tag{20}$$

for some absolute constants $\mathfrak{c}_5, \mathfrak{c}_6 > 0$. As in the statement of the proof, using $\chi := \mathrm{lr}(C_1 dm/\delta)$ to ease the notation, denote the event that

$$\begin{aligned} f(x_\tau) - f(x_0) \leq & -\frac{3\eta}{4} \sum_{t=0}^{\tau-1} \frac{1}{m} \sum_{i=1}^{m} \left| Z_{t,i}^\top \nabla f(x_t) \right|^2 + \left(\frac{\eta}{\alpha} + \frac{\mathfrak{c}_2 L\eta^2 \chi^3 d}{m} \right) \sum_{t=0}^{\tau-1} \|\nabla f(x_t)\|^2 \\ & + \frac{\tau \eta u^4 \rho^2}{2} \cdot \mathfrak{c}_5 \mathfrak{c}_6^3 d^3 \left(\log \frac{1}{\delta} \right)^3 + \tau L\eta^2 u^4 \rho^2 \cdot \mathfrak{c}_5 \mathfrak{c}_6^4 d^4 \left(\log \frac{1}{\delta} \right)^4 \\ & + \eta(\mathfrak{c}_3 \alpha r^2 + 2\mathfrak{c}_4 \eta L r^2) \log \frac{1}{\delta} + 2\mathfrak{c}_4 L\eta^2 \tau r^2 \end{aligned}$$

holds as $\mathcal{H}_{0,\tau}(\delta)$.

Plugging Eq. (18), Eq. (19), and Eq. (20) into Eq. (17), by union bound, we see that

$$\mathbb{P}(\cap_{\tau=1}^{\tau'} \mathcal{H}_{0,\tau}(\delta)) \geq 1 - (\tau' + 4\tau')\delta = 1 - 5\tau'\delta, \qquad \mathbb{P}(\mathcal{H}_{0,\tau}) \geq 1 - (\tau + 4)\delta.$$

The final result then follows by rescaling δ to $\frac{\delta}{T}$ and denoting $c_1 := \max\{\mathfrak{c}_2, \mathfrak{c}_3, 2\mathfrak{c}_4, \mathfrak{c}_5 \mathfrak{c}_6^3/2, \mathfrak{c}_5 \mathfrak{c}_6^4\}$. □

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 $\|\nabla f(x_t)\| \geq \epsilon$. Roughly speaking, as

Eq. (2) shows, we need to prove a lower bound of the form

$$\sum_{t=0}^{T-1} \frac{1}{m} \sum_{i=1}^{m} \left\| Z_{t,i}^{\top} \nabla f(x_t) \right\|^2 \geq \Omega \left(\frac{1}{\alpha} + \frac{c_1 L \eta \chi^3 d}{m} \right) \sum_{t=0}^{T-1} \| \nabla f(x_t) \|^2 \tag{21}$$

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, $\frac{1}{m} \sum_{i=1}^{m} \left\| Z_{t,i}^{\top} \nabla f(x_t) \right\|^2$ 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 careful analysis to prove a bound similar to Eq. (21). We do so using the following approach.

1. Intuitively, whilst for each individual iteration t, $\frac{1}{m} \sum_{i=1}^{m} \left\| Z_{t,i}^{\top} \nabla f(x_t) \right\|^2$ could be small even as $\| \nabla f(x_t) \|^2$ is large, in a small number of (consecutive) iterations $\{t_0, \ldots, t_0 + t_f\}$, with high probability, there will be at least one iteration t within $\{t_0, \ldots, t_0 + t_f - 1\}$, such that $\frac{1}{m} \sum_{i=1}^{m} \left\| Z_{t,i}^{\top} \nabla f(x_t) \right\|^2 = \Omega(\| \nabla f(x_t) \|^2)$. 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 $\{t_0, \ldots, t_0 + t_f - 1\}$. 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 η, we can show that
   $$\frac{1}{2} \| \nabla f(x_{t_0}) \| \leq \| \nabla f(x_t) \| \leq 2 \| \nabla f(x_{t_0}) \| \qquad \forall t \in \{t_0, \ldots, t_0 + t_f - 1\}.$$
   As a result, combined with point 1, we see that
   $$\sum_{t=t_0}^{t_0+t_f-1} \frac{1}{m} \sum_{i=1}^{m} \left\| Z_{t,i}^{\top} \nabla f(x_t) \right\|^2 \geq \Omega(\| \nabla f(x_{t_0}) \|^2).$$
   Thus, by choosing α and η judiciously, for such intervals, it is possible to show that
   $$\begin{aligned} \sum_{t=t_0}^{t_0+t_f-1} \frac{1}{m} \sum_{i=1}^{m} \left\| Z_{t,i}^{\top} \nabla f(x_t) \right\|^2 \geq \Omega(\| \nabla f(x_{t_0}) \|^2) \geq & \Omega \left(\frac{1}{\alpha} + \frac{c_1 L \eta \chi^3 d}{m} \right) \sum_{t=t_0}^{t_0+t_f-1} \| \nabla f(x_t) \|^2 \\ = & \Omega \left(\frac{1}{\alpha} + \frac{c_1 L \eta \chi^3 d}{m} \right) \Omega \left(t_f \| \nabla f(x_{t_0}) \|^2 \right) \end{aligned}$$
   Thus, in these intervals, it is possible to obtain function improvement on the order of $\eta \Omega(\| \nabla f(x_{t_0}) \|^2)$.
   (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, u and η 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 $\eta \Omega(\epsilon^2)$ for such intervals. By appropriately choosing η, 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 5.

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*

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

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

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

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

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

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

$$t_f(\delta) = \left\lceil \frac{c_2}{m} \log \frac{T}{\delta} \right\rceil, \qquad \delta > 0,$$

and defining the event

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

we have

$$\mathbb{P}\left(\mathcal{B}_{t_0}(\delta; k)\right) \ge 1 - \frac{\delta}{T}.$$

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

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 $\le c\|\nabla f(x_t)\|^2$ for some absolute constant $c > 0$. Hence

$$\begin{aligned}
\mathbb{P}_{\mathcal{F}_{t-1}}(E_t) &= \mathbb{P}_{\mathcal{F}_{t-1}}\left(\frac{1}{m}\sum_{i=1}^m |Z_{t,i}^\top \nabla f(x_t)|^2 < \frac{1}{2}\|\nabla f(x_t)\|^2 \right) \\
&= \mathbb{P}_{\mathcal{F}_{t-1}}\left(\sum_{i=1}^m \left(\|\nabla f(x_t)\|^2 - \left|Z_{t,i}^\top \nabla f(x_t)\right|^2 \right) > \frac{m}{2}\|\nabla f(x_t)\|^2 \right) \\
&\le \exp\left(-c' m\right),
\end{aligned}$$

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

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

$$= \mathbb{E}\left[\prod_{t=t_0}^{t_0+k-1} \mathbb{1}_{E_t}\right] = \mathbb{E}\left[\prod_{t=t_0}^{t_0+k-2} \mathbb{1}_{E_t} \cdot \mathbb{E}_{\mathcal{F}_{t_0+k-2}}\left[\mathbb{1}_{E_{t_0+k-1}}\right]\right]$$
$$\leq \exp(-c'm) \cdot \mathbb{E}\left[\prod_{t=t_0}^{t_0+k-2} \mathbb{1}_{E_t}\right] \leq \cdots \leq \exp(-c'mk).$$

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

$$k \geq t_f(\delta) = \left\lceil \frac{c_2}{m} \log \frac{T}{\delta} \right\rceil,$$

we get

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

which completes the proof. □

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

$$t_f(\delta) \coloneqq \left\lceil \frac{c_2}{m} \log \frac{T}{\delta} \right\rceil, \qquad \delta \in (0, 1/e], \tag{22}$$

where $c_2 \geq 1$ 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 $\left\|\frac{u}{2m}\sum_{i=1}^m Z_{t,i} Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i}\right\|$ as well as $\|Y_t\|$.

Lemma 15. *Define*

$$\mathcal{A}_t(\delta) \coloneqq \left\{\|\nabla f(x_{t+1}) - \nabla f(x_t)\| \leq \frac{\|\nabla f(x_t)\|}{8t_f(\delta)} + \eta L\left(\left\|\frac{u}{2m}\sum_{i=1}^m Z_{t,i} Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i}\right\| + \|Y_t\|\right)\right\} \tag{23}$$

where $t_f(\delta)$ is defined in Eq. (22)*, and let $C_1 \geq 1$ be the corresponding absolute constants defined in Lemma 1. Then there exists an absolute constant $c_4 > 0$ such that, whenever η satisfies*

$$\eta L \frac{c_4(\mathrm{lr}(C_1 dmT/\delta))^{3/2}\sqrt{d}}{\sqrt{m}} \leq \frac{1}{8t_f(\delta)}, \tag{24}$$

we have

$$\mathbb{P}(\mathcal{A}_t(\delta)) \geq 1 - \frac{\delta}{T}$$

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

$$\|\nabla f(x_{t+1}) - \nabla f(x_t)\| \leq L\|x_{t+1} - x_t\| \tag{25}$$
$$= \eta L\left\|\frac{1}{m}\sum_{i=1}^m Z_{t,i} Z_{t,i}^\top \nabla f(x_t) + \frac{u}{2m}\sum_{i=1}^m Z_{t,i} Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i} + Y_t\right\|. \tag{26}$$

Now, it follows from Eq. (16) (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

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

Hence, continuing from Eq. (26), it follows that with probability at least $1 - \delta/T$,

$$\|\nabla f(x_{t+1}) - \nabla f(x_t)\|$$

$$\leq \eta L \left(c_4 (\mathrm{lr}(C_1 dmT/\delta))^{3/2} \sqrt{\frac{d}{m}} \|\nabla f(x_t)\| + \left\| \frac{u}{2m} \sum_{i=1}^{m} Z_{t,i} Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i} \right\| + \|Y_t\| \right),$$

and by plugging in the condition Eq. (24), we see that the event

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

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 η 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

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

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 $t'_f \leq 2t_f(\delta)$, and any $t_0 \in \{0, \ldots, T - 1 - t'_f\}$. Suppose η satisfies the condition Eq.* (24)*. Then, on the event*

$$\mathcal{E}(t_0, t'_f, \delta) \cap \left(\bigcap_{t=t_0}^{t_0+t'_f-1} \mathcal{A}_t(\delta) \right),$$

we have

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

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

Proof. By plugging

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

into the definition of $\mathcal{A}_t(\delta)$, we see that, on the event $\mathcal{E}(t_0, t'_f, \delta) \cap \left(\bigcap_{t=t_0}^{t_0+t'_f-1} \mathcal{A}_t(\delta) \right)$, we have

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

and consequently,

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

which leads to

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

for all $t \in \{t_0, \ldots, t_0 + t'_f\}$. Then, since $(1 + 1/(4x))^{2x} \leq 2$ and $(1 - 1/(4x))^{2x} \geq 1/2$ for any $x \geq 1$, noting that $t'_f \leq 2t_f(\delta)$, 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 $\delta \in (0, 1/e]$ and $T \in \mathbb{Z}^+$ be such that $T > 2t_f(\delta) + 1$. Consider any positive integer $t'_f \leq 2t_f(\delta)$, and any $t_0 \in \{0, \ldots, T - 1 - t'_f\}$. Suppose η satisfies the condition Eq.* (24)*. Then, on the event*

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

we have

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

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

$$\|\nabla f(x_{t'})\| \leq 8t_f(\delta)\eta L \left(\frac{u}{2} \left\| \frac{1}{m} \sum_{i=1}^{m} Z_{t',i} Z_{t',i}^\top \tilde{H}_{t',i} Z_{t',i} \right\| + \|Y_{t'}\| \right). \tag{27}$$

Since we are working on an event which is a subset of $\mathcal{E}^c(t_0, t'_f, \delta)$, t' is well-defined. By $\|\tilde{H}_{t',i}\| \leq \rho u \|Z_{t',i}\|$, we see that

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

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

Recall that t' is the first time step such that Eq. (27) holds. By deriving similarly as in the proof of Lemma 16, we can show that for any $j \in \{t_0, t_0 + 1, \ldots, t' - 1\}$,

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

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

$$\begin{aligned}
\|\nabla f(x_{t+1})\| \leq\ & \left(1 + \frac{1}{8t_f(\delta)}\right) \|\nabla f(x_t)\| + \eta L c_3 \left(u^2 d^2 \rho \left(\log \frac{T}{\delta} \right)^2 + \sqrt{1 + \frac{\log(T/\delta)}{d}} r \right) \\
=\ & \left(1 + \frac{1}{8t_f(\delta)}\right)^{t+1-t'} \|\nabla f(x_{t'})\| \\
& + \sum_{i=0}^{t-t'} \left(1 + \frac{1}{8t_f(\delta)}\right)^{t-t'-i} \eta L c_3 \left(u^2 d^2 \rho \left(\log \frac{T}{\delta} \right)^2 + \sqrt{1 + \frac{\log(T/\delta)}{d}} r \right) \\
\leq\ & \left(1 + \frac{1}{8t_f(\delta)}\right)^{t'_f} \|\nabla f(x_{t'})\| \\
& + 8t_f(\delta) \left(\left(1 + \frac{1}{8t_f(\delta)}\right)^{t'_f} - 1 \right) \eta L c_3 \left(u^2 d^2 \rho \left(\log \frac{T}{\delta} \right)^2 + \sqrt{1 + \frac{\log(T/\delta)}{d}} r \right) \\
\leq\ & e^{1/4} \cdot 8t_f(\delta)\eta L c_3 \left(u^2 d^2 \rho \left(\log \frac{T}{\delta} \right)^2 + \sqrt{1 + \frac{\log(T/\delta)}{d}} r \right) \\
& + 8t_f(\delta)(e^{1/4} - 1) \cdot \eta L c_3 \left(u^2 d^2 \rho \left(\log \frac{T}{\delta} \right)^2 + \sqrt{1 + \frac{\log(T/\delta)}{d}} r \right)
\end{aligned}$$

$$\leq 16 t_f(\delta)\eta L c_3 \left(u^2 d^2 \rho \left(\log \frac{T}{\delta}\right)^2 + \sqrt{1 + \frac{\log(T/\delta)}{d}} r \right),$$

where we used $t'_f \leq 2t_f(\delta)$ and the fact that $(1 - 1/(8x))^{2x} \leq e^{1/4}$ for all $x > 0$. By defining $c_5 \coloneqq 16c_3$, we complete the proof. □

We next derive a useful result showing that the function change $f(x_\tau) - f(x_0)$ 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 η, u and r. For now, we focus on the case $\tau \geq t_f(\delta)$, 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 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 let $\tau \geq t_f(\delta))$ be arbitrary. Consider splitting $\{0, 1 \ldots, \tau - 1\}$ into $K \coloneqq \lfloor \tau / t_f(\delta) \rfloor$ intervals:*

$$J_k = \{k t_f(\delta), \ldots, (k+1)t_f(\delta) - 1\}, \;\; 0 \leq k < K-1,$$
$$J_{K-1} = \{(K-1)t_f(\delta), \ldots, \tau - 1\}.$$

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

$$\|\nabla f(x_t)\| > 8 t_f(\delta)\eta L \left(\frac{u}{2} \left\| \frac{1}{m} \sum_{i=1}^{m} Z_{t,i} Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i} \right\| + \|Y_t\| \right). \tag{28}$$

Suppose we choose η such that

$$\eta \leq \frac{1}{L t_f(\delta)} \cdot \min \left\{ \frac{\sqrt{m}}{8c_4 (\mathrm{lr}(C_1 dmT/\delta))^{3/2}\sqrt{d}}, \frac{m}{128 c_1 (\mathrm{lr}(C_1 dmT/\delta))^3 d} \right\}. \tag{29}$$

Then, on the event

$$\mathcal{E}_\tau(\delta) \coloneqq \mathcal{H}_\tau(\delta) \cap \left(\bigcap_{t=0}^{\tau-1} \mathcal{A}_t(\delta) \right) \cap \left(\bigcap_{t=0}^{\tau-1} \mathcal{G}_t(\delta) \right) \cap \left(\bigcap_{k=0}^{K-2} \mathcal{B}_{k t_f(\delta)}(\delta; t_f(\delta)) \right) \cap \mathcal{B}_{(K-1)t_f(\delta)}(\delta; \tau - (K-1)t_f(\delta)),$$

we have the following upper bound on function value change:

$$\begin{aligned} f(x_\tau) - f(x_0) \leq & - \sum_{k \in I_1} \frac{\eta}{2} \min_{t \in J_k} \|\nabla f(x_t)\|^2 + \tau \frac{c_5^2}{64} \eta^3 t_f(\delta)^2 L^2 \left(u^2 d^2 \rho \left(\log \frac{T}{\delta}\right)^2 + \sqrt{2\log(T/\delta)} r \right)^2 \\ & + \tau \eta u^4 \rho^2 \cdot c_1 d^3 \left(\log \frac{T}{\delta}\right)^3 + \tau L \eta^2 u^4 \rho^2 \cdot c_1 d^4 \left(\log \frac{T}{\delta}\right)^4 \\ & + \eta c_1 r^2 (128 t_f(\delta) + \eta L) \log \frac{T}{\delta} + \tau c_1 L \eta^2 r^2. \end{aligned} \tag{30}$$

Moreover, $\mathbb{P}(\mathcal{E}_\tau(\delta)) \geq 1 - \frac{(5\tau+4)\delta}{T}$.

Proof. Without loss of generality, we may assume that τ is a multiple of $t_f(\delta)$.[4] Then, any interval $J_k = \{t_0, \ldots, t_0 + t_f(\delta) - 1\}$ belongs to one of the following two cases:

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

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

[4]To accommodate the last interval which has length at most $2t_f(\delta) - 1$, we note that the results we require for the proof, namely Lemma 14, Lemma 16 and Lemma 17, all hold for any interval length $t'_f \leq 2t_f(\delta)$.

By our choice of η in Eq. (29), we can apply Lemma 16 to get

$$\min_{t\in J_k}\|\nabla f(x_t)\| \geq \frac{1}{4}\max_{t\in J_k}\|\nabla f(x_t)\|.$$

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

Note also that on the event $\mathcal{B}_{kt_f(\delta)}(\delta; t_f(\delta))$, there exists some $t \in J_k$ such that

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

This implies then that

$$\begin{aligned}\frac{1}{4}\sum_{t\in J_k}\frac{1}{m}\sum_{i=1}^{m}\left|Z_{t,i}^\top \nabla f(x_t)\right|^2 &\geq \frac{1}{4}\min_{t\in J_k}\|\nabla f(x_t)\|^2 \geq \frac{1}{64}\max_{t\in J_k}\|\nabla f(x_t)\|^2 \\ &\geq \frac{1}{64t_f(\delta)}\sum_{t\in J_k}\|\nabla f(x_t)\|^2. \end{aligned} \tag{31}$$

Thus by setting $\alpha = 128t_f(\delta)$ in Eq. (2) and by choosing η such that

$$\frac{c_1 L\eta^2\chi^3 d}{m} \leq \frac{\eta}{\alpha} = \frac{\eta}{128t_f(\delta)} \quad \Longleftrightarrow \quad \eta \leq \frac{m}{128c_1 L t_f(\delta) d\chi^3},$$

it follows that

$$\begin{aligned} &-\frac{3\eta}{4}\sum_{t\in J_k}\frac{1}{m}\sum_{i=1}^{m}\left|Z_{t,i}^\top \nabla f(x_t)\right|^2 + \left(\frac{\eta}{128t_f(\delta)} + \frac{c_1 L\eta^2\chi^3 d}{m}\right)\sum_{t\in J_k}\|\nabla f(x_t)\|^2 \\ =& -\frac{3\eta}{4}\sum_{t\in J_k}\frac{1}{m}\sum_{i=1}^{m}\left|Z_{t,i}^\top \nabla f(x_t)\right|^2 + \frac{\eta}{64t_f(\delta)}\sum_{t\in J_k}\|\nabla f(x_t)\|^2 \\ \leq& -\frac{\eta}{2}\sum_{t\in J_k}\frac{1}{m}\sum_{i=1}^{m}\left|Z_{t,i}^\top \nabla f(x_t)\right|^2 \\ \leq& -\frac{\eta}{2}\min_{t\in J_k}\|\nabla f(x_t)\|^2 \end{aligned} \tag{32}$$

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

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

By our choice of η in Eq. (29), we can apply Lemma 17 to get

$$\|\nabla f(x_t)\| \leq c_5 t_f(\delta)\eta L\left(u^2d^2\rho\left(\log\frac{T}{\delta}\right)^2 + \sqrt{1+\frac{\log(T/\delta)}{d}}r\right) \qquad \forall t \in J_k.$$

Hence, by setting $\alpha = 128t_f(\delta)$ in Eq. (2) and choosing η such that

$$\frac{c_1 L\eta^2\chi^3 d}{m} \leq \frac{\eta}{\alpha} = \frac{\eta}{128t_f(\delta)},$$

it follows that

$$\begin{aligned} &\left(\frac{\eta}{128t_f(\delta)} + \frac{c_1 L\eta^2\chi^3 d}{m}\right)\sum_{t\in J_k}\|\nabla f(x_t)\|^2 \\ \leq& \frac{\eta}{64t_f(\delta)}\sum_{t\in J_k}\left(c_5 t_f(\delta)\eta L\left(u^2d^2\rho\left(\log\frac{T}{\delta}\right)^2 + \sqrt{1+\frac{\log(T/\delta)}{d}}r\right)\right)^2 \\ \leq& \frac{c_5^2}{64}t_f(\delta)^2\eta^3L^2\left(u^2d^2\rho\left(\log\frac{T}{\delta}\right)^2 + \sqrt{1+\frac{\log(T/\delta)}{d}}r\right)^2 \end{aligned} \tag{33}$$

Without loss of generality, we may assume that τ is a multiple of $t_f(\delta)$.[5] Then, any interval $J_k = \{t_0, \ldots, t_0 + t_f(\delta) - 1\}$ 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 $\{0, 1, \ldots, K-1\}$. Then,

$$
\begin{aligned}
&-\frac{3\eta}{4}\sum_{t=0}^{\tau-1}\frac{1}{m}\sum_{i=1}^{m}\left|Z_{t,i}^{\top}\nabla f(x_t)\right|^2 + \left(\frac{\eta}{\alpha} + \frac{c_1 L\eta^2\chi^3 d}{m}\right)\sum_{t=0}^{\tau-1}\|\nabla f(x_t)\|^2 \\
&= \sum_{k\in I_1}\left(-\frac{3\eta}{4}\sum_{t=\in J_k}\frac{1}{m}\sum_{i=1}^{m}\left|Z_{t,i}^{\top}\nabla f(x_t)\right|^2 + \left(\frac{\eta}{128 t_f(\delta)} + \frac{c_1 L\eta^2\chi^3 d}{m}\right)\sum_{t\in J_k}\|\nabla f(x_t)\|^2\right) \\
&\quad + \sum_{k\in I_1^c}\left(-\frac{3\eta}{4}\sum_{t\in J_k}\frac{1}{m}\sum_{i=1}^{m}\left|Z_{t,i}^{\top}\nabla f(x_t)\right|^2 + \left(\frac{\eta}{128 t_f(\delta)} + \frac{c_1 L\eta^2\chi^3 d}{m}\right)\sum_{t\in J_k}\|\nabla f(x_t)\|^2\right) \\
&\leq -\sum_{k\in I_1}\frac{\eta}{2}\min_{t\in J_k}\|\nabla f(x_t)\|^2 + \sum_{k\in I_1^c} t_f(\delta)\left(\frac{c_5^2}{64}t_f(\delta)^2\eta^3 L^2\left(u^2 d^2\rho\left(\log\frac{T}{\delta}\right)^2 + \sqrt{1+\frac{\log(T/\delta)}{d}}\,r\right)^2\right) \\
&\leq -\sum_{k\in I_1}\frac{\eta}{2}\min_{t\in J_k}\|\nabla f(x_t)\|^2 + \tau\frac{c_5^2}{64}t_f(\delta)^2\eta^3 L^2\left(u^2 d^2\rho\left(\log\frac{T}{\delta}\right)^2 + \sqrt{1+\frac{\log(T/\delta)}{d}}\,r\right)^2.
\end{aligned}
\tag{34}
$$

and so by Eq. (2),

$$
\begin{aligned}
f(x_\tau) - f(x_0) \leq &-\sum_{k\in I_1}\frac{\eta}{2}\min_{t\in J_k}\|\nabla f(x_t)\|^2 + \tau\frac{c_5^2}{64}t_f(\delta)^2\eta^3 L^2\left(u^2 d^2\rho\left(\log\frac{T}{\delta}\right)^2 + \sqrt{1+\frac{\log(T/\delta)}{d}}\,r\right)^2 \\
&+ \tau\eta u^4\rho^2\cdot c_1 d^3\left(\log\frac{T}{\delta}\right)^3 + \tau L\eta^2 u^4\rho^2\cdot c_1 d^4\left(\log\frac{T}{\delta}\right)^4 \\
&+ \eta c_1 r^2(\alpha + \eta L)\log\frac{T}{\delta} + \tau c_1 L\eta^2 r^2.
\end{aligned}
$$

Note that we choose $\alpha = 128 t_f(\delta)$. In addition, observe that by our choice of δ (such that $\delta \leq \frac{1}{e}$), it follows that $\sqrt{1+\frac{\log(T/\delta)}{d}} \leq \sqrt{2\log(T/\delta)}$.

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

$$
\begin{aligned}
\mathbb{P}(\mathcal{E}_\tau^c) &\leq \mathbb{P}(\mathcal{H}_\tau^c) + \sum_{t=0}^{\tau-1}\mathbb{P}(\mathcal{A}_t^c) + \sum_{t=0}^{\tau-1}\mathbb{P}(\mathcal{G}_t^c) + \sum_{k=0}^{K-1}\mathbb{P}(\mathcal{B}_{kt_f(\delta)}^c(\delta; t_f(\delta))) \\
&\leq \frac{(\tau+4)\delta}{T} + \frac{\tau}{T}\delta + 2\frac{\tau}{T}\delta + \frac{K\delta}{T} \leq \frac{(5\tau+4)}{T}\delta. \qquad \square
\end{aligned}
$$

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 5. *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*

$$
\begin{aligned}
&u \leq \frac{\sqrt{\epsilon}}{d\sqrt{\rho}\log(T/\delta)}\cdot\min\left\{\frac{1}{64c_5^2 c_2}, \frac{1}{2048 c_1 c_2}\right\}^{1/4}, \qquad r \leq \epsilon\cdot\min\left\{\frac{1}{8c_5\sqrt{2c_2}}, \frac{1}{32\sqrt{c_1}}\right\}, \\
&\eta \leq \frac{1}{L t_f(\delta)}\min\left\{\frac{1}{\log(T/\delta)}, \frac{\sqrt{m}}{8c_4(\mathrm{lr}(C_1 dmT/\delta))^{3/2}\sqrt{d}}, \frac{m}{128 c_1(\mathrm{lr}(C_1 dmT/\delta))^3 d}\right\},
\end{aligned}
$$

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

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

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 $t_f(\delta)$, and we similarly split $\{0,1,\ldots,T\}$ into $K = \lfloor T/t_f(\delta)\rfloor$ intervals $J_0,\ldots,J_{K-1}$. Let I_1 denote the set of indices k such that for every $t \in J_k$,

$$\|\nabla f(x_t)\| > 8 t_f(\delta)\eta L\left[\left(\frac{u}{2}\left\|\frac{1}{m}\sum_{i=1}^m Z_{t,i}Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i}\right\|\right) + \|Y_t\|\right]. \tag{35}$$

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

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

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 δ in the statement of the lemma, we have

$$\begin{aligned} f(x_T) - f(x_0) \leq & -\sum_{k\in I_1}\frac{\eta}{2}\min_{t\in J_k}\|\nabla f(x_t)\|^2 + T\frac{c_5^2}{64}t_f(\delta)^2\eta^3L^2\left(u^2d^2\rho\left(\log\frac{T}{\delta}\right)^2 + \sqrt{2\log(T/\delta)}r\right)^2 \\ & + T\eta u^4\rho^2\cdot c_1 d^3\left(\log\frac{T}{\delta}\right)^3 + TL\eta^2u^4\rho^2\cdot c_1d^4\left(\log\frac{T}{\delta}\right)^4 \\ & + \eta c_1 r^2(128t_f(\delta)+\eta L)\log\frac{T}{\delta} + Tc_1L\eta^2r^2. \end{aligned} \tag{36}$$

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 $\lceil T/(4t_f(\delta))\rceil$ members. Note that, by our choices of the parameters η, u, r, it can be shown that

$$c_5 t_f(\delta)\eta L\left(u^2d^2\rho\left(\log\frac{T}{\delta}\right)^2 + \sqrt{1+\frac{\log(T/\delta)}{d}}r\right) < \epsilon, \tag{37}$$

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

$$\|\nabla f(x_t)\| \leq c_5 t_f(\delta)\eta L\left(u^2d^2\rho\log(T/\delta) + \sqrt{1+\frac{\log(T/\delta)}{d}}r\right), \qquad \forall t \in J_k.$$

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

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

$$\frac{1}{2}\left\|\nabla f(x_{kt_f(\delta)})\right\| \leq \|\nabla f(x_t)\| \leq 2\left\|\nabla f(x_{kt_f(\delta)})\right\|, \qquad \forall t \in J_k.$$

This implies in particular that for any $k \in I_\epsilon$, we have $\min_{t\in J_k}\|\nabla f(x_t)\|^2 \geq \frac{1}{16}\epsilon^2$, and consequently

$$-\sum_{k\in I_1}\frac{\eta}{2}\min_{t\in J_k}\|\nabla f(x_t)\|^2 \leq -\sum_{k\in I_\epsilon}\frac{\eta}{2}\cdot\frac{\epsilon^2}{16} \leq -\frac{T}{4t_f(\delta)}\cdot\frac{\eta}{2}\cdot\frac{\epsilon^2}{16} = -\frac{T\eta\epsilon^2}{128t_f(\delta)}.$$

Hence, by Eq. (36),

$$f(x_T) - f(x_0) \leq -\frac{T\eta\epsilon^2}{128t_f(\delta)} + T\frac{c_5^2}{64}t_f(\delta)^2\eta^3L^2\left(u^2d^2\rho\left(\log\frac{T}{\delta}\right)^2 + \sqrt{2\log(T/\delta)}r\right)^2$$

$$+ T\eta u^4\rho^2 \cdot c_1 d^3 \left(\log \frac{T}{\delta}\right)^3 + T\eta \cdot (\eta L) u^4 \rho^2 \cdot c_1 d^4 \left(\log \frac{T}{\delta}\right)^4$$
$$+ \eta c_1 r^2 (128 t_f(\delta) + \eta L) \log \frac{T}{\delta} + T\eta \cdot c_1 \eta L r^2. \tag{38}$$

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

$$T \frac{c_5^2}{64} t_f(\delta)^2 \eta^3 L^2 \left(u^2 d^2 \rho \left(\log \frac{T}{\delta}\right)^2 + \sqrt{2\log(T/\delta)} r \right)^2$$
$$\leq T\eta \cdot \frac{c_5^2}{32} t_f(\delta)^2 (\eta L)^2 \left(u^4 d^4 \rho^2 \left(\log \frac{T}{\delta}\right)^4 + 2\log(T/\delta) r^2 \right)$$
$$\leq T\eta \cdot \left(\frac{\epsilon^2}{2048 c_2 \left(\log \frac{T}{\delta}\right)^2} + \frac{\epsilon^2}{2048 c_2 \log(T/\delta)} \right) \leq \frac{T\eta\epsilon^2}{512 t_f(\delta)},$$

where we used $\log(T/\delta) \geq 1$ and $2c_2 \log(T/\delta) \geq t_f(\delta)$. We also have

$$T\eta u^4 \rho^2 \cdot c_1 d^3 \left(\log \frac{T}{\delta}\right)^3 + T\eta \cdot (\eta L) u^4 \rho^2 \cdot c_1 d^4 \left(\log \frac{T}{\delta}\right)^4 + T c_1 L \eta^2 r^2$$
$$\leq T\eta \cdot \frac{\epsilon^2}{2048 c_2 d \log(T/\delta)} + T\eta \cdot \frac{\epsilon^2}{2048 c_2 t_f(\delta) \log(T/\delta)} + T\eta \cdot \frac{\epsilon^2}{1024 t_f(\delta) \log(T/\delta)}$$
$$\leq \frac{T\eta\epsilon^2}{512 t_f(\delta)},$$

where we used $c_2 d \log(T/\delta) \geq t_f(\delta)$, $c_2 \geq 1$ and $\log(T/\delta) \geq 1$. Finally,

$$\eta c_1 r^2 (128 t_f(\delta) + \eta L) \log \frac{T}{\delta} \leq \frac{(128 t_f(\delta) + 1)\epsilon^2}{1024 L t_f(\delta)} < \frac{\epsilon^2}{L}.$$

By plugging these bounds into Eq. (38), we get

$$f(x_T) - f(x_0) < -\frac{T\eta\epsilon^2}{128 t_f(\delta)} + \frac{T\eta\epsilon^2}{512 t_f(\delta)} + \frac{T\eta\epsilon^2}{512 t_f(\delta)} + \frac{\epsilon^2}{L} \leq -\frac{T\eta\epsilon^2}{256 t_f(\delta)} + \frac{\epsilon^2}{L}.$$

Therefore, as long as

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

we will get $f(x_T) < f^*$, which is a contradiction. Thus, we can conclude that on the event $\mathcal{E}_T(\delta)$, there are at most $T/4$ iterations for which $\|\nabla f(x_t)\| \geq \epsilon$.

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

$$\mathbb{P}(\mathcal{E}_T^c) \leq \mathbb{P}(\mathcal{H}_T^c) + \sum_{t=0}^{T-1} \mathbb{P}(\mathcal{A}_t^c) + \sum_{t=0}^{T-1} \mathbb{P}(\mathcal{G}_t^c) + \sum_{k=0}^{K-1} \mathbb{P}(\mathcal{B}_{k t_f(\delta)}^c(\delta; t_f(\delta)))$$
$$\leq \frac{(T+4)\delta}{T} + \delta + 2\delta + \frac{K\delta}{T} \leq 6\delta. \qquad \square$$

E ESCAPING SADDLE POINT

E.1 KEY QUANTITIES AND NOTATION

We will use γ to denote $-\lambda_{\min}(\nabla^2 f(x_0))$, where we know that $\gamma \geq \sqrt{\rho\epsilon}$.

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 $x_{t_0+\tau}$ and x_{t_0} into a few different terms.

Lemma 19. *Consider the perturbed zeroth-order update Algorithm 1. Then, for any $t_0 \in \mathbb{N}$ and $\tau \in \mathbb{N}$,*

$$\|x_{t_0+\tau} - x_{t_0}\|^2 \leq V_1(t_0,\tau) + V_2(t_0,\tau) + V_3(t_0,\tau) + V_4(t_0,\tau), \tag{39}$$

where

$$\begin{aligned} V_1(t_0,\tau) &\coloneqq 8\eta^2\tau \textstyle\sum_{t=t_0}^{t_0+\tau-1} \|\nabla f(x_t)\|^2, & V_2(t_0,\tau) &\coloneqq 8\eta^2 \left\| \textstyle\sum_{t=t_0}^{t_0+\tau-1} \frac{1}{m} \sum_{i=1}^m (Z_{t,i}Z_{t,i}^\top - I)\nabla f(x_t) \right\|^2 \\ V_3(t_0,\tau) &\coloneqq 4\eta^2 \left\| \textstyle\sum_{t=t_0}^{t_0+\tau-1} Y_t \right\|^2, & V_4(t_0,\tau) &\coloneqq 4\eta^2 \left\| \textstyle\sum_{t=t_0}^{t_0+\tau-1} \frac{1}{m} \sum_{i=1}^m u Z_{t,i} Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i} \right\|^2. \end{aligned} \tag{40}$$

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

$$\begin{aligned}
&\|x_\tau - x_0\|^2 \\
&= \left\| \sum_{t=0}^{\tau-1} x_{t+1} - x_t \right\|^2 \\
&= \eta^2 \left\| \sum_{t=0}^{\tau-1} \frac{1}{m} \sum_{i=1}^m Z_{t,i}Z_{t,i}^\top \nabla f(x_t) + \frac{1}{m}\sum_{i=1}^m u Z_{t,i} Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i} + Y_t \right\|^2 \\
&\leq 4\eta^2 \left\| \sum_{t=0}^{\tau-1} \frac{1}{m} \sum_{i=1}^m Z_{t,i}Z_{t,i}^\top \nabla f(x_t) \right\|^2 + 4\eta^2 \left\| \sum_{t=0}^{\tau-1} \frac{1}{m}\sum_{i=1}^m u Z_{t,i} Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i} \right\|^2 + 4\eta^2 \left\| \sum_{t=0}^{\tau-1} Y_t \right\|^2 \\
&\leq 4\eta^2 \left\| \sum_{t=0}^{\tau-1} \frac{1}{m} \sum_{i=1}^m (Z_{t,i}Z_{t,i}^\top - I) \nabla f(x_t) + \sum_{t=0}^{\tau-1} \nabla f(x_t) \right\|^2 + 4\eta^2 \left\| \sum_{t=0}^{\tau-1} \frac{1}{m}\sum_{i=1}^m u Z_{t,i} Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i} \right\|^2 + 4\eta^2 \left\| \sum_{t=0}^{\tau-1} Y_t \right\|^2 \\
&\leq \underbrace{8\eta^2\tau \sum_{t=0}^{\tau-1} \|\nabla f(x_t)\|^2}_{V_1(0,\tau)} + \underbrace{8\eta^2 \left\| \sum_{t=0}^{\tau-1} \frac{1}{m} \sum_{i=1}^m (Z_{t,i}Z_{t,i}^\top - I)\nabla f(x_t) \right\|^2}_{V_2(0,\tau)} + \underbrace{4\eta^2 \left\| \sum_{t=0}^{\tau-1} Y_t \right\|^2}_{V_3(0,\tau)} + \underbrace{4\eta^2 \left\| \sum_{t=0}^{\tau-1} \frac{1}{m}\sum_{i=1}^m u Z_{t,i} Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i} \right\|^2}_{V_4(0,\tau)}.
\end{aligned}$$

□

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

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

Lemma 20. *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 η such that

$$\eta \leq \frac{1}{Lt_f(\delta)} \cdot \min\left\{ \frac{\sqrt{m}}{8c_4(\mathrm{lr}(C_1 dmT/\delta))^{3/2}\sqrt{d}}, \frac{m}{128c_1(\mathrm{lr}(C_1 dmT/\delta))^3 d} \right\}.$$

There are two cases to consider.

1. *The first is when $\tau \geq t_f(\delta)$. In this case, split $\{t_0, t_0+1, \ldots, t_0+\tau-1\}$ into $K \coloneqq \lfloor \tau/t_f(\delta) \rfloor$ intervals:*

$$J_k = \{t_0 + kt_f(\delta), \ldots, t_0 + (k+1)t_f(\delta) - 1\}, \quad 0 \leq k < K-1,$$

$$J_{K-1} = \{t_0 + (K-1)t_f(\delta), \ldots, t_0 + \tau - 1\}.$$

Then, on the event

$$\mathcal{E}_{t_0,\tau}(\delta) := \mathcal{H}_{t_0,\tau}(\delta) \cap \left(\bigcap_{t=t_0}^{t_0+\tau-1} \mathcal{A}_t(\delta) \right) \cap \left(\bigcap_{t=t_0}^{t_0+\tau-1} \mathcal{G}_t(\delta) \right) \cap \left(\bigcap_{k=0}^{K-2} \mathcal{B}_{t_0+kt_f(\delta)}(\delta; t_f(\delta)) \right) \cap \mathcal{B}_{t_0+(K-1)t_f(\delta)}(\delta; \tau - (K-1)t_f(\delta)),$$

we have that

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

where

$$\begin{aligned} N_{u,r}(\tau; \delta) := \tau \frac{c_5^2}{64} \eta^3 t_f(\delta)^2 L^2 &\left(u^2 d^2 \rho \left(\log \frac{T}{\delta} \right)^2 + \sqrt{2\log(T/\delta)} r \right)^2 \\ &+ \tau\eta u^4 \rho^2 \cdot c_1 d^3 \left(\log \frac{T}{\delta} \right)^3 + \tau L \eta^2 u^4 \rho^2 \cdot c_1 d^4 \left(\log \frac{T}{\delta} \right)^4 \\ &+ \eta c_1 r^2 (128 t_f(\delta) + \eta L) \log \frac{T}{\delta} + \tau c_1 L \eta^2 r^2 \\ &+ c_5^2 t_f^3(\delta) \eta^3 L^2 \left(u^2 d^2 \rho \log(T/\delta) + \sqrt{2\log(T/\delta)} r \right)^2. \end{aligned} \tag{41}$$

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

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

Suppose the event $\cap_{t=t_0}^{t_0+\tau-1}(\mathcal{A}_t(\delta) \cap \mathcal{G}_t(\delta)$ holds. Suppose also that $\|\nabla f(x_{t_0})\| \leq \epsilon$. Then,

$$V_1(t_0, \tau) \leq 32\eta^2\tau^2\epsilon^2 \leq 32\eta^2 (t_f(\delta))^2 \epsilon^2$$

Proof. 1. We first consider the case where $\tau \geq t_f(\delta)$. 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

$$\|\nabla f(x_t)\| > 8t_f(\delta)\eta L \left(\frac{u}{2} \left\| \frac{1}{m} \sum_{i=1}^m Z_{t,i} Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i} \right\| + \|Y_t\| \right). \tag{42}$$

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 $t_f(\delta)$. From Lemma 18, on the event that $\mathcal{E}_\tau(\delta)$ holds and by our choice of η, we have

$$\begin{aligned} f(x_\tau) - f(x_0) \leq &- \sum_{k \in I_1} \frac{\eta}{2} \min_{t \in J_k} \|\nabla f(x_t)\|^2 + \tau \frac{c_5^2}{64} \eta^3 t_f(\delta)^2 L^2 \left(u^2 d^2 \rho \left(\log \frac{T}{\delta} \right)^2 + \sqrt{2\log(T/\delta)} r \right)^2 \\ &+ \tau\eta u^4 \rho^2 \cdot c_1 d^3 \left(\log \frac{T}{\delta} \right)^3 + \tau L \eta^2 u^4 \rho^2 \cdot c_1 d^4 \left(\log \frac{T}{\delta} \right)^4 \\ &+ \eta c_1 r^2 (128 t_f(\delta) + \eta L) \log \frac{T}{\delta} + \tau c_1 L \eta^2 r^2. \end{aligned}$$

By Lemma 16 (and our choice of η), it follows that for any $k \in I_1$, on the event $\cap_{t \in J_k} \mathcal{A}_t(\delta)$, we have

$$\sum_{t \in J_k} \|\nabla f(x_t)\|^2 \leq 4t_f \min_{t \in J_k} \|\nabla f(x_t)\|^2.$$

Thus, on the event that $\mathcal{E}_\tau(\delta)$ holds, for our choice of η, we have

$$\eta \sum_{k \in I_1} \sum_{t \in J_k} \|\nabla f(x_t)\|^2 \leq 4t_f(\delta)\eta \sum_{k \in I_1} \min_{t \in J_k} \|\nabla f(x_t)\|^2$$

$$
\begin{aligned}
&\leq 8t_f(\delta)\sum_{k\in I_1}\frac{\eta}{2}\min_{t\in J_k}\|\nabla f(x_t)\|^2 \\
&\leq 8t_f(\delta)\left((f(x_0)-f(x_\tau))+\tau\frac{c_5^2}{64}\eta^3 t_f(\delta)^2L^2\left(u^2d^2\rho\left(\log\frac{T}{\delta}\right)^2+\sqrt{2\log(T/\delta)}r\right)^2\right) \\
&\quad + \ 8t_f(\delta)\left(\tau\eta u^4\rho^2\cdot c_1d^3\left(\log\frac{T}{\delta}\right)^3+\tau L\eta^2u^4\rho^2\cdot c_1d^4\left(\log\frac{T}{\delta}\right)^4\right) \\
&\quad + \ 8t_f(\delta)\left(\eta c_1r^2(128t_f(\delta)+\eta L)\log\frac{T}{\delta}+\tau c_1L\eta^2r^2\right).
\end{aligned}
$$

Similarly, for any $k\in I_1^c$ (where I_1^c denotes the complement of I_1 in $\{0,1,\ldots,K-1\}$, i.e. intervals where the gradient is smaller than than the perturbation terms in some iteration), on the event $(\cap_{t\in J_k}\mathcal{A}_t(\delta))\cap(\cap_{t\in J_k}\mathcal{G}_t(\delta))$, by Lemma 17 (and our choice of η), we have

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

On the event that $\mathcal{E}_\tau(\delta)$ holds, this gives us then

$$
\eta\sum_{k\in I_1^c}\sum_{t\in J_k}\|\nabla f(x_t)\|^2\leq\eta\tau\left(c_5^2t_f^2(\delta)\eta^2L^2\left(u^2d^2\rho\log(T/\delta)+\sqrt{2\log(T/\delta)}r\right)^2\right).
$$

Hence, on the event that $\mathcal{E}_\tau(\delta)$ holds, we have that

$$
\begin{aligned}
\eta\sum_{t=0}^{\tau-1}\|\nabla f(x_t)\|^2 &= \eta\sum_{k\in I_1}\sum_{t\in J_k}\|\nabla f(x_t)\|^2+\eta\sum_{k\in I_1^c}\sum_{t\in J_k}\|\nabla f(x_t)\|^2 \\
&\leq 8t_f(\delta)\left((f(x_0)-f(x_\tau))+\tau\frac{c_5^2}{64}\eta^3t_f(\delta)^2L^2\left(u^2d^2\rho\left(\log\frac{T}{\delta}\right)^2+\sqrt{2\log(T/\delta)}r\right)^2\right) \\
&\quad + \ 8t_f(\delta)\left(\tau\eta u^4\rho^2\cdot c_1d^3\left(\log\frac{T}{\delta}\right)^3+\tau L\eta^2u^4\rho^2\cdot c_1d^4\left(\log\frac{T}{\delta}\right)^4\right) \\
&\quad + \ 8t_f(\delta)\left(\eta c_1r^2(128t_f(\delta)+\eta L)\log\frac{T}{\delta}+\tau c_1L\eta^2r^2\right) \\
&\quad + \ 8t_f(\delta)\eta\tau\left(c_5^2t_f^2(\delta)\eta^2L^2\left(u^2d^2\rho\log(T/\delta)+\sqrt{2\log(T/\delta)}r\right)^2\right).
\end{aligned}
$$

This yields the final result for the case $\tau\geq t_f(\delta)$.

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

$$
\mathcal{E}(t_0,t_0+\tau,\delta):=\cap_{t=t_0}^{t_0+\tau-1}\left\{\|\nabla f(x_t)\|>8t_f(\delta)\eta L\left(\frac{u}{2}\left\|\frac{1}{m}\sum_{i=1}^m Z_{t,i}Z_{t,i}^\top\tilde{H}_{t,i}Z_{t,i}\right\|+\|Y_t\|\right)\right\}
$$

There are two cases to consider.

(a) On the event $\mathcal{E}(t_0,t_0+\tau,\delta)\cap\left(\cap_{t=t_0}^{t_0+\tau-1}\mathcal{A}_t(\delta)\right)$, we have by Lemma 16 that $\|\nabla f(x_t)\|\leq 2\|\nabla f(x_0)\|$ for each $t\in\{0,1,\ldots,\tau-1\}$. Then,

$$
V_1(t_0,\tau)=8\eta^2\tau\sum_{t=t_0}^{t_0+\tau-1}\|\nabla f(x_t)\|^2\leq 8\eta^2\tau^2\left(4\|\nabla f(x_0)\|^2\right)\leq 32\eta^2\tau^2\epsilon^2,
$$

where the final inequality uses the assumption that $\|\nabla f(x_0)\|\leq\epsilon$.

(b) Suppose the event $\mathcal{E}^c(t_0,t_0+\tau,\delta)\cap\left(\cap_{t=t_0}^{t_0+\tau-1}\mathcal{A}_t(\delta)\right)\cap\left(\cap_{t=t_0}^{t_0+\tau-1}\mathcal{G}_t(\delta)\right)$ holds. In this case, by Lemma 17, we have that for each $t\in\{t_0,t_0+1,\ldots,t_0+\tau-1\}$

$$
\|\nabla f(x_t)\|\leq c_5t_f(\delta)\eta L\left(u^2d^2\rho\left(\log\frac{T}{\delta}\right)^2+\sqrt{1+\frac{\log(T/\delta)}{d}}r\right)
$$

$$\leq \epsilon,$$

where the final inequality follows by our choice of η, u and r (cf. Eq. (37)). Hence,

$$\begin{aligned} V_1(t_0,\tau) &= 8\eta^2\tau \sum_{t=t_0}^{t_0+\tau-1} \|\nabla f(x_t)\|^2 \\ &\leq 8\eta^2\tau^2 \left(c_5 t_f(\delta)\eta L \left(u^2 d^2 \rho \left(\log \frac{T}{\delta}\right)^2 + \sqrt{1+\frac{\log(T/\delta)}{d}} r \right)\right)^2 \\ &\leq 8\eta^2\tau^2\epsilon^2 < 32\eta^2\tau^2\epsilon^2. \end{aligned}$$

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

□

We proceed to bound $V_2(t_0,\tau)$.

Lemma 21. *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 and $\tau > 0$ be arbitrary. Suppose we choose η such that*

$$\eta \leq \frac{1}{L t_f(\delta)} \cdot \min\left\{ \frac{\sqrt{m}}{8c_4(\mathrm{lr}(C_1 dmT/\delta))^{3/2}\sqrt{d}}, \frac{m}{128 c_1 (\mathrm{lr}(C_1 dmT/\delta))^3 d} \right\}.$$

Let T_s denote an integer such that $T_s \geq \max\{\tau, t_f(\delta)\}$, and for any $F > 0$, define

$$B(\delta;F) := \frac{8t_f(\delta)(F+N_{u,r}(T_s,\delta))}{\eta}\left(T_s + \frac{d}{m}\right)(\mathrm{lr}(CT^2/\delta))^2, \qquad b_\tau(\delta;F) := \frac{t_f(\delta)\tau F}{\eta}.$$

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

$$\text{either} \quad \sum_{t=t_0}^{t_0+\tau-1} \frac{d}{m}(\mathrm{lr}(CT^2/\delta))^2\|\nabla f(x_t)\|^2 \geq B(\delta;F)$$

$$\text{or} \quad \sqrt{\frac{V_2(t_0,\tau)}{8\eta^2}} \leq c' \sqrt{\max\left\{\sum_{t=t_0}^{t_0+\tau-1}\frac{d}{m}(\mathrm{lr}(CT^2/\delta))^2\|\nabla f(x_t)\|^2, b_\tau(\delta;F)\right\}\left(\log\left(\frac{CT^2}{\delta}\right)+\log\left(\log\left(\frac{B(\delta;F)}{b_\tau(\delta;F)}\right)+1\right)\right)}$$

holds as $\mathcal{L}_{t_0,\tau}(\delta;F)$[6]. We show that $\mathbb{P}(\mathcal{L}_{t_0,\tau}(\delta;F)) \geq 1 - \frac{\delta}{T}$. Finally, denote the event $\mathcal{M}_{t_0,T_s}(F)$ as the event that $f(x_{t_0}) - f(x_{t_0+T_s}) < F$.

Then, on the event $\mathcal{L}_{t_0,\tau}(\delta) \cap \mathcal{E}_{t_0,T_s}(\delta) \cap \mathcal{M}_{t_0,T_s}(F)$ (where $\mathcal{E}_{0,T_s}(\delta)$ is as defined in Lemma 20),

$$V_2(t_0,\tau) \leq 8c'^2\beta_1(\delta;F)\eta t_f(\delta)\max\left\{\frac{8d}{m}(\mathrm{lr}(CT^2/\delta))^2\left(F+N_{u,r}(T_s,\delta)\right), \tau F\right\}, \tag{43}$$

where

$$\beta_1(\delta;F) := \log\left(\frac{CT^2}{\delta}\right) + \log\left(\log\left(\frac{B(\delta;F)}{b_1(\delta;F)}\right)+1\right).$$

Proof. We note that $\mathbb{P}(\mathcal{L}_{t_0,\tau}(\delta;F)) \geq 1 - \frac{\delta}{T}$. 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 $\mathcal{L}_{0,\tau}(\delta;F) \cap \mathcal{E}_{0,T_s}(\delta) \cap \mathcal{M}_{t_0,T_s}(F)$, suppose that

$$\begin{aligned} &\sum_{t=0}^{\tau-1}\frac{d}{m}(\mathrm{lr}(CT^2/\delta))^2\|\nabla f(x_t)\|^2 \geq B(\delta;F) = \frac{8t_f(\delta)(F+N_{u,r}(T_s,\delta))}{\eta}\left(T_s+\frac{d}{m}\right)(\mathrm{lr}(CT^2/\delta))^2 \\ &\implies \eta\sum_{t=0}^{\tau-1}\|\nabla f(x_t)\|^2 \geq 8t_f(\delta)(F+N_{u,r}(T_s,\delta)) \end{aligned}$$

[6]We note that by construction, $B(\delta;F) \geq b_\tau(\delta;F)$

$$\implies \eta \sum_{t=0}^{T_s-1} \|\nabla f(x_t)\|^2 \geq 8t_f(\delta)(F + N_{u,r}(T_s,\delta))$$
$$\implies 8\eta^2 T_s \sum_{t=0}^{T_s-1} \|\nabla f(x_t)\|^2 \geq 64\eta T_s t_f(\delta)(F + N_{u,r}(T_s,\delta))$$
$$\implies 8\eta^2 T_s \sum_{t=0}^{T_s-1} \|\nabla f(x_t)\|^2 \geq 64\eta T_s t_f(\delta)(f(x_0) - f(x_{T_s}) + N_{u,r}(T_s,\delta)), \quad \text{since } f(x_0) - f(x_{T_s}) \leq F$$
$$\iff V_1(0,T_s) \geq 64\eta T_s t_f(\delta)(f(x_0) - f(x_{T_s}) + N_{u,r}(T_s,\delta)),$$

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

$$\beta_\tau(\delta;F) := \log\left(\frac{CT^2}{\delta}\right) + \log\left(\log\left(\frac{B(\delta;F)}{b_\tau(\delta;F)}\right) + 1\right).$$

Observe that β_1 is larger than β_τ for every $\tau \geq 1$. Since $\mathcal{L}_{t_0,\tau}(\delta;F)$ holds, we must have then that

$$\sqrt{\frac{V_2(0,\tau)}{8\eta^2}} \leq c'\sqrt{\max\left\{\sum_{t=0}^{\tau-1}\frac{d}{m}(\mathrm{lr}(CT^2/\delta))^2\|\nabla f(x_t)\|^2, b_\tau(\delta;F)\right\}\beta_1(\delta;F)}.$$

Now, continuing, recalling the definition of $V_1(0,T_s) = 8\eta^2 T_s \sum_{t=0}^{T_s-1}\|\nabla f(x_t)\|^2$

$$\begin{aligned}
V_2(0,\tau) &\leq c'^2\beta_1(\delta;F)\max\left\{8\eta^2\sum_{t=0}^{\tau-1}\frac{d}{m}(\mathrm{lr}(CT^2/\delta))^2\|\nabla f(x_t)\|^2, 8\eta^2 b_\tau(\delta;F)\right\} \\
&\leq c'^2\beta_1(\delta;F)\max\left\{8\eta^2\sum_{t=0}^{T_s-1}\frac{d}{m}(\mathrm{lr}(CT^2/\delta))^2\|\nabla f(x_t)\|^2, 8\eta^2 b_\tau(\delta;F)\right\} \\
&\leq c'^2\beta_1(\delta;F)\max\left\{\frac{d}{m}(\mathrm{lr}(CT^2/\delta))^2\frac{V_1(0,T_s)}{T_s}, 8\eta t_f(\delta)\tau F\right\} \\
&\overset{\text{(i)}}{\leq} c'^2\beta_1(\delta;F)\max\left\{\frac{d}{m}(\mathrm{lr}(CT^2/\delta))^2\left(64\eta t_f(\delta)(f(x_0) - f(x_{T_S}) + N_{u,r}(T_s,\delta))\right), 8\eta t_f(\delta)\tau F\right\} \\
&\overset{\text{(ii)}}{\leq} c'^2\beta_1(\delta;F)\max\left\{\frac{d}{m}(\mathrm{lr}(CT^2/\delta))^2\left(64\eta t_f(\delta)(F + N_{u,r}(T_s,\delta))\right), 8\eta t_f(\delta)\tau F\right\} \\
&= c'^2\beta_1(\delta;F)(8\eta t_f(\delta))\max\left\{\frac{d}{m}(\mathrm{lr}(CT^2/\delta))^2\left(8(F + N_{u,r}(T_s,\delta))\right), \tau F\right\}.
\end{aligned}$$

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

□

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

Lemma 22. *Let $c > 0$ denote the same constant in Lemma 7. Consider any arbitrary $0 < \delta \leq 1/e$, and let $\tau \geq t_f(\delta)$ be arbitrary. Let $\mathcal{N}_{t_0,\tau}(\delta)$ denote the event that*

$$V_3(t_0,\tau) := 4\eta^2\left\|\sum_{t=t_0}^{t_0+\tau-1} Y_t\right\|^2 \leq 4c_6\eta^2\tau\log(2dT/\delta)r^2,$$

where $c_6 > 0$ is an absolute constant. Then, by Lemma 7, $\mathbb{P}(\mathcal{N}_{t_0,\tau}(\delta)) \geq 1 - \frac{\delta}{T}$. Denote the event

$$\mathcal{O}_t(\delta) := \left\{\frac{1}{m}\sum_{i=1}^m \|Z_{t,i}\|^8 \leq c_7 d^4\left(\log\left(\frac{T}{\delta}\right)\right)^4\right\},$$

where $c_7 > 0$ is an absolute constant. Then, on the event $\cap_{t=t_0}^{t_0+\tau-1}\mathcal{O}_t(\delta)$, we have

$$V_4(t_0,\tau) \leq 4c_7\eta^2\tau^2\rho^2u^4d^4\left(\log\left(\frac{T}{\delta}\right)\right)^4.$$

Moreover, for each t, $\mathbb{P}(\mathcal{O}_t(\delta)) \geq 1 - \frac{\delta}{T}$.

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

$$\begin{aligned} V_4(t_0,\tau) &= 4\eta^2 \left\| \sum_{t=0}^{\tau-1} \frac{1}{m} \sum_{i=1}^{m} u Z_{t,i} Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i} \right\|^2 \\ &\leq 4\eta^2 \tau \left(\sum_{t=t_0}^{t_0+\tau-1} \left\| \frac{1}{m} \sum_{i=1}^{m} u Z_{t,i} Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i} \right\|^2 \right) \\ &\overset{\text{(iii)}}{\leq} 4\eta^2 \tau \sum_{t=t_0}^{t_0+\tau-1} \frac{1}{m} \sum_{i=1}^{m} \rho^2 u^4 \|Z_{t,i}\|^8 \\ &\leq 4c_7 \eta^2 \tau^2 \rho^2 u^4 d^4 \left(\log\left(\frac{T}{\delta}\right)\right)^4. \end{aligned}$$

Above, to derive (iii), we used the bound that $\left\|\tilde{H}_{t,i}\right\| \leq \rho u \|Z_{t,i}\|$. The final inequality is a consequence of our assumption that $\cap_{t=t_0}^{t_0+\tau-1} \mathcal{O}_t(\delta)$ holds. Finally, the result that $\mathbb{P}(\mathcal{O}_t(\delta)) \geq 1 - \frac{\delta}{T}$ 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 $c' > 0, c_1 > 0, c_2 \geq 1, c_4 > 0, c_5 > 0, c_6 > 0, c_7 > 0, C_1 \geq 1$ be the absolute constants defined in the statements of the previous lemmas, and let $\delta \in (0, 1/e]$ be arbitrary. Consider any $T_s \geq t_f(\delta)$. For any $F > 0$, suppose $f(x_{T_s}) - f(x_0) > -F$, i.e. $f(x_0) - f(x_{T_s}) < F$. Suppose that the event*

$$\mathcal{P}_{t_0,T_s}(\delta, F) := \cap_{\tau=1}^{T_s} \left(\mathcal{L}_{t_0,\tau}(\delta; F) \cap \mathcal{N}_{t_0,\tau}(\delta)\right) \cap \left(\cap_{t=t_0}^{t_0+T_s-1} \mathcal{O}_t(\delta) \cap \mathcal{A}_t(\delta) \cap \mathcal{G}_t(\delta)\right) \cap \left(\cap_{\tau=t_f(\delta)}^{T_s-1} \mathcal{E}_{t_0,\tau}(\delta)\right)$$

holds, where the events $\mathcal{E}_{t_0,\tau}(\delta), \mathcal{L}_{t_0,\tau}(\delta), \mathcal{N}_{t_0,\tau}(\delta), \mathcal{O}_t(\delta)$ are as defined in Lemma 20, Lemma 21 and Lemma 22, and $\mathcal{G}_t(\delta)$ and $\mathcal{A}_t(\delta)$ are as defined in Lemma 13 and Lemma 15. Suppose we choose u, r and η such that

$$\begin{aligned} &u \leq \frac{\sqrt{\epsilon}}{d\sqrt{\rho}\log(T/\delta)} \cdot \min\left\{\frac{1}{64c_5^2 c_2}, \frac{1}{2048 c_1 c_2}\right\}^{1/4}, \qquad r \leq \epsilon \cdot \min\left\{\frac{1}{8c_5\sqrt{2c_2}}, \frac{1}{32\sqrt{c_1}}\right\}, \\ &\eta \leq \frac{1}{L t_f(\delta)} \min\left\{\frac{1}{\log(T/\delta)}, \frac{\sqrt{m}}{8c_4(\mathrm{lr}(C_1 dmT/\delta))^{3/2}\sqrt{d}}, \frac{m}{128 c_1 (\mathrm{lr}(C_1 dmT/\delta))^3 d}\right\}. \end{aligned}$$

Suppose $\eta \leq \min\left\{1, \frac{1}{t_f(\delta)}, \frac{1}{t_f \delta L}\right\}$. Suppose also we pick u and r small enough such that

$$u \leq \frac{r^{1/2}}{d\log(T/\delta)\rho^{1/2}}, \quad r^2 \leq \min\left\{\frac{F}{\eta T_s \log(T/\delta)\left(\frac{65c_5^2}{8} + 132c_1 + 1\right)}, \frac{F}{4c_6\log(2dT/\delta) + 4c_7\eta T_s}\right\}.$$

Then, for each $\tau \in \{0, 1, \ldots, T_s\}$, we have that

$$\|x_{t_0+\tau} - x_{t_0}\|^2 \leq \phi_{T_s}(\delta, F),$$

where

$$\phi_{T_s}(\delta, F) \leq \max\left\{128\eta T_s t_f(\delta) F, 32\eta^2 (t_f(\delta))^2 \epsilon^2\right\} + 8c'^2 \beta_1(\delta; F)\eta t_f(\delta) \max\left\{\frac{16d}{m}(\mathrm{lr}(CT^2/\delta))^2 F, T_s F\right\} + T_s \eta t_f(\delta) F,$$

where $\beta_1(\delta; F)$ is defined as in Lemma 21. Moreover, $\mathbb{P}(\mathcal{P}_{t_0,T_s}(\delta, F)) \geq 1 - \frac{12T_s\delta}{T}$.

Proof. We recall that

$$\|x_{t_0+\tau} - x_{t_0}\|^2 \leq \underbrace{8\eta^2\tau \sum_{t=t_0}^{t_0+\tau-1} \|\nabla f(x_t)\|^2}_{V_1(t_0,\tau)} + \underbrace{8\eta^2 \left\| \sum_{t=t_0}^{t_0+\tau-1} \frac{1}{m} \sum_{i=1}^{m} (Z_{t,i} Z_{t,i}^\top - I) \nabla f(x_t) \right\|^2}_{V_2(t_0,\tau)}$$
$$+ \underbrace{4\eta^2 \left\| \sum_{t=t_0}^{t_0+\tau-1} Y_t \right\|^2}_{V_3(t_0,\tau)} + \underbrace{4\eta^2 \left\| \sum_{t=t_0}^{t_0+\tau-1} \frac{1}{m} \sum_{i=1}^{m} u Z_{t,i} Z_{t,i}^\top \tilde{H}_{t,i} Z_{t,i} \right\|^2}_{V_4(t_0,\tau)}$$

By Lemma 20, Lemma 21, and Lemma 22, which bound $V_1(t_0,\tau)$, $V_2(t_0,\tau)$, and $V_3(t_0,\tau)$, $V_4(t_0,\tau)$ respectively, on the event $\mathcal{P}_{t_0,T_s}(\delta, F)$, we have, for any $0 \leq \tau \leq T_s$,

$$\begin{aligned}
\|x_\tau - x_0\|^2 &\leq V_1(0,\tau) + V_2(0,\tau) + V_3(0,\tau) + V_4(0,\tau) \\
&\leq \max\left\{64\eta\tau t_f(\delta)(F + N_{u,r}(\tau;\delta)), 32\eta^2 (t_f(\delta))^2 \epsilon^2\right\} \\
&\quad + 8c'^2 \beta_1(\delta; F)\eta t_f(\delta) \max\left\{\frac{8d}{m}(\mathrm{lr}(CT^2/\delta))^2 \left(F + N_{u,r}(T_s,\delta)\right), \tau F\right\} \\
&\quad + 4c_6\eta^2\tau \log(2dT/\delta) r^2 + 4c_7\eta^2\tau^2\rho^2 u^4 d^4 \left(\log(T/\delta)\right)^4,
\end{aligned}$$

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

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

$$4c_6\eta^2 T_s \log(2dT/\delta) r^2 + 4c_7\eta^2 T_s^2 \rho^2 u^4 d^4 \left(\log(T/\delta)\right)^4 \leq \eta T_s t_f(\delta) F.$$

First, our choice of u ensures that

$$u^4 d^4 \rho^2 (\log(T/\delta))^4 \leq r^2.$$

Next, recall that

$$\begin{aligned}
N_{u,r}(\tau;\delta) \coloneqq\ & \tau \frac{c_5^2}{64} \eta^3 t_f(\delta)^2 L^2 \left(u^2 d^2 \rho \left(\log\frac{T}{\delta}\right)^2 + \sqrt{2\log(T/\delta)} r\right)^2 \\
&+ \tau\eta u^4 \rho^2 \cdot c_1 d^3 \left(\log\frac{T}{\delta}\right)^3 + \tau L \eta^2 u^4 \rho^2 \cdot c_1 d^4 \left(\log\frac{T}{\delta}\right)^4 \\
&+ \eta c_1 r^2 (128 t_f(\delta) + \eta L) \log\frac{T}{\delta} + \tau c_1 L \eta^2 r^2 \\
&+ c_5^2 t_f^3(\delta) \eta^3 L^2 \left(u^2 d^2 \rho \log(T/\delta) + \sqrt{2\log(T/\delta)} r\right)^2.
\end{aligned}$$

Recalling our choice of η such that

$$\eta \leq \min\{1, \frac{1}{t_f(\delta)}, \frac{1}{t_f(\delta) L}\},$$

it follows that

$$\begin{aligned}
N_{u,r}(T_s;\delta) &\leq \eta T_s r^2 \left(\frac{8c_5^2}{64}\log(T/\delta) + 2c_1 + 2c_1 + (128c_1 + 1)\log(T/\delta) + c_1 + 8c_5^2 \log(T/\delta)\right) \\
&\leq \eta T_s r^2 \log(T/\delta) \left(\frac{65c_5^2}{8} + 132c_1 + 1\right) \leq F,
\end{aligned}$$

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

$$4c_6\eta^2 T_s \log(2dT/\delta) r^2 + 4c_7\eta^2 T_s^2 \rho^2 u^4 d^4 \left(\log(T/\delta)\right)^4$$

$$\leq \eta T_s t_f(\delta) \left(4c_6\eta \log(2dT/\delta) r^2 + 4c_7\eta T_s \rho^2 u^4 d^4 (\log(T/\delta))^4\right)$$
$$\leq \eta T_s t_f(\delta) \left(4c_6\eta \log(2dT/\delta) r^2 + 4c_7\eta T_s r^2\right)$$

By choosing r such that

$$r^2 \leq \frac{F}{4c_6 \log(2dT/\delta) + 4c_7\eta T_s},$$

it follows that

$$4c_6\eta^2 T_s \log(2dT/\delta) r^2 + 4c_7\eta^2 T_s^2 \rho^2 u^4 d^4 \left(\log(T/\delta)\right)^4 \leq \eta T_s t_f(\delta) F,$$

as desired.

We next lower bound the probability of

$$\mathcal{P}_{t_0,T_s}(\delta, F) := \cap_{\tau=1}^{T_s} \left(\mathcal{L}_{t_0,\tau}(\delta; F) \cap \mathcal{N}_{t_0,\tau}(\delta)\right) \cap \left(\cap_{t=t_0}^{t_0+T_s-1} \mathcal{O}_t(\delta) \cap \mathcal{A}_t(\delta) \cap \mathcal{G}_t(\delta)\right) \cap \left(\cap_{\tau=t_f(\delta)}^{T_s} \mathcal{E}_{t_0,\tau}(\delta)\right).$$

Observe that

$$\begin{aligned}
&\cap_{\tau=t_f(\delta)}^{T_s} \mathcal{E}_{t_0,\tau}(\delta) \\
&= \cap_{\tau=t_f(\delta)}^{T_s} \left(\mathcal{H}_{t_0,\tau}(\delta) \cap \left(\bigcap_{t=t_0}^{t_0+\tau-1} \mathcal{A}_t(\delta) \cap \mathcal{G}_t(\delta)\right) \cap \left(\bigcap_{k=0}^{K-2} \mathcal{B}_{t_0+kt_f(\delta)}(\delta; t_f(\delta))\right) \cap \mathcal{B}_{t_0+(K-1)t_f(\delta)}(\delta; \tau-(K-1)t_f(\delta))\right) \\
&= \cap_{\tau=t_f(\delta)}^{T_s} \left(\mathcal{H}_{t_0,\tau}(\delta) \cap \left(\bigcap_{k=0}^{K-2} \mathcal{B}_{t_0+kt_f(\delta)}(\delta; t_f(\delta))\right) \cap \mathcal{B}_{t_0+(K-1)t_f(\delta)}(\delta; \tau-(K-1)t_f(\delta))\right) \cap \left(\bigcap_{t=t_0}^{T_s-1} \mathcal{A}_t(\delta) \cap \mathcal{G}_t(\delta)\right).
\end{aligned}$$

Note this implies that $\cap_{\tau=t_f(\delta)}^{T_s} \mathcal{E}_{t_0,\tau}(\delta) \cap \left(\cap_{t=t_0}^{T_s-1} \mathcal{A}_t(\delta) \cap \mathcal{G}_t(\delta)\right) = \cap_{\tau=t_f(\delta)}^{T_s} \mathcal{E}_{t_0,\tau}(\delta)$ We note that by Lemma 1,

$$\mathbb{P}\left(\left(\cap_{\tau=t_f(\delta)}^{T_s} \mathcal{H}_{t_0,\tau}(\delta)\right)^c\right) \leq \frac{5T_s\delta}{T}.$$

Meanwhile, we note that

$$\cap_{t=t_0}^{T_s-1} \mathcal{B}_t(\delta; t_f(\delta)) \subseteq \cap_{\tau=t_f(\delta)}^{T_s} \left(\left(\bigcap_{k=0}^{K-2} \mathcal{B}_{t_0+kt_f(\delta)}(\delta; t_f(\delta))\right) \cap \mathcal{B}_{t_0+(K-1)t_f(\delta)}(\delta; \tau-(K-1)t_f(\delta))\right).$$

Hence, by Lemma 14, we have that

$$\begin{aligned}
&\mathbb{P}\left(\left(\cap_{\tau=t_f(\delta)}^{T_s} \left(\left(\bigcap_{k=0}^{K-2} \mathcal{B}_{t_0+kt_f(\delta)}(\delta; t_f(\delta))\right) \cap \mathcal{B}_{t_0+(K-1)t_f(\delta)}(\delta; \tau-(K-1)t_f(\delta))\right)\right)^c\right) \\
&\leq \mathbb{P}\left(\left(\cap_{t=t_0}^{T_s-1} \mathcal{B}_t(\delta; t_f(\delta))\right)^c\right) \leq \frac{T_s\delta}{T}.
\end{aligned}$$

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

$$\mathbb{P}\left(\left(\bigcap_{t=t_0}^{T_s-1} \mathcal{A}_t(\delta) \cap \mathcal{G}_t(\delta)\right)^c\right) \leq \frac{T_s\delta}{T} + \frac{2T_s\delta}{T} = \frac{3T_s\delta}{T}.$$

Hence, it follows that

$$\mathbb{P}\left(\left(\cap_{\tau=t_f(\delta)}^{T_s} \mathcal{E}_{t_0,\tau}(\delta) \cap \left(\cap_{t=t_0}^{T_s-1} \mathcal{A}_t(\delta) \cap \mathcal{G}_t(\delta)\right)\right)^c\right) \leq \frac{5T_s\delta}{T} + \frac{T_s\delta}{T} + \frac{3T_s\delta}{T} = \frac{9T_s\delta}{T}.$$

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

$$\mathbb{P}\left(\left(\cap_{\tau=1}^{T_s} \left(\mathcal{L}_{t_0,\tau}(\delta; F) \cap \mathcal{N}_{t_0,\tau}(\delta)\right) \cap \left(\cap_{t=t_0}^{T_s-1} \mathcal{O}_t(\delta)\right)\right)^c\right) \leq \frac{3T_s\delta}{T}.$$

Hence, it follows that $\mathbb{P}(\mathcal{P}_{t_0,T_s}(\delta, F)) \geq 1 - \frac{12T_s\delta}{T}$.

□

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 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 ϕ 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 ϕ 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 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, 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 $j \neq 2$. 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*

$$x_{t+1}' = x_t' - \eta\left(\frac{1}{m}\sum_{i=1}^{m}\left(Z_{t,i}Z_{t,i}^\top \nabla f(x_t') + \frac{u}{2}Z_{t,i}Z_{t,i}^\top \tilde{H}_{t,i}' Z_{t,i}\right) + Y_t'\right), \quad x_0' = x_0,$$

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

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

where

$$\xi_{g_0}(t) = \frac{1}{m}\sum_{i=1}^{m}(Z_{t,i}Z_{t,i}^\top - I)\nabla f(x_t),\ \xi_{g_0}'(t) = \frac{1}{m}\sum_{i=1}^{m}(Z_{t,i}(Z_{t,i})^\top - I)\nabla f(x_t'),\ \hat{\xi}_{g_0}(t) = \xi_{g_0}(t) - \xi_{g_0}'(t),$$

$$\xi_u(t) = \frac{1}{m}\sum_{i=1}^{m}\frac{u}{2}Z_{t,i}Z_{t,i}\tilde{H}_{t,i}Z_{t,i}, \quad \xi_u'(t) = \frac{1}{m}\sum_{i=1}^{m}\frac{u}{2}Z_{t,i}Z_{t,i}\tilde{H}_{t,i}'Z_{t,i}, \quad \hat{\xi}_u(t) = \xi_u(t) - \xi_u'(t),$$

$$\hat{Y}_t = Y_t - Y_t', \quad \bar{H}_t = \int_0^1 \nabla^2 f(a x_t + (1-a)x_t')da.$$

Proof. Observe that

$$\begin{aligned}
&\hat{x}_{t+1} := x_{t+1} - x_{t+1}' \\
&= x_t - \eta\left(\nabla f(x_t) + \xi_{g_0}(t) + \xi_u(t)Y_t\right) - \left[x_t' - \eta\left(\nabla f(x_t') + \xi_{g_0}'(t) + \xi_u'(t) + Y_t'\right)\right] \\
&= \hat{x}_t - \eta\left[(\nabla f(x_t) - \nabla f(x_t')) + \left(\xi_{g_0}(t) - \xi_{g_0}'(t)\right) + (\xi_u(t) - \xi_u'(t)) + (Y_t - Y_t')\right] \\
&= \hat{x}_t - \eta H\hat{x}_t - \eta(\bar{H}_t - H)\hat{x}_t - \eta\hat{\xi}_{g_0}(t) - \eta\hat{\xi}_u(t) - \eta\hat{Y}_t
\end{aligned}$$

$$= \underbrace{-\eta \sum_{\tau=0}^{t} (I - \eta H)^{t-\tau} \hat{\xi}_{g_0}(\tau)}_{W_{g_0}(t+1)} \underbrace{- \eta \sum_{\tau=0}^{t} (I - \eta H)^{t-\tau} (\bar{H}_\tau - H) \hat{x}_\tau}_{W_H(t+1)} \underbrace{- \eta \sum_{\tau=0}^{t} (I - \eta H)^{t-\tau} \hat{\xi}_u(\tau)}_{W_u(t+1)} \underbrace{- \eta \sum_{\tau=0}^{t} (I - \eta H)^{t-\tau} \hat{Y}_\tau}_{W_p(t+1)}$$

where

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

$$\xi_u(t) = \frac{1}{m} \sum_{i=1}^{m} \frac{u}{2} Z_{t,i} Z_{t,i} \tilde{H}_{t,i} Z_{t,i}, \quad \xi'_u(t) = \frac{1}{m} \sum_{i=1}^{m} \frac{u}{2} Z_{t,i} Z_{t,i} \tilde{H}'_{t,i} Z_{t,i}, \quad \hat{\xi}_u(t) = \xi_u(t) - \xi'_u(t),$$

$$\hat{Y}_t = Y_t - Y'_t, \quad \bar{H}_t = \int_0^1 \nabla^2 f(a x_t + (1-a) x'_t) da.$$

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

Suppose x_0 is an ϵ-saddle point. Recall that $\gamma > 0$ denotes $-\lambda_{\min}(\nabla^2 f(x_0))$, where we know that $\gamma \geq \sqrt{\rho\epsilon}$. In the sequel, for any $t \geq 0$, it is helpful to define the quantities

$$\beta(t)^2 := \frac{(1+\eta\gamma)^{2t}}{(\eta\gamma)^2 + 2\eta\gamma}, \quad \alpha(t)^2 := \frac{(1+\eta\gamma)^{2t} - 1}{(\eta\gamma)^2 + 2\eta\gamma}. \tag{44}$$

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 $\delta \in (0, 1/e]$ throughout the lemma. Suppose that we pick u, r and η as specified in Lemma 23. Suppose $T_s \geq t_f(\delta)$. Suppose also that*

$$f(x_{T_s}) - f(x_0) > -F, \quad f(x'_{T_s}) - f(x_0) > -F.$$

Then, we have the following results.

1. *Let $\mathcal{S}_\phi(\delta)$ denote the event*

$$\mathcal{S}_\phi(\delta) := \left\{ \max\{\|x_t - x_0\|^2, \|x'_t - x_0\|^2\} \leq \phi_{T_s}(\delta, F), \quad \forall 0 \leq t \leq T_s \right\}.$$

In addition, let $\mathcal{S}_u(\delta)$ denote the event

$$\mathcal{S}_u(\delta) := \left\{ \|W_u(t+1)\| \leq \eta \beta(t+1) \frac{\sqrt{3}}{\sqrt{\eta\sqrt{\rho\epsilon}}} \left(2 c_3 \rho d^2 (\log(T/\delta))^2\right) u^2, \quad \forall 0 \leq t \leq T_s - 1 \right\},$$

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

$$\mathbb{P}(\mathcal{S}_\phi(\delta) \cap \mathcal{S}_u(\delta)) \geq 1 - \frac{24 T_s \delta}{T}.$$

2. *Consider defining the event $\mathcal{R}_t(\delta)$, which is the event where*

either $\displaystyle\sum_{\tau=0}^{t} (1+\eta\gamma)^{2(t-\tau)} \frac{dL^2}{m} \|x_\tau - x'_\tau\|^2 (\mathrm{lr}(CT^2/\delta))^2 \geq G_{T_s}(\delta, F)$, ***or***

$$\|W_{g_0}(t+1)\| \leq c' \eta \sqrt{\max\left\{ \left(\mathrm{lr}\left(\frac{CT^2}{\delta}\right)\right)^2 \sum_{\tau=0}^{t} \frac{dL^2}{m} (1+\eta\gamma)^{2(t-\tau)} \|x_\tau - x'_\tau\|^2, g(t+1) \right\} \left(\log\left(\frac{CdT^2}{\delta}\right) + \log\left(\log\left(\frac{G_{T_s}(\delta, F)}{g(t+1)}\right)\right) + 1 \right)}$$

normalsize holds. Above, c', C refer to the same constants as in Proposition 2, and

$$G_{T_s}(\delta, F) := 8 \sum_{\tau=0}^{T_s - 1} (1+\eta\gamma)^{2\tau} \frac{dL^2}{m} (\mathrm{lr}(CT^2/\delta))^2 \phi_{T_s}(\delta, F) + \left(\frac{\beta(T_s)\eta r}{60\sqrt{d}}\right)^2, \qquad g(t+1) := \left(\frac{\beta(t+1)\eta r}{60\sqrt{d}}\right)^2.$$

Then, $\mathbb{P}(\mathcal{R}_t(\delta)) \geq 1 - \frac{\delta}{T}$. *Suppose the event*

$$\left(\cap_{t=0}^{T_s-1}\mathcal{R}_t(\delta)\right) \cap \mathcal{S}_\phi(\delta)$$

holds. Then, the event $\mathcal{S}_{g_0}(\delta)$ *holds, where*

$$\mathcal{S}_{g_0}(\delta) := \cap_{t=0}^{T_s-1}\mathcal{S}_{g_0,t}(\delta),$$

and $\mathcal{S}_{g_0,t}(\delta)$ *is defined as*

$$\mathcal{S}_{g_0,t}(\delta) := \left\{\|W_{g_0}(t+1)\| \leq \zeta_1(\delta,F)c'\eta\sqrt{\max\left\{\left(\mathrm{lr}\left(\frac{CT^2}{\delta}\right)\right)^2\sum_{\tau=0}^{t}\frac{dL^2}{m}(1+\eta\gamma)^{2(t-\tau)}\|x_\tau - x'_\tau\|^2, g(t+1)\right\}}\right\}$$

where

$$\zeta_1(\delta,F) := \left(\log\left(\frac{CdT^2}{\delta}\right) + \log\left(\log\left(\frac{G_{T_s}(\delta,F)}{g(1)}\right)\right) + 1\right).$$

3. *In addition, let* $\mathcal{S}_p(\delta)$ *denote the event*

$$\mathcal{S}_p(\delta) := \left\{\|W_p(t+1)\| \leq \frac{2\sqrt{2\log(T/\delta)}\beta(t+1)\eta r}{\sqrt{d}} \quad \forall 0 \leq t \leq T_s - 1\right\}.$$

Then, $\mathbb{P}(\mathcal{S}_p(\delta)) \geq 1 - \frac{T_s\delta}{T}$.

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 $\mathcal{P}_{0,T_s}(\delta,F)$, we have that $\|x_\tau - x_0\|^2 \leq \phi_{T_s}(\delta,F)$. Simultaneously, on the event $\mathcal{P}_{0,T_s}(\delta,F)$, we know that $\cap_{t=0}^{T_s-1}\mathcal{G}_t(\delta)$ holds, i.e.

$$\frac{1}{m}\sum_{i=1}^{m}\|Z_{t,i}\|^4 \leq 2c_3d^2\left(\log(T/\delta)\right)^2, \quad \forall 0 \leq t \leq T_s - 1. \tag{45}$$

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

$$\begin{aligned}
\|W_u(t+1)\| &= \left\|\eta\sum_{\tau=0}^{t}(I-\eta H)^{t-\tau}\hat{\xi}_u(\tau)\right\| \\
&\leq \left\|\eta\sum_{\tau=0}^{t}(I-\eta H)^{t-\tau}\xi_u(\tau)\right\| + \left\|\eta\sum_{\tau=0}^{t}(I-\eta H)^{t-\tau}\hat{\xi}'_u(\tau)\right\| \\
&\leq \eta\sum_{\tau=0}^{t}(1+\eta\gamma)^{t-\tau}\left(\left\|\frac{1}{m}\sum_{i=1}^{m}\frac{u}{2}Z_{t,i}Z_{t,i}\tilde{H}_{t,i}Z_{t,i}\right\| + \left\|\frac{1}{m}\sum_{i=1}^{m}\frac{u}{2}Z_{t,i}Z_{t,i}\tilde{H}'_{t,i}Z_{t,i}\right\|\right) \\
&\leq \eta\sum_{\tau=0}^{t}(1+\eta\gamma)^{t-\tau}\frac{\rho}{m}\sum_{i=1}^{m}\|Z_{t,i}\|^4u^2 \\
&\overset{\text{(iv)}}{\leq} \eta\sum_{\tau=0}^{t}(1+\eta\gamma)^{t-\tau}\rho(2c_3)d^2(\log(T2/\delta))^2u^2 \\
&\leq \eta\frac{(1+\eta\gamma)^{t+1}}{\eta\gamma}\left(2c_3\rho Cd^2(\log(T/\delta))^2\right)u^2 \\
&\overset{\text{(v)}}{=} \eta\beta(t+1)\frac{\sqrt{(\eta\gamma)^2+2\eta\gamma}}{\eta\gamma}\left(2c_3\rho d^2(\log(T/\delta))^2\right)u^2 \\
&\leq \eta\beta(t+1)\frac{\sqrt{3}}{\sqrt{\eta\gamma}}\left(2c_3\rho d^2(\log(T/\delta))^2\right)u^2
\end{aligned}$$

$$\overset{\text{(vi)}}{\leq} \eta\beta(t+1)\frac{\sqrt{3}}{\sqrt{\eta\sqrt{\rho\epsilon}}}\left(2c_3\rho d^2(\log(T/\delta))^2\right)u^2$$

where the inequality in (iv) holds due to Eq. (45), the equality in (v) holds due to the definition of $\beta(t+1)$, and the inequality in (vi) used the fact that $\gamma \geq \sqrt{\rho\epsilon}$.

Hence the event

$$\cap_{t=0}^{T_s}\left\{\|x_t - x_0\|^2 \leq \phi_{T_s}(\delta, F) \text{ and } \right\} \cap \mathcal{S}_u(\delta)$$

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

Note that by the coupling, the distribution of x'_τ is the same as that of x_τ. Thus, by the assumption $f(x'_{T_s}) - f(x_0) > -F$, it follows by a similar argument that the bound $\|x'_\tau - x_0\|^2 \leq \phi_{T_s}(\delta, F)$ also holds with probability at least $1 - \frac{12T_s\delta}{T}$. The claim then follows by an application of the union bound.

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

$$\|x_\tau - x_0\|^2 \leq \phi_{T_s}(\delta, F), \qquad \|x'_\tau - x_0\|^2 \leq \phi_{T_s}(\delta, F)$$

where $\phi_{T_s}(\delta, F)$ is as defined in Lemma 23.

Suppose now that $\mathcal{R}_t(\delta)$ holds true, and suppose for contradiction that

$$\begin{aligned}
&\sum_{\tau=0}^{t}(1+\eta\gamma)^{2(t-\tau)}\frac{dL^2}{m}\|x_\tau - x'_\tau\|^2(\mathrm{lr}(CT^2/\delta))^2 \\
&\geq G_{T_s}(\delta, F) \\
&= 8\sum_{\tau=0}^{T_s-1}(1+\eta\gamma)^{2\tau}\frac{dL^2}{m}(\mathrm{lr}(CT^2/\delta))^2\phi_{T_s}(\delta, F) + \left(\frac{\beta(T_s)\eta r}{60\sqrt{d}}\right)^2.
\end{aligned}$$

This implies that there exists some $0 \leq \tau \leq t \leq T_s$ such that $\|x_\tau - x'_\tau\|^2 \geq 8\phi_{T_s}(\delta, F)$. However, we also know that on the event $\mathcal{S}_\phi(\delta)$,

$$\|x_\tau - x'_\tau\|^2 \leq 2\|x_\tau - x_0\|^2 + 2\|x'_\tau - x_0\|^2 \leq 4\phi_{T_s}(\delta, F).$$

This leads to a contradiction. We must then have that

$$\|W_{g_0}(t+1)\| \leq \zeta_1(\delta, F)c'\eta\sqrt{\max\left\{\left(\mathrm{lr}\left(\frac{CT^2}{\delta}\right)\right)^2\sum_{\tau=0}^{t}\frac{dL^2}{m}(1+\eta\gamma)^{2(t-\tau)}\|x_\tau - x'_\tau\|^2, g(t+1)\right\}}$$

where

$$\zeta_1(\delta, F) := \sqrt{\log\left(\frac{CdT^2}{\delta}\right) + \log\left(\log\left(\frac{G(\delta, F)}{g(1)}\right) + 1\right)}$$

3. Observe that

$$W_p(t+1) = \eta\sum_{\tau=0}^{t}(I - \eta H)^{t-\tau}\hat{Y}_\tau = \eta\sum_{\tau=0}^{t}(1+\eta\gamma)^{t-\tau}(2(Y_\tau)_1),$$

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

$$\eta^2\sum_{\tau=0}^{t}(1+\eta\gamma)^{2(t-\tau)}\frac{4r^2}{d} = \frac{4\eta^2r^2}{d}\frac{(1+\eta\gamma)^{2(t+1)} - 1}{2\eta\gamma + (\eta\gamma)^2} = \frac{4\eta^2r^2\alpha(t+1)^2}{d}. \tag{46}$$

Since $\alpha(t+1) \leq \beta(t+1)$, using the subGaussianity of a Gaussian distribution, it follows that for any t, with probability at least $1 - \delta/T$,

$$\|W_p(t+1)\| \leq \frac{2\sqrt{2\log(T/\delta)}\beta(t+1)\eta r}{\sqrt{d}}.$$

□

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

Proposition 6. *Suppose that x_{t_0} is an ϵ-approximate saddle point. Let $c' > 0, c_1 > 0, c_2 \geq 1, c_4 > 0, c_5 > 0, c_6 > 0, c_7 > 0, C_1 \geq 1$ be the absolute constants defined in the statements of the previous lemmas, and let $\delta \in (0, 1/e]$ be arbitrary. Consider any $F > 0$. As in the statement of Lemma 23, suppose we choose u, r and η such that*

$$u \leq \frac{\sqrt{\epsilon}}{d\sqrt{\rho}\log(T/\delta)} \cdot \min\left\{\frac{1}{64c_5^2c_2}, \frac{1}{2048c_1c_2}\right\}^{1/4}, \qquad r \leq \epsilon \cdot \min\left\{\frac{1}{8c_5\sqrt{2c_2}}, \frac{1}{32\sqrt{c_1}}\right\},$$
$$\eta \leq \frac{1}{Lt_f(\delta)} \min\left\{\frac{1}{\log(T/\delta)}, \frac{\sqrt{m}}{8c_4(\mathrm{lr}(C_1dmT/\delta))^{3/2}\sqrt{d}}, \frac{m}{128c_1(\mathrm{lr}(C_1dmT/\delta))^3 d}\right\}.$$

Suppose we pick

$$T_s = \max\left\{\lceil\frac{\iota}{\eta\sqrt{\rho\epsilon}}\rceil, t_f(\delta), 4\right\}, \tag{47}$$

where

$$\iota = \max\left\{\log\left(2\sqrt{\phi_{T_s}(\delta, F)}\frac{20\sqrt{d}\sqrt{\eta^2\gamma^2 + 2\eta\gamma}}{\eta r}\right), 1\right\},$$

Suppose in addition that u, η also satisfy the conditions

$$u \leq \sqrt{\frac{r\sqrt{\eta\sqrt{\rho\epsilon}}}{120\sqrt{3}c_3\sqrt{d}\rho d^2(\log(T/\delta))^2}},$$
$$\eta \leq \max\left\{\frac{1}{c'c_9\zeta_1(\delta, F)}, \frac{m\sqrt{\rho\epsilon}}{360\iota(c')^2c_9^2dL^2\left(\mathrm{lr}\left(\frac{CT^2}{\delta}\right)\right)^2\zeta_1(\delta, F)^2}, \frac{1}{2\sqrt{\rho\epsilon}}\right\},$$

where $\zeta_1(\delta, F)$ is as defined in Lemma 23, $c', c_3, C > 0$ are the same constants as in the previous results, and $c_9 = 2\sqrt{2} + \frac{1}{20}$. Suppose also that $\phi_{T_s}(\delta, F)$ satisfies the bound

$$\phi_{T_s}(\delta, F) \leq \left(\frac{\sqrt{\rho\epsilon}}{60c_9\iota\rho\log(T/\delta)}\right)^2. \tag{48}$$

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

Proof of Proposition 6. Without loss of generality, we assume that $t_0 = 0$. By Lemma 3, we have

$$\begin{aligned}
&\hat{x}_{t+1} \\
&:= x_{t+1} - x'_{t+1} \\
&= -\underbrace{\eta\sum_{\tau=t_0}^{t}(I - \eta H)^{t-\tau}\hat{\xi}_{g_0}(\tau)}_{W_{g_0}(t+1)} - \underbrace{\eta\sum_{\tau=t_0}^{t}(I - \eta H)^{t-\tau}(\bar{H}_\tau - H)\hat{x}_\tau}_{W_H(t+1)} - \underbrace{\eta\sum_{\tau=t_0}^{t}(I - \eta H)^{t-\tau}\hat{\xi}_u(\tau)}_{W_u(t+1)} - \underbrace{\eta\sum_{\tau=t_0}^{t}(I - \eta H)^{t-\tau}\hat{Y}_\tau}_{W_p(t+1)}
\end{aligned}$$

where

$$\xi_{g_0}(t) = \frac{1}{m}\sum_{i=1}^{m}(Z_{t,i}Z_{t,i}^\top - I)\nabla f(x_t), \quad \xi'_{g_0}(t) = \frac{1}{m}\sum_{i=1}^{m}(Z_{t,i}(Z_{t,i})^\top - I)\nabla f(x'_t), \quad \hat{\xi}_{g_0}(t) = \xi_{g_0}(t) - \xi'_{g_0}(t),$$
$$\xi_u(t) = \frac{1}{m}\sum_{i=1}^{m}\frac{u}{2}Z_{t,i}Z_{t,i}\tilde{H}_{t,i}Z_{t,i}, \quad \xi'_u(t) = \frac{1}{m}\sum_{i=1}^{m}\frac{u}{2}Z_{t,i}Z_{t,i}\tilde{H}'_{t,i}Z_{t,i}, \quad \hat{\xi}_u(t) = \xi_u(t) - \xi'_u(t),$$
$$\hat{Y}_t = Y_t - Y'_t, \quad \bar{H}_t = \int_0^1 \nabla^2 f(ax_t + (1-a)x'_t)da.$$

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

$$\beta(t)^2 := \frac{(1+\eta\gamma)^{2t}}{(\eta\gamma)^2 + 2\eta\gamma}, \quad \alpha(t)^2 := \frac{(1+\eta\gamma)^{2t} - 1}{(\eta\gamma)^2 + 2\eta\gamma}.$$

Throughout the proof, we suppose for contradiction that

$$f(x_{T_s}) - f(x_0) > -F, \quad f(x'_{T_s}) - f(x_0) > -F,$$

and assume the event

$$\left(\cap_{t=0}^{T_s-1} \mathcal{R}_t(\delta)\right) \cap \mathcal{S}_\phi(\delta) \cap \mathcal{S}_u(\delta) \cap \mathcal{S}_p(\delta)$$

holds, where the events intersected are defined in Lemma 24. Then, by Lemma 24, the event $\mathcal{S}_{g_0}(\delta)$ (also defined in Lemma 24) holds[7].

Consider the following induction argument, where we seek to show that there exists an absolute constant $c_9 > 0$ such that for every $t \in \{0, 1, \ldots, T_s\}$,

$$\|x_t - x'_t\| \leq c_9 \log(T/\delta) \frac{\beta(t)\eta r}{\sqrt{d}}, \text{ and } \max\left\{\|W_{g_0}(t)\|, \|W_H(t)\|, \|W_u(t)\|\right\} \leq \frac{\beta(t+1)\eta r}{\sqrt{d}} \tag{49}$$

Combined with a lower bound on $\|W_p(t+1)\|$ (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. (49) to show that

$$\|W_p(T_s)\| \geq 2\left(\left\|W_{g_0(T_s)}\right\| + \|W_H(T_s)\| + \|W_u(T_s)\|\right).$$

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 $\left\|x_{T_s} - x'_{T_s}\right\|$ is larger than what expect (since our assumptions imply that $\max\left\{\|x_{T_s} - x_0\|^2, \left\|x'_{T_s} - x_0\right\|^2\right\} \leq \phi_{T_s}(\delta, F)$, i.e. x_{T_s} and x'_{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.

$$\left(\cap_{t=0}^{T_s-1} \mathcal{R}_t(\delta)\right) \cap \mathcal{S}_\phi(\delta) \cap \mathcal{S}_p(\delta)$$

the assumption

$$f(x_{T_s}) - f(x_0) > -F, \quad \text{and } f(x'_{T_s}) - f(x_0) > -F$$

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

We proceed to prove Eq. (49). Observe that the claim holds for the base case $t = 0$; this is true since $x_0 = x'_0$. Now suppose that this holds for all $\tau \leq t$. We will seek to show that Eq. (49) holds for $t+1$ 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.

1. (Bounding $\|W_{g_0}(t+1)\|$) Since the event $\mathcal{S}_{g_0}(\delta)$ holds, it follows that for each $0 \leq t \leq T_s - 1$, we have that

$$\|W_{g_0}(t+1)\| \leq \zeta_1(\delta, F) c' \eta \sqrt{\max\left\{\left(\text{lr}\left(\frac{CT^2}{\delta}\right)\right)^2 \sum_{\tau=0}^{t} \frac{dL^2}{m}(1+\eta\gamma)^{2(t-\tau)}\|x_\tau - x'_\tau\|^2, g(t+1)\right\}}$$

where

$$\zeta_1(\delta, F) := \left(\log\left(\frac{CdT^2}{\delta}\right) + \log\left(\log\left(\frac{G_{T_s}(\delta, F)}{g(1)}\right)\right) + 1\right),$$

and the terms $G_{T_s}(\delta, F)$ and $g(1)$ are defined as in Lemma 24. Recall by the inductive claim in Eq. (49) that there exists $c_9 > 0$ such that

$$\|x_\tau - x'_\tau\| \leq c_9 \log(T/\delta) \frac{\beta(t)\eta r}{\sqrt{d}} \quad \forall\, 0 \leq \tau \leq t.$$

[7] We may also directly assume that $\mathcal{S}_{g_0}(\delta)$ also holds, but our way of reasoning prevents double counting of probabilities.

Hence, it follows that

$$\|W_{g_0}(t+1)\| \leq c'\zeta_1(\delta,F)\eta \max\left\{\sqrt{t+1}\left(\mathrm{lr}\left(\frac{CT^2}{\delta}\right)\right)\frac{c_9\sqrt{d}L}{\sqrt{m}}\frac{\beta(t)\eta r}{\sqrt{d}}, \frac{\beta(t+1)\eta r}{60\sqrt{d}}\right\}.$$

Hence, noting the choice of T_s in Eq. (47), by choosing η such that

$$c'c_9\zeta_1(\delta,F)\eta\sqrt{T_s}\left(\mathrm{lr}\left(\frac{CT^2}{\delta}\right)\right)\frac{\sqrt{d}L}{\sqrt{m}} \leq \frac{1}{60} \iff \eta \leq \frac{m\sqrt{\rho\epsilon}}{360\iota(c')^2c_9^2dL^2\left(\mathrm{lr}\left(\frac{CT^2}{\delta}\right)\right)^2\zeta_1(\delta,F)^2}, \text{ and}$$

$c'c_9\zeta_1(\delta,F)\eta \leq 1$.

it follows that

$$\|W_{g_0}(t+1)\| \leq \frac{\beta(t+1)\eta r}{60\sqrt{d}}.$$

2. Meanwhile, the term $W_H(t+1)$ can be bounded as follows. By the inductive assumption in Eq. (49), we have that

$$\|\hat{x}_\tau\| = \|x_\tau - x'_\tau\| \leq c_9\log(T/\delta)\frac{\beta(\tau)\eta r}{\sqrt{d}} \quad \forall\, 0 \leq \tau \leq t.$$

Moreover, on the event our proof assumes, we know that

$$\max\left\{\|x_\tau - x_0\|^2, \|x'_\tau - x_0\|^2\right\} \leq \phi_{T_s}(\delta,F).$$

Thus, using the ρ-Hessian Lipschitz property, we have

$$\begin{aligned}
\|W_H(t+1)\| &= \eta\left\|\sum_{\tau=0}^{t}(I-\eta H)^{t-\tau}(\bar{H}_\tau - H)\hat{x}_\tau\right\| \\
&\leq \eta\sum_{\tau=0}^{t}(1+\eta\gamma)^{t-\tau}\rho\sqrt{\phi_{T_s}(\delta,F)}\frac{c_9\log(T/\delta)\beta(\tau)\eta r}{\sqrt{d}} \\
&\leq c_9(t+1)\log(T/\delta)\eta\rho\sqrt{\phi_{T_s}(\delta,F)}\frac{\beta(t)\eta r}{\sqrt{d}} \\
&\leq c_9T_s\log(T/\delta)\eta\rho\sqrt{\phi_{T_s}(\delta,F)}\frac{\beta(t)\eta r}{\sqrt{d}}.
\end{aligned}$$

Given our choice of T_s in Eq. (47), if

$$c_9T_s\log(T/\delta)\eta\rho\sqrt{\phi_{T_s}(\delta,F)} \leq \frac{1}{60} \iff \phi_{T_s}(\delta,F) \leq \left(\frac{\sqrt{\rho\epsilon}}{60c_9\iota\rho\log(T/\delta)}\right)^2$$

it follows that

$$\|W_H(t+1)\| \leq \frac{\beta(t+1)\eta r}{60\sqrt{d}}.$$

3. Meanwhile, for $W_u(t+1)$, since the event $\mathcal{S}_u(\delta)$ holds, we have that

$$\|W_u(t+1)\| \leq \eta\beta(t+1)\frac{\sqrt{3}}{\sqrt{\eta\sqrt{\rho\epsilon}}}\left(2c_3\rho d^2(\log(T/\delta))^2\right)u^2.$$

Now, by picking

$$\eta\beta(t+1)\frac{\sqrt{3}}{\sqrt{\eta\sqrt{\rho\epsilon}}}\left(2c_3\rho d^2(\log(T/\delta))^2\right)u^2 \leq \frac{\beta(t+1)\eta r}{60\sqrt{d}} \iff u \leq \sqrt{\frac{r\sqrt{\eta\sqrt{\rho\epsilon}}}{120\sqrt{3}c_3\sqrt{d}\rho d^2(\log(T/\delta))^2}},$$

it follows that with probability $1-\delta/T$, $\|W_u(t+1)\| \leq \frac{\beta(t+1)\eta r}{60\sqrt{d}}$.

4. Meanwhile, observe that since $\mathcal{S}_p(\delta)$ holds, it follows that

$$W_p(t+1) \leq \frac{2\sqrt{2\log(T/\delta)}\beta(t+1)\eta r}{\sqrt{d}}.$$

Combining the bounds for W_{g_0}, W_p, W_H and W_u, it follows that

$$\begin{aligned}\|\hat{x}_{t+1}\| &\leq \|W_{g_0}(t+1)\| + \|W_p(t+1)\| + \|W_H(t+1)\| + \|W_u(t+1)\| \\ &\leq \frac{\beta(t+1)\eta r}{\sqrt{d}}\left(\frac{1}{60} + \frac{1}{60} + \frac{1}{60} + 2\sqrt{2\log(T/\delta)}\right) \\ &\leq \frac{\beta(t+1)\eta r}{\sqrt{d}}\left(\frac{1}{20} + 2\sqrt{2}\right)\log(T/\delta),\end{aligned}$$

where the final inequality uses the fact that $0 < \delta \leq 1/e$ (which implies $\log(T/\delta) \geq 1$). Hence, we see that the first part of the inductive claim of Eq. (49) holds with the constant $c_9 := \frac{1}{20} + 2\sqrt{2}$, and the second part follows naturally as a consequence of our argument above.

Meanwhile, observe that for any η such that $\eta\sqrt{\rho\epsilon} \leq \frac{1}{2}$, we have that $(1+\eta\gamma)^{\frac{1}{\eta\sqrt{\rho\epsilon}}} \geq 2$. Thus, by choosing η such that $\eta\sqrt{\rho\epsilon} \leq \frac{1}{2}$, we have that for any $t \geq \frac{1}{\eta\sqrt{\rho\epsilon}}$,

$$\alpha(t+1)^2 \geq \frac{1}{2}\beta(t+1)^2.$$

Hence, following Eq. (46), by choosing $T_s \geq \frac{1}{\eta\sqrt{\rho\epsilon}}$, $W_p(T_s)$ is a 1-dimensional Gaussian with variance at least $\frac{2\eta^2 r^2 \beta(T_s)}{d}$, such that with probability at least 2/3,

$$\|W_p(T_s)\| \geq \frac{\beta(T_s)\eta r}{10\sqrt{d}}.$$

Simultaneously, we know that on the event

$$\left(\cap_{t=0}^{T_s-1}\mathcal{R}_t(\delta)\right) \cap \mathcal{S}_\phi(\delta) \cap \mathcal{S}_u(\delta) \cap \mathcal{S}_p(\delta),$$

we have

$$\|W_{g_0}(T_s)\| + \|W_H(T_s)\| + \|W_u(T_s)\| \leq \frac{3\beta(T_s)\eta r}{60\sqrt{d}} = \frac{\beta(T_s)\eta r}{20\sqrt{d}}.$$

We note that by Lemma 24, we have

$$\mathbb{P}\left(\left(\cap_{t=0}^{T_s-1}\mathcal{R}_t(\delta)\right) \cap \mathcal{S}_\phi(\delta) \cap \mathcal{S}_u(\delta) \cap \mathcal{S}_p(\delta)\right) \geq 1 - \left(\frac{24T_s\delta}{T} + \frac{T_s\delta}{T} + \frac{T_s\delta}{T}\right) = 1 - \frac{26T_s\delta}{T}.$$

Thus, with probability at least $2/3 - \frac{26T_s\delta}{T}$, we have

$$\|\hat{x}_{T_s}\| \geq \frac{1}{2}\|W_p(T_s)\| \geq \frac{\beta(T_s)\eta r}{20\sqrt{d}}$$

Thus, choosing $T_s \geq \frac{\iota}{\eta\sqrt{\rho\epsilon}}$, where

$$\iota = \max\left\{\log\left(2\sqrt{\phi_{T_s}(\delta, F)}\frac{20\sqrt{d}\sqrt{\eta^2\gamma^2 + 2\eta\gamma}}{\eta r}\right), 1\right\},$$

noting that if $\eta\sqrt{\rho\epsilon} \leq 1/2$, then $(1+\eta\gamma)^{\frac{1}{\eta\sqrt{\rho\epsilon}}} \geq (1+\eta\sqrt{\rho\epsilon})^{\frac{1}{\eta\sqrt{\rho\epsilon}}} \geq 2$, we have that with probability at least $2/3 - \frac{26T_s\delta}{T}$,

$$\|\hat{x}_{T_s}\| \geq \frac{\beta(T_s)\eta r}{20\sqrt{d}} = \frac{\eta r}{20\sqrt{d}}\frac{(1+\eta\gamma)^{T_s}}{\sqrt{2\eta\gamma + (\eta\gamma)^2}}$$

$$\geq \frac{\eta r}{20\sqrt{d}} \frac{(1+\eta\gamma)^{\frac{\log\left(2\sqrt{\phi_{T_s}(\delta,F)}\frac{20\sqrt{d}\sqrt{\eta^2\gamma^2+2\eta\gamma}}{\eta r}\right)}{\eta\sqrt{\rho\epsilon}}}}{\sqrt{2\eta\gamma+(\eta\gamma)^2}}$$

$$\geq \frac{\eta r}{20\sqrt{d}\sqrt{2\eta\gamma+(\eta\gamma)^2}} 2^{\log\left(2\sqrt{\phi_{T_s}(\delta,F)}\frac{20\sqrt{d}\sqrt{\eta^2\gamma^2+2\eta\gamma}}{\eta r}\right)} > 2\sqrt{\phi_{T_s}(\delta,F)} > 2\sqrt{\phi(T_s,\delta)}.$$

Thus, at least one of $\|x_{T_s} - x_0\|$ and $\left\|x'_{T_s} - x_0\right\|$ is larger than $\sqrt{\phi(T_s,\delta)}$, a contradiction. Since the two sequences have the same distribution, it follows that with probability at least $1/3 - \frac{13T_s\delta}{T}$, $f(x_{T_s}) - f(x_0) \leq -F$. □

In the result above, we require an upper bound on the norm of $\phi_{T_s}(\delta, F)$ to hold (i.e. equation 48), 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 48, assuming that u, r and η are chosen appropriately small such that the dominant term of $\|\phi_{T_s}(\delta, F)\|$ scales with F.

Lemma 25. *Consider choosing F such that*

$$F = \frac{1}{2}\left(\frac{\sqrt{\rho\epsilon}}{60c_9\iota\rho\log(T/\delta)}\right)^2 \frac{1}{\eta T_s t_f(\delta)\left(129 + 8c'^2\beta_1(\delta;F)\left(16(\mathrm{lr}(CT^2/\delta))^2+1\right)\right)}.$$

Suppose $\eta \leq \min\left\{1, \frac{1}{t_f(\delta)}, \frac{1}{t_f\delta L}\right\}$. Suppose we pick u and r small enough such that

$$u \leq \frac{r^{1/2}}{d\log(T/\delta)\rho^{1/2}}, \quad r^2 \leq \min\left\{\frac{F\sqrt{\rho\epsilon}}{2\iota\log(T/\delta)\left(\frac{65c_5^2}{8}+6c_1+1\right)}, \frac{F}{4c_6\log(2dT/\delta)+\frac{8c_7\iota}{\sqrt{\rho\epsilon}}}\right\}.$$

Then, $N_{u,r}(T_s,\delta) \leq F$, and that

$$4c_6\eta^2 T_s\log(2dT/\delta)r^2 + 4c_7\eta^2T_s^2\rho^2u^4d^4\left(\log(T/\delta)\right)^4 \leq \eta T_s t_f(\delta)F.$$

Suppose in addition η is small enough so that

$$32\eta^2(t_f(\delta))^2\epsilon^2 \leq \frac{1}{2}\left(\frac{\sqrt{\rho\epsilon}}{60c_9\iota\rho\log(T/\delta)}\right)^2.$$

Suppose also that $\sqrt{\rho\epsilon} \leq 1$ and $\eta \leq \frac{m}{d}$, so that $T_s \geq \frac{\iota}{\eta\sqrt{\rho\epsilon}} \geq \frac{d}{m}$. Then, the condition in Eq. (48) *will be satisfied.*

Proof. We note that since $\frac{\iota}{\sqrt{\rho\epsilon}} \leq T_s \leq \frac{2\iota}{\sqrt{\rho\epsilon}}$, 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. (48) requires that

$$\phi_{T_s}(\delta,F) \leq \left(\frac{\sqrt{\rho\epsilon}}{60c_9\iota\rho\log(T/\delta)}\right)^2.$$

By our choice of η such that

$$32\eta^2(t_f(\delta))^2\epsilon^2 \leq \frac{1}{2}\left(\frac{\sqrt{\rho\epsilon}}{60c_9\iota\rho\log(T/\delta)}\right)^2,$$

it suffices for us to show that

$$\frac{1}{2}\left(\frac{\sqrt{\rho\epsilon}}{60c_9\iota\rho\log(T/\delta)}\right)^2 \geq 128\eta T_s t_f(\delta)F + 8c'^2\beta_1(\delta;F)\eta t_f(\delta)\max\left\{\frac{16d}{m}(\mathrm{lr}(CT^2/\delta))^2F, T_sF\right\} + \eta T_s t_f(\delta)F$$
$$= 129\eta T_s t_f(\delta)F + 8c'^2\beta_1(\delta;F)\eta t_f(\delta)\max\left\{\frac{16d}{m}(\mathrm{lr}(CT^2/\delta))^2F, T_sF\right\}.$$

By our assumption, we know that $T_s \geq \frac{d}{m}$. Thus, further simplifying indicates that it suffices for us to show

$$\frac{1}{2}\left(\frac{\sqrt{\rho\epsilon}}{60c_9\iota\rho\log(T/\delta)}\right)^2 \geq 129\eta T_s t_f(\delta)F + 8c'^2\beta_1(\delta;F)\eta t_f(\delta)\max\left\{16T_s(\mathrm{lr}(CT^2/\delta))^2F, T_sF\right\}. \tag{50}$$

By choosing F such that

$$F \leq \frac{1}{2}\left(\frac{\sqrt{\rho\epsilon}}{60c_9\iota\rho\log(T/\delta)}\right)^2 \frac{1}{\eta T_s t_f(\delta)\left(129 + 8c'^2\beta_1(\delta;F)\left(16(\mathrm{lr}(CT^2/\delta))^2+1\right)\right)},$$

we see that Eq. (50) is satisfied.

□

Remark 1. Suppose without loss of generality that $T_s = \frac{\iota}{\eta\sqrt{\rho\epsilon}}$. 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

$$\frac{F}{T_s} = \frac{1}{2}\left(\frac{\sqrt{\rho\epsilon}}{60c_9\iota\rho\log(T/\delta)}\right)^2 \frac{1}{\eta T_s^2 t_f(\delta)\left(129 + 8c'^2\beta_1(\delta;F)\left(16(\mathrm{lr}(CT^2/\delta))^2+1\right)\right)}$$
$$= \eta\frac{\sqrt{\rho\epsilon}^4}{\rho^2}\left(\frac{1}{2\iota^2}\frac{1}{(60c_9\iota\log(T/\delta))^2\left(t_f(\delta)\right)\left(129 + 8c'^2\beta_1(\delta;F)\left(16(\mathrm{lr}(CT^2/\delta))^2+1\right)\right)}\right)$$
$$= \eta\epsilon^2\left(\frac{1}{2\iota^2}\frac{1}{(60c_9\iota\log(T/\delta))^2\left(t_f(\delta)\right)\left(129 + 8c'^2\beta_1(\delta;F)\left(16(\mathrm{lr}(CT^2/\delta))^2+1\right)\right)}\right)$$

Before we proceed to prove our main result, we need an additional result 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 $\tau \geq t_f(\delta)$.

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

$$\eta \leq \frac{1}{Lt_f(\delta)} \cdot \min\left\{\frac{\sqrt{m}}{8c_4(\mathrm{lr}(C_1dmT/\delta))^{3/2}\sqrt{d}}, \frac{m}{128c_1(\mathrm{lr}(C_1dmT/\delta))^3 d}\right\}. \tag{51}$$

Suppose also we pick u, r and η as prescribed in the statement of Proposition 5.

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:

$$f(x_\tau) - f(x_0) \leq \frac{\eta}{4}\epsilon^2 + t_f(\delta)\eta u^4\rho^2 \cdot c_1 d^3\left(\log\frac{T}{\delta}\right)^3 + t_f(\delta)L\eta^2u^4\rho^2 \cdot c_1 d^4\left(\log\frac{T}{\delta}\right)^4$$
$$+ \eta c_1 r^2(128t_f(\delta) + \eta L)\log\frac{T}{\delta} + t_f(\delta)c_1L\eta^2r^2.$$

Moreover, $\mathbb{P}(\mathcal{D}_\tau(\delta)) \geq 1 - \frac{(4t_f(\delta)+4)\delta}{T}$.

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

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

By our choice of η in Eq. (29), 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 $\alpha = 128 t_f(\delta)$ in Eq. (2) and by choosing η such that

$$\frac{c_1 L \eta^2 \chi^3 d}{m} \leq \frac{\eta}{\alpha} = \frac{\eta}{128 t_f(\delta)} \quad \Longleftrightarrow \quad \eta \leq \frac{m}{128 c_1 L t_f(\delta) d \chi^3},$$

it follows that

$$\begin{aligned}
&- \frac{3\eta}{4} \sum_{t \in J} \frac{1}{m} \sum_{i=1}^{m} \left| Z_{t,i}^\top \nabla f(x_t) \right|^2 + \left(\frac{\eta}{128 t_f(\delta)} + \frac{c_1 L \eta^2 \chi^3 d}{m} \right) \sum_{t \in J} \|\nabla f(x_t)\|^2 \\
&= - \frac{3\eta}{4} \sum_{t \in J} \frac{1}{m} \sum_{i=1}^{m} \left| Z_{t,i}^\top \nabla f(x_t) \right|^2 + \frac{\eta}{64 t_f(\delta)} \sum_{t \in J} \|\nabla f(x_t)\|^2 \\
&\leq \frac{\eta}{64 t_f(\delta)} \sum_{t \in J} \|\nabla f(x_t)\|^2 \\
&\leq \frac{\eta}{64 t_f(\delta)} \sum_{t \in J} \max_{t \in J} \|\nabla f(x_t)\|^2 \\
&\leq \frac{16\eta}{64 t_f(\delta)} \sum_{t \in J} \min_{t \in J} \|\nabla f(x_t)\|^2 \leq \frac{\eta}{4} \min_{t \in J} \|\nabla f(x_t)\|^2 \leq \frac{\eta}{4} \epsilon^2,
\end{aligned} \tag{52}$$

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

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

By our choice of η in Eq. (29), we can apply Lemma 17 to get

$$\|\nabla f(x_t)\| \leq c_5 t_f(\delta)\eta L \left(u^2 d^2 \rho \left(\log \frac{T}{\delta} \right)^2 + \sqrt{1 + \frac{\log(T/\delta)}{d}} r \right) \qquad \forall t \in J.$$

Note that, by our choices of the parameters η, u, r, it can be shown that

$$c_5 t_f(\delta)\eta L \left(u^2 d^2 \rho \left(\log \frac{T}{\delta} \right)^2 + \sqrt{1 + \frac{\log(T/\delta)}{d}} r \right) < \epsilon,$$

Hence, by setting $\alpha = 128 t_f(\delta)$ in Eq. (2) and choosing η such that

$$\frac{c_1 L \eta^2 \chi^3 d}{m} \leq \frac{\eta}{\alpha} = \frac{\eta}{128 t_f(\delta)},$$

it follows that

$$\left(\frac{\eta}{128 t_f(\delta)} + \frac{c_1 L \eta^2 \chi^3 d}{m} \right) \sum_{t \in J} \|\nabla f(x_t)\|^2$$

$$
\begin{aligned}
&\leq \frac{\eta}{64 t_f(\delta)} \sum_{t \in J} \left(c_5 t_f(\delta) \eta L \left(u^2 d^2 \rho \left(\log \frac{T}{\delta} \right)^2 + \sqrt{1 + \frac{\log(T/\delta)}{d} r} \right) \right)^2 \\
&\leq \frac{\eta}{64 t_f(\delta)} \sum_{t \in J} \epsilon^2 \\
&\leq \frac{\eta}{64} \epsilon^2
\end{aligned} \tag{53}
$$

Combining both cases above (Eq. (52) and Eq. (53)), we see that for the choice $\alpha = 128 t_f(\delta)$, the bound

$$
-\frac{3\eta}{4} \sum_{t \in J} \frac{1}{m} \sum_{i=1}^{m} \left| Z_{t,i}^{\top} \nabla f(x_t) \right|^2 + \left(\frac{\eta}{128 t_f(\delta)} + \frac{c_1 L \eta^2 \chi^3 d}{m} \right) \sum_{t \in J} \|\nabla f(x_t)\|^2 \leq \frac{\eta}{4} \epsilon^2 \tag{54}
$$

always holds.

Recall by Eq. (2) that we have

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

By plugging in Eq. (54) above, as well as the choice $\alpha = 128 t_f(\delta)$, we see that

$$
\begin{aligned}
f(x_\tau) - f(x_0) \leq & \frac{\eta}{4} \epsilon^2 + t_f(\delta) \eta u^4 \rho^2 \cdot c_1 d^3 \left(\log \frac{T}{\delta} \right)^3 + t_f(\delta) L \eta^2 u^4 \rho^2 \cdot c_1 d^4 \left(\log \frac{T}{\delta} \right)^4 \\
& + \eta c_1 r^2 (128 t_f(\delta) + \eta L) \log \frac{T}{\delta} + t_f(\delta) c_1 L \eta^2 r^2.
\end{aligned}
$$

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

$$
\begin{aligned}
\mathbb{P}(\mathcal{D}_\tau^c) &\leq \mathbb{P}(\mathcal{H}_\tau^c) + \sum_{t=0}^{\tau-1} \mathbb{P}(\mathcal{A}_t^c) + \sum_{t=0}^{\tau-1} \mathbb{P}(\mathcal{G}_t^c) \\
&\leq \frac{(\tau+4)\delta}{T} + \frac{\tau}{T}\delta + 2\frac{\tau}{T}\delta \leq \frac{(4 t_f(\delta) + 4)}{T} \delta
\end{aligned}
$$

□

Armed with Proposition 6 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 5, this yields the following result.

Theorem 2. *Suppose we pick u, r, η such that they satisfy the conditions in Proposition 6 and Lemma 25. Suppose F is chosen as prescribed in Lemma 25. Suppose that $\sqrt{\rho\epsilon} \leq 1$, so that $T_s \geq \frac{\iota}{\eta\sqrt{\rho\epsilon}} \geq \frac{d}{mL}$[8]. Suppose we pick T_s as prescribed in Proposition 6. Suppose in addition we pick r such that*

$$
r^2 \leq \min \left\{ \frac{\epsilon^2}{4(130 c_1 t_f(\delta) + c_1 \log(T/\delta) + c_1)}, \frac{F\sqrt{\rho\epsilon}}{80\iota \log(T/\delta) \left(\frac{65 c_5^2}{8} + 132 c_1 + 1 \right)} \right\}.
$$

Suppose also that we choose η such that

[8]Recall we focus on the case $\sqrt{\rho\epsilon} \leq L$, since otherwise, by the L-Lipschitz assumption, $\lambda_{\min}(\nabla^2 f(x)) \geq -L$ for all $x \in \mathbb{R}^d$, i.e. ϵ-first order stationary poits are also ϵ-second order stationary points.

$$\eta \leq \frac{0.1}{2\epsilon^2}\frac{\sqrt{\rho\epsilon}}{2\iota}\frac{1}{2}\left(\frac{\sqrt{\rho\epsilon}}{60c_9\iota\rho\log(T/\delta)}\right)^2 \frac{1}{t_f(\delta)\left(129+8c'^2\beta_1(\delta;F)\left(16(\mathrm{lr}(CT^2/\delta))^2+1\right)\right)}.$$

Suppose

$$T \geq \left\{\frac{256t_f(\delta)\left((f(x_0)-f^*)+\epsilon^2/L\right)}{\eta\epsilon^2}, \frac{\varphi\left(f(x_0)-f^*\right)}{\eta\epsilon^2}, 256\lceil\frac{\iota}{\eta\sqrt{\rho\epsilon}}\rceil, 256t_f(\delta), 1024\right\},$$

where

$$\varphi := 20\left(2\iota^2(60c_9\iota\log(T/\delta))^2\left(t_f(\delta)\right)\left(129+8c'^2\beta_1(\delta;F)\left(16(\mathrm{lr}(CT^2/\delta))^2+1\right)\right)\right).$$

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:

$$\tau_1 = \inf_t\{t \leq T : \|\nabla f(x_t)\| < \epsilon, \lambda_{\min}(\nabla^2 f(x_t)) \leq -\sqrt{\rho\epsilon}\},$$
$$\tau_{i+1} = \inf_t\{t \leq T : t > \tau_i + T_s, \|\nabla f(x_t)\| < \epsilon, \lambda_{\min}(\nabla^2 f(x_t)) \leq -\sqrt{\rho\epsilon}\}, \quad \forall 1 \leq i \leq \lfloor T/T_s \rfloor.$$

□

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

$$\begin{aligned}
&f(x_T) - f(x_0)\\
&= (f(x_{\tau_{N_s}}) - f(x_0)) + (f(x_T) - f(x_{\tau_{N_s}}))\\
&= (f(x_{\tau_1}) - f(x_0)) + \sum_{i=1}^{N_s}(f(x_{\tau_i+T_s}) - f(x_{\tau_i})) + \sum_{i=1}^{N_s-1}\left(f(x_{\tau_{i+1}}) - f(x_{\tau_i+T_s})\right) + (f(x_T) - f(x_{\tau_{N_s}}))\\
&= \underbrace{\sum_{i=1}^{N_s}(f(x_{\tau_i+T_s}) - f(x_{\tau_i}))}_{U_1} + \underbrace{(f(x_{\tau_1}) - f(x_0)) + \sum_{i=1}^{N_s-1}\left(f(x_{\tau_{i+1}}) - f(x_{\tau_i+T_s})\right)}_{U_2} + (f(x_T) - f(x_{\tau_{N_s}})).
\end{aligned}$$

We first consider U_1. Letting $x_j := x_T$ for any $j \geq T$, we have that

$$\sum_{i=1}^{N_s} f(x_{\tau_i+T_s}) - f(x_{\tau_i}) = \sum_{i=1}^{\lfloor T/T_s \rfloor}\left(f(x_{\tau_i+T_s}) - f(x_{\tau_i})\right)1_{\tau_i<T}$$

Now, by Eq. (30), 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

$$\begin{aligned}
\left(f(x_{\tau_i+T_s}) - f(x_{\tau_i})\right)1_{\tau_i<T} \leq\ & \tau\frac{c_5^2}{64}\eta^3 t_f(\delta)^2 L^2\left(u^2d^2\rho\left(\log\frac{T}{\delta}\right)^2 + \sqrt{2\log(T/\delta)}r\right)^2\\
&+ \tau\eta u^4\rho^2\cdot c_1 d^3\left(\log\frac{T}{\delta}\right)^3 + \tau L\eta^2u^4\rho^2\cdot c_1d^4\left(\log\frac{T}{\delta}\right)^4\\
&+ \eta c_1 r^2(128t_f(\delta)+\eta L)\log\frac{T}{\delta} + \tau c_1 L\eta^2 r^2\\
&:= M_{u,r,T_s}.
\end{aligned}$$

Suppose we pick u, r such that $M_{u,r,T_s} \leq 0.1F$. Recall from Proposition 6 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 $\mathbb{P}(\mathcal{E}_\tau) \leq 6\delta$. Then, by applying the weakened supermartingale inequality in Proposition 4, we have

$$\mathbb{P}\left(\sum_{i=1}^{T/T_s} (f(x_{\tau_i+T_s}) - f(x_{\tau_i}))\, 1_{\tau_i<T} \geq -N_s 0.9F + s\right) \leq \mathbb{E}\left[\exp\left(-\frac{s^2}{4N_sF^2}\right)\right] + \mathbb{P}(\mathcal{E}_\tau) \leq \exp\left(-\frac{s^2}{4(T/T_s)F^2}\right) + 6\delta.$$

Now, pick $s = 2F\sqrt{\sqrt{\log(1/\delta)}T/T_s}$, then

$$\mathbb{P}\left(\sum_{i=1}^{T/T_s} (f(x_{\tau_i+T_s}) - f(x_{\tau_i}))\, 1_{\tau_i<T} \geq -N_s 0.9F + 2F\sqrt{\sqrt{\log(1/\delta)}T/T_s}\right) \leq 7\delta.$$

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 $1 - 7\delta$,

$$U_1 = \sum_{i=1}^{T/T_s} (f(x_{\tau_i+T_s}) - f(x_{\tau_i}))\, 1_{\tau_i<T} \leq -(0.1T/T_s)F.$$

Next, we bound the summand U_2. Recall that

$$U_2 = (f(x_{\tau_1}) - f(x_0)) + \sum_{i=1}^{N_s-1} \left(f(x_{\tau_{i+1}}) - f(x_{\tau_i+T_s})\right).$$

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

$$f(x_{\tau_1}) - f(x_0) \leq \frac{\eta}{4}\epsilon^2 + t_f(\delta)\eta u^4\rho^2 \cdot c_1 d^3 \left(\log\frac{T}{\delta}\right)^3 + t_f(\delta)L\eta^2u^4\rho^2 \cdot c_1 d^4 \left(\log\frac{T}{\delta}\right)^4 + \eta c_1 r^2(128t_f(\delta) + \eta L)\log\frac{T}{\delta} + t_f(\delta)c_1 L\eta^2 r^2$$

   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

$$f(x_{\tau_1}) - f(x_0) \leq \tau_1 \frac{c_5^2}{64}\eta^3 t_f(\delta)^2 L^2 \left(u^2d^2\rho\left(\log\frac{T}{\delta}\right)^2 + \sqrt{2\log(T/\delta)}r\right)^2$$

$$+ \tau_1 \eta u^4 \rho^2 \cdot c_1 d^3 \left(\log \frac{T}{\delta}\right)^3 + \tau_1 L \eta^2 u^4 \rho^2 \cdot c_1 d^4 \left(\log \frac{T}{\delta}\right)^4$$
$$+ \eta c_1 r^2 (128 t_f(\delta) + \eta L) \log \frac{T}{\delta} + \tau_1 c_1 L \eta^2 r^2.$$

with probability at least $1 - \frac{(5\tau_1+4)\delta}{T}$.

By our choice of u, we know that

$$t_f(\delta)\eta u^4 \rho^2 \cdot c_1 d^3 \left(\log \frac{T}{\delta}\right)^3 + t_f(\delta) L \eta^2 u^4 \rho^2 \cdot c_1 d^4 \left(\log \frac{T}{\delta}\right)^4 + \eta c_1 r^2 (128 t_f(\delta) + \eta L) \log \frac{T}{\delta} + t_f(\delta) c_1 L \eta^2 r^2$$
$$\leq t_f(\delta) r^2 c_1 + t_f(\delta) r^2 c_1 + c_1 r^2 (128 t_f(\delta) + 1) \log(T/\delta) + c_1 r^2$$
$$= r^2 (130 c_1 t_f(\delta) + c_1 \log(T/\delta) + c_1).$$

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

$$\frac{\eta \epsilon^2}{4} \geq t_f(\delta)\eta u^4 \rho^2 \cdot c_1 d^3 \left(\log \frac{T}{\delta}\right)^3 + t_f(\delta) L \eta^2 u^4 \rho^2 \cdot c_1 d^4 \left(\log \frac{T}{\delta}\right)^4$$
$$+ \eta c_1 r^2 (128 t_f(\delta) + \eta L) \log \frac{T}{\delta} + t_f(\delta) c_1 L \eta^2 r^2.$$

Then, if $\tau_1 < t_f(\delta)$, with probability at least $1 - \frac{(5 t_f(\delta)+4)}{\delta}$,

$$f(x_{\tau_1}) - f(x_0) \leq \frac{\eta \epsilon^2}{2}.$$

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

$$\frac{F}{40}\frac{T}{T_s} \geq T\frac{c_5^2}{64}\eta^3 t_f(\delta)^2 L^2 \left(u^2 d^2 \rho \left(\log \frac{T}{\delta}\right)^2 + \sqrt{2\log(T/\delta)} r\right)^2$$
$$+ T\eta u^4 \rho^2 \cdot c_1 d^3 \left(\log \frac{T}{\delta}\right)^3 + T L \eta^2 u^4 \rho^2 \cdot c_1 d^4 \left(\log \frac{T}{\delta}\right)^4$$
$$+ \frac{T}{T_s}\eta c_1 r^2 (128 t_f(\delta) + \eta L) \log \frac{T}{\delta} + T c_1 L \eta^2 r^2.$$

Then, by a union bound, it follows that with probability at least $1 - 9\delta$,

$$U_2 = (f(x_{\tau_1}) - f(x_0)) + \sum_{i=1}^{N_s - 1} \left(f(x_{\tau_{i+1}}) - f(x_{\tau_i + T_s})\right)$$
$$\leq \frac{T}{T_s}\frac{\eta\epsilon^2}{2} + \frac{F}{40}\frac{T}{T_s}$$

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

$$\eta \leq \frac{0.1}{2\epsilon^2}\frac{\sqrt{\rho\epsilon}}{2\iota}\frac{1}{2}\left(\frac{\sqrt{\rho\epsilon}}{60 c_9 \iota \rho \log(T/\delta)}\right)^2 \frac{1}{t_f(\delta)\left(129 + 8c'^2 \beta_1(\delta; F)\left(16(\mathrm{lr}(CT^2/\delta))^2 + 1\right)\right)}$$

$$\leq \frac{0.1}{2\epsilon^2}\frac{1}{2}\left(\frac{\sqrt{\rho\epsilon}}{60c_9\iota\rho\log(T/\delta)}\right)^2 \frac{1}{\eta T_s t_f(\delta)\left(129+8c'^2\beta_1(\delta;F)\left(16(\mathrm{lr}(CT^2/\delta))^2+1\right)\right)} = \frac{0.1F}{2\epsilon^2},$$

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

$$-(0.05T/T_s)F \leq -(f(x_0)-f^*) \iff T \geq \frac{20T_s(f(x_0)-f^*)}{F} \geq \frac{\varphi\left(f(x_0)-f^*\right)}{\eta\epsilon^2}$$

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 5, 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.