

# Boundary Conditions for Linear Exit Time Gradient Trajectories Around Saddle Points: Analysis and Algorithm

Rishabh Dixit, Mert Gürbüzbalaban, and Waheed U. Bajwa

**Abstract**—Gradient-related first-order methods have become the workhorse of large-scale numerical optimization problems. Many of these problems involve nonconvex objective functions with multiple saddle points, which necessitates an understanding of the behavior of discrete trajectories of first-order methods within the geometrical landscape of these functions. This paper concerns convergence of first-order discrete methods to a local minimum of nonconvex optimization problems that comprise strict-saddle points within the geometrical landscape. To this end, it focuses on analysis of discrete gradient trajectories around saddle neighborhoods, derives sufficient conditions under which these trajectories can escape strict-saddle neighborhoods in linear time, explores the contractive and expansive dynamics of these trajectories in neighborhoods of strict-saddle points that are characterized by gradients of moderate magnitude, characterizes the non-curving nature of these trajectories, and highlights the inability of these trajectories to re-enter the neighborhoods around strict-saddle points after exiting them. Based on these insights and analyses, the paper then proposes a simple variant of the vanilla gradient descent algorithm, termed Curvature Conditioned Regularized Gradient Descent (CCRGD) algorithm, which utilizes a check for an initial boundary condition to ensure its trajectories can escape strict-saddle neighborhoods in linear time. Convergence analysis of the CCRGD algorithm, which includes its rate of convergence to a local minimum, is also presented in the paper. Numerical experiments are then provided on a test function as well as a low-rank matrix factorization problem to evaluate the efficacy of the proposed algorithm.

**Index Terms**—Boundary conditions, gradient descent, linear-time exit, Morse function, nonconvex optimization, saddle escape, strict-saddle property.

## I. INTRODUCTION

The gradient descent method and its (stochastic) variants have been at the forefront of nonconvex optimization for nearly a decade. Many of these variants stem from the earliest works like [1]–[3], the interior-point method [4]–[6], and their stochastic counterparts. But the highly complicated geometrical landscape of many nonconvex functions often puts the efficacy of these algorithms to question, which otherwise

have robust performance in convex settings. Indeed, problems involving matrix factorization [7], neural networks [8], rank minimization [9], etc., can be highly nonconvex, wherein the function geometry can possess many saddle points that create regions of very small magnitude gradients, something which the gradient-related methods rely upon heavily. As a consequence, travel times for trajectories generated by these methods in such regions could be exponentially large, thereby defeating the purpose of optimization. However, the large travel times around saddle points for gradient-based methods is not always the case; see, e.g., [10] that gives a linear exit-time bound for first-order approximations of gradient trajectories provided some necessary boundary conditions are satisfied by the trajectories. Such analysis suggests existence of gradient-based methods capable of ‘fast’ traversal of geometrical landscapes of nonconvex functions under appropriate conditions. Development of such methods, however, necessitates a deeper geometric analysis of the saddle neighborhoods so as to leverage any initial boundary conditions required by the faster gradient trajectories around saddle points in order to reduce the total travel time on the entire function landscape.

To this end, we *first* study in this paper the problem of developing sufficient boundary conditions for gradient trajectories around any saddle point  $\mathbf{x}^*$  of some nonconvex function  $f(\mathbf{x})$  that can guarantee linear exit time, i.e.,  $K_{\text{exit}} = \mathcal{O}(\log(\varepsilon^{-1}))$ , from the open saddle neighborhood  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$ . This problem focuses on a closed neighborhood  $\bar{\mathcal{B}}_\varepsilon(\mathbf{x}^*)$  around the saddle point  $\mathbf{x}^*$ , with the current iterate  $\mathbf{x}_0$  sitting on the boundary of this neighborhood, i.e.,  $\mathbf{x}_0 \in \bar{\mathcal{B}}_\varepsilon(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*)$ . Suppose also that the gradient trajectory starting at  $\mathbf{x}_0$  has approximately linear exit time from this region  $\bar{\mathcal{B}}_\varepsilon(\mathbf{x}^*)$ . (Existence of such trajectories is guaranteed because of the analysis in [10].) Then, the question posed here is what are the sufficient conditions on  $\mathbf{x}_0$  such that the trajectory can escape  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  in almost linear time of order  $\mathcal{O}(\log(\varepsilon^{-1}))$ . Once the sufficient conditions have been derived, we *next* study the question of whether it is possible to get linear rates of travel by the same gradient trajectory in some bigger neighborhood  $\mathcal{B}_\xi(\mathbf{x}^*) \supset \mathcal{B}_\varepsilon(\mathbf{x}^*)$ . Note that unlike the matrix perturbation-based analysis in [10], the radius  $\xi$  of the bigger neighborhood needs to be characterized by a fundamentally different proof technique. This is since the eigenspace of the Hessian  $\nabla^2 f(\mathbf{x})$  for any  $\mathbf{x} \in \mathcal{B}_\xi(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*)$  cannot be obtained by perturbing the eigenspace of  $\nabla^2 f(\mathbf{x}^*)$  since the series expansion of  $\nabla^2 f(\mathbf{x})$  about  $\nabla^2 f(\mathbf{x}^*)$  may not necessarily converge from matrix perturbation theory. *Third*, after such linear rates have

R. Dixit (Department of Electrical and Computer Engineering), M. Gürbüzbalaban (Departments of Electrical & Computer Engineering, Management Science and Information Systems, and Statistics), and W. U. Bajwa (Departments of Electrical & Computer Engineering and Statistics) are at Rutgers University–New Brunswick, NJ 08854, USA (Emails: {rishabh.dixit, mg1366, waheed.bajwa}@rutgers.edu).

This work was supported in part by the National Science Foundation under grants CCF-1453073, CCF-1907658, CCF-1814888, DMS-2053485, and CCF-1910110, by the Army Research Office under grants W911NF-17-1-0546 and W911NF-21-1-0301, by the Office of Naval Research under grant N00014-21-1-2244, and by the DARPA Lagrange Program under ONR/SPAWAR contract N660011824020.

been obtained, we then study whether it is possible to develop a robust algorithm that leverages the boundary conditions so as to steer the gradient trajectory away from  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  in almost linear time. *Finally*, we seek an answer to the question of whether the developed algorithm converges to a neighborhood of a local minimum and, if so, what would be its rate of convergence within the global landscape of the nonconvex function.

To address all these problems effectively, we engage in a rigorous analysis of trajectories of the vanilla gradient descent method, starting off directly where we left in [10].<sup>1</sup> First, we utilize tools from the matrix perturbation theory to develop sufficient conditions on  $\mathbf{x}_0 \in \tilde{\mathcal{B}}_\varepsilon(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*)$  for which the subsequent gradient trajectory has linear exit time from  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$ . Next, we prove a rather intuitive yet extremely powerful result, termed the *sequential monotonicity of gradient trajectories*, which establishes that the gradient trajectories in a neighborhood of the saddle point first exhibit contractive dynamics up to some point and there onward strictly expansive dynamics. Next, we provide an analysis of the travel time for the gradient trajectory in the region  $\mathcal{B}_\xi(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*)$  using the sequential monotonicity result. Finally, we develop a novel gradient-based algorithm, termed Curvature Conditioned Regularized Gradient Descent (CCRGD), around the idea of sufficient boundary conditions with a robust check condition guaranteeing almost linear exit time from  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$ . In doing so, we also prove certain qualitative lemmas about the local behavior of gradient trajectories around saddle points. Thereafter, the asymptotic convergence and the rate of convergence for CCRGD to a local minimum is proved using these lemmas. Finally, the performance of CCRGD is evaluated on two problems: a test function for nonconvex optimization and a low-rank matrix factorization problem.

### A. Relation to Prior Work

Since this work directly extends the results in [10], we steer away from repeating the discussion in [10, Sec. 1.1] in relation to existing convergence guarantees for gradient-related methods in nonconvex settings. Instead, we primarily focus in this section on presenting comparisons and highlighting key differences between our contributions and the existing literature. In addition, given the vast interest of the optimization community in nonconvex optimization using gradient-related methods, we also discuss some additional relevant works in here.

Similar to [11], which focuses on the first order methods, we prove in Theorem 5 that the trajectories generated by the proposed CCRGD algorithm (see Algorithm 1) converge to a local minimum. But unlike [11], which fundamentally uses the Stable Manifold Theorem [12], we also develop in this paper a proof of convergence of CCRGD to a local minimum and obtain algorithmic convergence rates using the geometry of function landscape near saddle points and in regions that have sufficiently large gradient magnitudes. Though this idea

of rate analysis has been well summarized in [13] for gradient-related sequences and more recently in [14] for Newton-type methods, yet these works do not utilize the nonconvex geometry to its fullest extent. Specifically, we categorize the function geometry in our work into ‘regions near’ and ‘regions away’ from the stationary points so as to better analyze ‘escape conditions’ from saddle neighborhoods and at the same time generate convergence guarantees to a local minimum. Within the regions of ‘moderate gradients’ around saddle points, i.e., the shell  $\mathcal{B}_\xi(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*)$ , we show using the sequential monotonicity property (detailed in Theorem 2) that the sequence  $\{\|\mathbf{x}_k - \mathbf{x}^*\|\}$  is strictly monotonic whenever the iterate  $\{\mathbf{x}_k\}$  has expansive dynamics with respect to  $\mathbf{x}^*$ , while the function value sequence  $\{f(\mathbf{x}_k)\}$  satisfies the Polyak–Łojasiewicz (PL) condition [15] whenever the iterate sequence  $\{\mathbf{x}_k\}$  has contractive dynamics with respect to  $\mathbf{x}^*$  (see Lemma 1). Consequently, linear rates of contraction to a point on the boundary  $\mathcal{B}_\varepsilon(\mathbf{x}^*) \setminus \mathcal{B}_\xi(\mathbf{x}^*)$  are derived using the PL condition and linear rates of expansion to a point on the boundary  $\mathcal{B}_\xi(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*)$  are obtained using the sequential monotonicity property from Theorem 2, both of which aid in our convergence analysis. Note that the PL condition cannot be applied directly around a saddle point since that would yield a trivial lower bound of 0 on the gradient norm (see Lemma 1). This particular analytical approach of separately analyzing the contractive and expansive dynamics locally around a saddle point and exploiting the PL condition restricted to contractive dynamics is in contrast to the existing works that focus on the problem of escaping saddle points for nonconvex optimization. In addition, while the PL condition or the more general Kurdyka–Łojasiewicz property [16] are often used for local or even global analysis such as in [17] and [18], they have not been used in the context of analyzing local contractive dynamics of iterates w.r.t. a strict saddle point. In terms of the analytical tools used, regions near the saddle points in this work are analysed using the matrix perturbation theory, yielding sharp bounds (‘sharp’ in terms of the condition number, problem dimension, and spectral gap) on the initial conditions, whereas regions away from the saddle points utilize properties like the sequential monotonicity (cf. Theorem 2). Such local analysis distinguishing sufficiently small saddle neighborhoods from moderately small saddle neighborhoods seems to be quite novel and has not been carried out in any previous work to our knowledge.

Next, to the best of our knowledge, no other work has provided sufficient boundary conditions for escape from saddle neighborhoods for the case of *discrete-time* gradient descent-related algorithms. Though the idea is not necessarily new and has been explored while dealing with continuous-time dynamical systems, specifically the boundary value problems, yet it is still nascent when it comes to analyzing saddle points. The continuous-time works such as [18]–[20] have been discussed in detail in [10]. However even these works do not analyze the boundary conditions for continuous trajectories. The work [20] does take into account cascaded saddles encountered by continuous trajectories, which gets a detailed treatment in our work in Theorems 6 and 7 for discrete trajectories.

The Stochastic Differential Equation (SDE) setup has also been utilized in a recent work [21] to study gradient-

<sup>1</sup>Since this work is a continuation of [10], we refrain from elaborating certain terminologies and definitions that were covered in detail in [10], though a summary of all the required concepts is provided in Sec. III-A to make this a self-contained paper.

based (stochastic) methods for nonconvex optimization in the continuous-time setting. Interestingly, this work considers the set of *index-1 saddle points* in the function's geometry and thereby obtains a stochastic rate of convergence to a global minimum, where the rate is of the order 'a constant term plus a geometric term'. While the rate is linear/geometric, [21] assumes the *coercivity condition* (sufficient growth condition on the function away from the origin) and the *Villani condition* (growth of gradient's norm), whereas only the former condition of coercivity is assumed in our work. Also, the constant in the non-geometric term of the rate is dependent on the horizon  $T$  obtained from discretization of the SDE, which could be large. Moreover, it is not clear how the SDE approach in [21] would apply to the discrete-time setting of this paper.

Recently, within the class of discrete-time non-acceleration-based methods, [22] studies saddle escape in the context of tensor decomposition, [23], [24] provide the rates for escaping saddles using perturbed gradient descent, [25] utilizes the notion of variational coherence between stochastic mirror gradient and descent direction in quasi convex and nonconvex problems for obtaining ergodic rates of convergence to a local/global minimum (under certain conditions), and [26] provides rates and escape guarantees under certain strong assumptions of high correlation between the negative curvature direction and a random perturbation vector. However, none of these stochastic variants explore the idea of initial boundary conditions near saddle points so as to obtain linear rates. It should be noted that the work in [23] shows the time to escape cascaded saddles scales exponentially with dimension, whereas we show in Theorem 7 that the time to escape cascaded saddles is not exponential in dimension. Rather, the number of cascaded saddles encountered by the trajectory is upper bounded and this bound scales only linearly with the inverse of the gradient norms in regions away from the stationary points of the objective. Further, this upper bound on the number of saddles encountered is independent of the problem dimension.

The next set of related discrete-time gradient-based methods includes first-order methods leveraging acceleration and momentum techniques. For instance, the work in [27] provides an extension of SGD to methods like the Stochastic Variance Reduced Gradient (SVRG) algorithm for escaping saddles. Recently, methods approximating the second-order information of the function that preserve the first-order nature of the algorithm have also been employed to escape the saddles. Examples include [28], where the authors prove that an acceleration step in gradient descent guarantees escape from saddle points, and the method in [29], which utilizes the second-order nature of the acceleration step combined with a stochastic perturbation to guarantee escape rates. Moreover, both [30], [31] build on the idea of utilizing acceleration as a source of finding the negative curvature direction. Due to the low computational cost of evaluating gradients, we also make use of such connections between the curvature magnitude and the gradient difference in our proposed algorithm (Algorithm 1). In the class of first-order algorithms, there also exist trust region-based methods. The work in [32] is one such method that presents a novel stopping criterion with a heavy ball controlled mechanism for escaping saddles using the SGD

method. If the SGD iterate escapes some neighborhood in a certain number of iterations, the algorithm is restarted with the next round of SGD, else the ergodic average of the iterate sequence is designated to be a second-order stationary solution. In a similar vein, we formally derive in Lemma 6 the escape guarantees from a neighborhood around a saddle point and utilize that result within the proposed Algorithm 1.

Lastly, higher-order methods are discussed in [33], [34], which utilize either Hessian-based approaches or a second-order step combined with first-order algorithms so as to reach local minimum with fast speed while trading off with computational costs. Going a step even further, the work in [35] poses the escape problem with second-order saddles, thereby motivating the use of higher-order methods. Though these techniques optimize well over certain pathological functions like those having 'degenerate' saddles or very ill-conditioned geometries, yet they suffer heavily in terms of complexity; e.g., the work [35] requires third-order methods to solve for a feasible descent direction. This further motivates us to develop a hybrid algorithm for the saddle escape problem that captures the advantages of a Hessian-based method and at the same time is low on computational complexity.

Table 1 draws comparisons between our work and other existing works within the realm of saddle escape in deterministic nonconvex optimization problems. Though there is a plethora of works that study the saddle escape problem, only those works are listed here that address the simple unconstrained optimization problem of minimizing a smooth nonconvex function  $f(\cdot)$  and propose perturbation of deterministic gradient-based methods for saddle escape. Many of the other related works discussed in this section tackle stochastic optimization problems and are therefore not included in the table.

## B. Our Contributions

This work starts off directly from the point where we left off in [10], where we obtained exit time bounds for  $\varepsilon$ -precision gradient descent trajectories around saddle points and derived a necessary condition on the initial unstable subspace projection value for linear exit time. The first novel result in this work is the development of a bound on the initial unstable subspace projection value in Theorem 1 that approximately guarantees the linear exit time bound from [10, Theorem 3.2]. Our second contribution is Theorem 2 in which we analyze the behavior of gradient descent trajectories in some region  $\mathcal{B}_\xi(\mathbf{x}^*) \supset \mathcal{B}_\varepsilon(\mathbf{x}^*)$  where the approximate analysis from matrix perturbation theory may not necessarily hold. In such augmented neighborhood of the strict saddle point  $\mathbf{x}^*$ , we prove that the gradient descent trajectories have a sequential monotonic behavior, i.e., there exists some  $\xi$  such that the trajectory inside  $\mathcal{B}_\xi(\mathbf{x}^*)$  first exhibits contractive dynamics moving towards  $\mathbf{x}^*$  and then has expansive dynamics for the remainder of the time as long as it stays inside  $\mathcal{B}_\xi(\mathbf{x}^*)$ . Though this property may appear to be trivial for trajectories around saddle points, yet it is extremely important in developing improved rates/travel times of the gradient descent trajectories inside  $\mathcal{B}_\xi(\mathbf{x}^*)$ , which follows from our next contribution. Our third contribution is Theorem 3 in which we obtain upper bounds on the travel

TABLE I  
SUMMARY OF THE SIMILARITIES AND DIFFERENCES BETWEEN THIS WORK AND SOME RELATED PRIOR WORKS.

| References       | Method of saddle escape                                          | Base algorithm                      | Explicit dependence on number of saddles | Convergence rate                                                                                                                                                                                                                                                                                                                                       | Type of convergence rate |
|------------------|------------------------------------------------------------------|-------------------------------------|------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------|
| [24]             | One-step noise                                                   | Gradient descent method             | $\times$                                 | $\mathcal{O}\left(\frac{1}{\varepsilon^2} \log^4\left(\frac{1}{\varepsilon^2}\right)\right)$                                                                                                                                                                                                                                                           | probabilistic            |
| [28]             | One-step noise with negative curvature search                    | Accelerated gradient method         | $\times$                                 | $\mathcal{O}\left(\frac{1}{\varepsilon^{7/4}} \log^6\left(\frac{1}{\varepsilon}\right)\right)$                                                                                                                                                                                                                                                         | probabilistic            |
| [33]             | One-step noise with negative curvature search                    | Second-order Newton method          | $\checkmark$                             | $\mathcal{O}\left(T \log\left(\frac{1}{\varepsilon}\right) + T \log \log\left(\frac{1}{\varepsilon}\right)\right)$ ;<br>$T$ is the number of saddles encountered                                                                                                                                                                                       | probabilistic            |
| [36]             | Multi-step noise with negative curvature search                  | Accelerated gradient method         | $\times$                                 | $\mathcal{O}\left(\frac{1}{\varepsilon^{7/4}} \log\left(\frac{1}{\varepsilon}\right)\right)$                                                                                                                                                                                                                                                           | probabilistic            |
| [37]             | Multi-step noise with negative curvature search                  | Adaptive negative curvature descent | $\times$                                 | $\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$                                                                                                                                                                                                                                                                                                      | probabilistic            |
| [38]             | One-step noise followed by multi-step negative curvature search  | Accelerated gradient method         | $\times$                                 | $\mathcal{O}\left(\frac{1}{\varepsilon^{7/4}} \log\left(\frac{1}{\varepsilon}\right)\right)$                                                                                                                                                                                                                                                           | probabilistic            |
| <b>This work</b> | One second-order step <i>only</i> when curvature condition fails | Gradient descent method             | $\checkmark$                             | $\mathcal{O}\left(T \log\left(\frac{1}{\varepsilon}\right)\right) + \mathcal{O}\left(T \log\left(\frac{\xi}{\varepsilon}\right)\right) + \mathcal{O}\left(\frac{1}{\varepsilon^{2v}}\right)$ ;<br>for locally analytic, coercive Morse functions;<br>$T = \mathcal{O}\left(\frac{1}{\varepsilon^v}\right)$ is the number of saddles and $v \in [0, 1)$ | deterministic            |

<sup>1</sup>The parameter  $v$  is defined in Proposition 5 and it controls the function geometry in regions away from its critical points.

time of gradient trajectory inside the shell  $\mathcal{B}_\xi(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*)$  that we denote by  $K_{shell}$ . This particular region is specifically of great importance since we can categorize it as a region of “moderate” gradients (gradient magnitude not too small) that still inherits certain geometric properties such as the minimum curvature from the smaller saddle neighborhood  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$ . Without taking such properties into consideration, the journey time in this shell could only be naively upper bounded as  $K_{shell} = \mathcal{O}(\varepsilon^{-2})$  using the gradient Lipschitz condition. Hence, it is imperative to separately analyze the journey time inside the shell  $\mathcal{B}_\xi(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*)$  so as to improve upon the standard nonconvex rate of  $\mathcal{O}(\varepsilon^{-2})$ .

Our next set of contributions corresponds to Lemmas 2-6, in which we provide insights into certain qualitative properties of the gradient descent trajectories around saddle points. Lemma 2 talks about the approximate hyperbolic nature of the gradient trajectories near saddle points, while Lemma 3 proves that trajectories with linear exit time approximately never curve around saddle points. Lemma 4 shows that the gradient trajectory can only exit  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  at those points where the function value is strictly less than  $f(\mathbf{x}^*)$ . Lemma 5 establishes that the gradient trajectory, once it exits the neighborhood  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$ , can never re-enter it, while Lemma 6 extends the same result to the bigger neighborhood  $\mathcal{B}_\xi(\mathbf{x}^*)$  under certain stricter conditions. Our next contribution is the development of the Curvature Conditioned Regularized Gradient Descent (CCRGD) algorithm (cf. Algorithm 1) that provably escapes saddle neighborhoods and gives second-order stationary solutions. The asymptotic convergence of the proposed algorithm is established from Theorem 5, which is proved using Lemmas 9, 10, 11 and the Global Convergence Theorem (Theorem 4) from [39]. The algorithm checks for a curvature condition near the saddle neighborhood and makes the decision of whether to perform a second-order iteration for one step or continue using the vanilla gradient descent method. The curvature condition (Step 15 in Algorithm 1) is derived from our proof of convergence of the algorithm; in addition, Algorithm 1 is

tested for its efficacy on a modified Rastrigin function (a test function for nonconvex optimization) and the matrix factorization problem as part of numerical experiments. Last, but not the least, the final contribution of this work is derivation of the rate of convergence of an iterate sequence generated from Algorithm 1 to a local minimum. The rates are obtained for a more general setting of cascaded saddles where the number of saddles encountered and the total time of convergence are bounded from Theorems 6 and 7, respectively. Finally, the convergence guarantees in Theorems 5 and 7 are extended to functions which are not globally gradient and Hessian Lipschitz continuous based on Theorems 8 and 9.

### C. Notation

All vectors in the paper are in bold lower-case letters, all matrices are in bold upper-case letters,  $\mathbf{0}$  is the  $n$ -dimensional null vector,  $\mathbf{I}$  represents the  $n \times n$  identity matrix, and  $\langle \cdot, \cdot \rangle$  represents the inner product of two vectors. In addition, unless otherwise stated, all vector norms  $\|\cdot\|$  are  $\ell_2$  norms, while the matrix norm  $\|\cdot\|_2$  denotes the operator norm. Further, the symbol  $(\cdot)^T$  is the transpose operator, the symbol  $\mathcal{O}$  represents the Big-O notation and sometimes we use  $a \ll b \iff a = \mathcal{O}(b)$ , the symbol  $\Omega$  is the Big-Omega notation and  $\Theta$  represents the Big-Theta notation,  $\otimes$  represents the kronecker product, i.o. means infinitely often,  $\text{id}$  represents the identity map, and  $W(\cdot)$  is the Lambert  $W$  function [40]. Throughout the paper,  $k$  and  $K$  are used for the discrete time.

Next,  $\gtrsim$  and  $\lesssim$  represent the ‘approximately greater than’ and ‘approximately less than’ symbols, respectively, where  $a \lesssim b$  implies  $a \leq b + g(\varepsilon)$  and  $a \gtrsim b$  implies  $a + g(\varepsilon) \geq b$  for some absolutely continuous function  $g(\cdot)$  of  $\varepsilon$  where  $g(\cdot) \geq 0$  and  $g(\varepsilon) \rightarrow 0$  as  $\varepsilon \rightarrow 0$ . Also, for any matrix expressed as  $\mathbf{Z} + \mathcal{O}(c)$  with  $c$  being a scalar, the matrix-valued perturbation term  $\mathcal{O}(c)$  is with respect to the Frobenius norm. Finally, the operator  $\text{dist}(\cdot, \cdot)$  gives the distance between two sets whereas  $\text{diam}(\cdot)$  gives the diameter of a set.

## II. PROBLEM FORMULATION

Consider a nonconvex smooth function  $f(\cdot)$  that has strict first-order saddle points in its geometry. By strict first-order saddle points, we mean that the Hessian of function  $f(\cdot)$  at these points has at least one negative eigenvalue, i.e., the function has negative curvature. Next, consider some (open) neighborhood  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  around a given saddle point  $\mathbf{x}^*$ , where the neighborhood radius  $\varepsilon$  is bounded above by  $\Theta(LM^{-1})$  (see [10, Theorem 3.2] for the exact form) with  $L$  and  $M$  being the gradient and Hessian Lipschitz constants of  $f(\cdot)$ . Also, it is given that the initial iterate  $\mathbf{x}_0$  of the gradient trajectory sits on the boundary of the neighborhood, i.e.,  $\mathbf{x}_0 \in \mathcal{B}_\varepsilon(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*)$ , and the gradient trajectory exits  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  in linear time bounded by [10, Theorem 3.2]. With this information, we are first interested in finding the sufficient conditions on  $\mathbf{x}_0$  that guarantee the linear exit time. In addition, we need to analyze the gradient trajectories in some larger neighborhood  $\mathcal{B}_\xi(\mathbf{x}^*) \supset \mathcal{B}_\varepsilon(\mathbf{x}^*)$  such that the trajectories first contract towards the saddle point and then expand away from it. More importantly, we are interested in finding such  $\xi > \varepsilon$  for which the gradient trajectory has linear travel time in the shell  $\mathcal{B}_\xi(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*)$ . Next, we are required to find certain local properties of  $f(\cdot)$  for which the gradient trajectories, having escaped it once, can never re-enter the neighborhood  $\mathcal{B}_\xi(\mathbf{x}^*)$ . Finally, we have to develop a robust low-complexity algorithm that utilizes the sufficient conditions to traverse the landscape of saddle neighborhoods in linear time and also provide its rate of convergence to some local minimum.

Having briefly stated the problem, we now formally state the set of assumptions that are required for this problem to be tackled in this work.

### A. Assumptions

- A1.** The function  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  is coercive, i.e.,  $\lim_{\|\mathbf{x}\| \rightarrow \infty} f(\mathbf{x}) = \infty$ , is globally  $\mathcal{C}^2$ , i.e., twice continuously differentiable, and is locally  $\mathcal{C}^\omega$  in sufficiently large neighborhoods of its saddle points, i.e., all the derivatives of this function are continuous around saddle points and the function  $f(\cdot)$  also admits Taylor series expansion in these neighborhoods<sup>2</sup>.
- A2.** The gradient of function  $f(\cdot)$  is  $L$ -Lipschitz continuous:  $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ .
- A3.** The Hessian of function  $f(\cdot)$  is  $M$ -Lipschitz continuous:  $\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\|_2 \leq M\|\mathbf{x} - \mathbf{y}\|$ .
- A4.** The function  $f(\cdot)$  has only well-conditioned first-order stationary points, i.e., no eigenvalue of the function's Hessian is close to zero around these points. Formally, if  $\mathbf{x}^*$  is the first-order stationary point for  $f(\cdot)$ , then

$$\nabla f(\mathbf{x}^*) = \mathbf{0}, \text{ and } \min_i |\lambda_i(\nabla^2 f(\mathbf{x}^*))| > \beta,$$

where  $\lambda_i(\nabla^2 f(\mathbf{x}^*))$  denotes the  $i^{\text{th}}$  eigenvalue of the matrix  $\nabla^2 f(\mathbf{x}^*)$  and  $\beta > 0$ . Note that such a function

is termed a Morse function. Also, there exists an open neighborhood  $\mathcal{W}$  of  $\mathbf{x}^*$  such that

$$\forall \mathbf{x} \in \mathcal{W}, \min_i |\lambda_i(\nabla^2 f(\mathbf{x}))| > \beta.$$

**Remark 1.** The coercivity of  $f(\cdot)$  is only required from Section IV onward, where we prove the convergence of Algorithm 1. Also, Section IV requires  $f(\cdot)$  to be only  $\mathcal{C}^2$  Hessian-Lipschitz Morse function, unlike Section III in which the additional assumption of local analyticity is required around saddle points.

Note that Assumption A1 may seem too restrictive since it requires  $f(\cdot)$  to be locally real analytic, while the theory of nonconvex optimization is often developed around only the assumption that  $f \in \mathcal{C}^2$  with Lipschitz-continuous Hessian. It is worth reminding the reader, however, that many practical nonconvex problems such as quadratic programs, low-rank matrix completion, phase retrieval, etc., with appropriate smooth regularizers satisfy this assumption of real analyticity around the saddle neighborhoods; see, e.g., the formulations discussed in [41]. Similarly, many of the loss functions in nonconvex optimization are coercive, i.e., they grow arbitrarily large asymptotically due to the presence of some form of regularization. As for the other assumptions, gradient Lipschitz continuity (Assumption A2) and Hessian Lipschitz continuity (Assumption A3) are invoked routinely in the nonconvex optimization literature<sup>3</sup>, while Assumption A4 implies  $f(\cdot)$  is a Morse function. In particular, since Morse functions are dense in the class of  $\mathcal{C}^2$  functions [42], we are not giving up much by making this assumption. We now state two propositions that follow from our assumptions and that will be routinely used in our analysis.

**Proposition 1.** Under Assumption A4, the function  $f(\cdot)$  has only first-order saddle points in its geometry. Moreover, these first-order saddle points are strict saddle, i.e., for any first-order saddle point  $\mathbf{x}^*$ , there exists at least one eigenvalue  $\lambda_i$  of  $\nabla^2 f(\mathbf{x}^*)$  that satisfies  $\lambda_i(\nabla^2 f(\mathbf{x}^*)) < -\beta$ .

*Proof.* For any  $\mathcal{C}^m$ -smooth function  $f(\cdot)$  with  $m \geq 2$ , if  $\mathbf{x}^*$  is its second- or higher-order saddle point then it must necessarily satisfy  $\nabla f(\mathbf{x}^*) = \mathbf{0}$  and  $\nabla^2 f(\mathbf{x}^*) \succeq \mathbf{0}$ , where at least one of the eigenvalues of  $\nabla^2 f(\mathbf{x}^*)$  is 0. But this is not possible in our case because of Assumption A4. The fact that such an eigenvalue  $\lambda_i$  exists is also a direct consequence of Assumption A4. ■

**Proposition 2.** Under Assumption A4, for any sufficiently small  $\varepsilon$  where  $\varepsilon \ll \beta$ , we can group the eigenvalues of the Hessian  $\nabla^2 f(\mathbf{x}^*)$  at any strict saddle point  $\mathbf{x}^*$  into  $m$  disjoint sets  $\{\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_m\}$  with  $2 \leq m \leq n$  based on the level of degeneracy of eigenvalues (closeness to one another) such that for some  $\delta = \Omega(\varepsilon^{1-a})$  where  $a \in (0, 1]$ , we have the following conditions:

$$\begin{aligned} \text{dist}(\mathcal{G}_p, \mathcal{G}_q) &\geq \delta \quad \forall \mathcal{G}_p, \mathcal{G}_q \text{ s.t. } p \neq q, \text{ and} \\ \max_p \{\text{diam}(\mathcal{G}_p)\} &= \mathcal{O}(\varepsilon^{1-a}). \end{aligned} \quad (1) \quad (2)$$

<sup>2</sup>By sufficiently large neighborhoods, we mean that the diameter of such neighborhoods is  $\Omega(1)$ .

<sup>3</sup>Note that later in this work, we will show that Assumptions A2, A3 can be removed based on Theorems 8 and 9.

*Proof.* From Assumption **A4**, the eigenvalues of the Hessian  $\nabla^2 f(\mathbf{x}^*)$  at any strict saddle point  $\mathbf{x}^*$  can always be separated into two distinct groups, one consisting of positive eigenvalues and the other comprising negative eigenvalues. By this construction, the distance between these groups will be at least  $2\beta$ . Since  $\varepsilon \ll \beta$ , we get a  $\delta = 2\beta$  for this construction which satisfies the constraint  $\delta = \Omega(1)$ . Next, we check whether the diameter of these two groups is larger than  $\Theta(\varepsilon^{1-a})$ ; if yes then we split that particular group into two more groups at the first eigenvalue where the consecutive eigenvalue gap within that group exceeds  $\Theta(\varepsilon^{1-a})$ . This eigenvalue gap becomes our new  $\delta$  and by construction it will satisfy the constraint  $\delta = \Omega(\varepsilon^{1-a})$  for some  $a > 0$  since  $\delta > \Theta(\varepsilon^{1-a})$ . Repeating this process recursively, we would have constructed the disjoint sets  $\{\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_m\}$  with  $2 \leq m \leq n$ . Since  $n$  is finite, this process will terminate in finite steps (maximum  $n-1$  steps) and therefore after the final splitting, we will obtain  $\delta = \Omega(\varepsilon^{1-a})$  for some  $a \in (0, 1]$  such that  $\max_p \{\text{diam}(\mathcal{G}_p)\} = \Theta(\varepsilon^{1-a})$ . ■

Proposition 2 describes a fundamental property of any  $\mathcal{C}^2$  function that arises due to the algebraic multiplicity / (approximate) degeneracy of the eigenvalues of its Hessian at the saddle points. Note that, as a consequence of the strict-saddle property (Assumption **A4** / Proposition 1) and Proposition 2, we get the following necessary condition:

$$\beta \geq \frac{\delta}{2}. \quad (3)$$

### III. BOUNDARY CONDITIONS FOR LINEAR EXIT TIME FROM A SADDLE NEIGHBORHOOD

#### A. Preface

Given a saddle neighborhood  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  for some strict saddle point  $\mathbf{x}^*$  and  $\varepsilon > 0$ , the goal is selecting those gradient trajectories in  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  for which the exit time is of the order  $K_{\text{exit}} = \mathcal{O}(\log(\varepsilon^{-1}))$ , i.e., of linear rate. Formally, the exit time for an iterate sequence  $\{\mathbf{x}_k\}$  of some trajectory in the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  is defined as the smallest positive index  $K$  such that  $\|\mathbf{x}_K - \mathbf{x}^*\| \geq \varepsilon$  and we are required to obtain such sequence  $\{\mathbf{x}_k\}$  generated by the gradient descent method for which the exit time from the saddle neighborhood  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  is linear. To conduct such analysis, certain essential concepts and definitions need to be elaborated, most of which were developed in a previous work (for reference see [10]).

First, due to the strict-saddle property, for any  $\mathbf{x}$  in an  $\varepsilon$ -neighborhood of  $\mathbf{x}^*$ , i.e.,  $\mathbf{x} \in \mathcal{B}_\varepsilon(\mathbf{x}^*)$ , the vector  $\mathbf{x} - \mathbf{x}^*$  belongs to a vector space  $\mathcal{E} = \mathcal{E}_S \oplus \mathcal{E}_{US}$ , where

$$\begin{aligned} \mathcal{E}_S &= \text{span}\{\mathbf{v}_i | \lambda_i > 0\}, \mathcal{N}_S = \{i | \lambda_i > 0\}, \\ \mathcal{E}_{US} &= \text{span}\{\mathbf{v}_i | \lambda_i < 0\}, \mathcal{N}_{US} = \{j | \lambda_j < 0\}, \end{aligned}$$

and  $(\lambda_i, \mathbf{v}_i)$  are the  $i^{\text{th}}$  eigenvalue–eigenvector pair of the Hessian  $\nabla^2 f(\mathbf{x}^*)$ .

Second, using the ‘degenerate’ matrix perturbation theory [43], [44], the Hessian  $\nabla^2 f(\mathbf{x})$  at any point  $\mathbf{x} = \mathbf{x}^* + p\mathbf{u}$ , where  $p \in [0, 1]$  and  $\|\mathbf{u}\| \leq \varepsilon$ , can be given as

$$\nabla^2 f(\mathbf{x}) = \nabla^2 f(\mathbf{x}^*) + p\|\mathbf{u}\| \mathbf{H}(\hat{\mathbf{u}}) + \mathcal{O}(\varepsilon^2), \quad (4)$$

where  $\mathbf{u} := \mathbf{x} - \mathbf{x}^*$  is termed the **radial vector**,  $\hat{\mathbf{u}} = \frac{\mathbf{u}}{\|\mathbf{u}\|}$  is the unit radial vector and we have that

$$\mathbf{H}(\hat{\mathbf{u}}) = \sum_{i=1}^n \left( \langle \mathbf{v}_i, \mathbf{H}(\hat{\mathbf{u}}) \mathbf{v}_i \rangle \mathbf{v}_i \mathbf{v}_i^T + \lambda_i \sum_{l \notin \mathcal{G}_i} \frac{\langle \mathbf{v}_l, \mathbf{H}(\hat{\mathbf{u}}) \mathbf{v}_l \rangle}{\lambda_l - \lambda_i} \left( \mathbf{v}_l \mathbf{v}_l^T + \mathbf{v}_i \mathbf{v}_i^T \right) \right) \quad (5)$$

with  $\mathcal{G}_i = \{j \mid \lambda_j = \lambda_i \pm \mathcal{O}(\varepsilon)\}$ . For details, see Lemma 3.3 from [10].

The third concept can be regarded as the most important tool for developing the proof machinery of linear exit time; see Lemmas 3.4 and 3.5 from [10] for details. Specifically, it can be summarized as the ‘Approximation Lemma’ for a linear dynamical system. Given some initialization of the radial vector  $\mathbf{u}_0$  and sufficiently small  $\varepsilon$ , we have for any iteration  $K$  that  $\mathbf{u}_K = \prod_{k=0}^{K-1} [\mathbf{A}_k + \varepsilon \mathbf{P}_k] \mathbf{u}_0$ , where  $\varepsilon \mathbf{P}_k = \mathbf{B}_k + \mathcal{O}(\varepsilon^2)$ ,  $\mathbf{B}_k = \mathcal{O}(\varepsilon)$  for  $\mathbf{x}_k \in \mathcal{B}_\varepsilon(\mathbf{x}^*)$ ,  $\{\mathbf{A}_k\}$  and  $\{\mathbf{B}_k\}$  are sequences of real symmetric matrices, and  $\mathbf{A}_k$ ’s are invertible.

When  $K\varepsilon \ll 1$  and  $\varepsilon < \|\mathbf{A}^{-1}\|_2^{-1} \|\mathbf{P}\|_2^{-1}$ , we have the condition

$$\begin{aligned} \|\mathbf{A}^{-1}\|_2^{-K} \left( 1 - K\varepsilon \frac{\|\mathbf{P}\|_2}{\|\mathbf{A}^{-1}\|_2^{-1}} - \mathcal{O}((K\varepsilon)^2) \right) \leq v_n \leq \dots \leq v_1 \leq \\ \|\mathbf{A}\|_2^K \left( 1 + K\varepsilon \frac{\|\mathbf{P}\|_2}{\|\mathbf{A}\|_2} + \mathcal{O}((K\varepsilon)^2) \right), \end{aligned}$$

where  $v_n \leq \dots \leq v_1$  are absolute values of the eigenvalues of matrix  $\prod_{k=0}^{K-1} [\mathbf{A}_k + \varepsilon \mathbf{P}_k]$  and  $\sup_{0 \leq k \leq K-1} \|\mathbf{A}_k\|_2 = \|\mathbf{A}\|_2$ ,  $\sup_{0 \leq k \leq K-1} \|\mathbf{A}_k^{-1}\|_2 = \|\mathbf{A}^{-1}\|_2$ ,  $\sup_{0 \leq k \leq K-1} \|\mathbf{P}_k\|_2 = \|\mathbf{P}\|_2$  for some matrices  $\mathbf{A}$  and  $\mathbf{P}$ . Hence,  $\mathbf{u}_K = \prod_{k=0}^{K-1} [\mathbf{A}_k + \varepsilon \mathbf{P}_k] \mathbf{u}_0$  can be expanded to first order in  $\varepsilon$  with the first-order approximation called  $\tilde{\mathbf{u}}_K$  and the trajectory generated by the sequence  $\{\tilde{\mathbf{u}}_k\}$  is termed  $\varepsilon$ -precision trajectory. Thus the gradient update  $\mathbf{x}_{K+1} = \mathbf{x}_K - \alpha \nabla f(\mathbf{x}_K)$  near  $\mathbf{x}^*$  can be written as  $\mathbf{u}_K = \prod_{k=0}^{K-1} [\mathbf{A}_k + \varepsilon \mathbf{P}_k] \mathbf{u}_0$  for  $\mathbf{u}_K = \mathbf{x}_K - \mathbf{x}^*$ ,  $\mathbf{A}_K = \mathbf{I} - \alpha \nabla^2 f(\mathbf{x}^*)$  and  $\varepsilon \mathbf{P}_K = -\frac{\alpha \|\mathbf{u}_K\|}{2} \mathbf{H}(\hat{\mathbf{u}}_K) + \mathcal{O}(\varepsilon^2)$ .

Fourth, from Lemma 3.6 of [10], the ‘minimal’  $\varepsilon$ -precision trajectory has the maximum exit time. More rigorously, let  $\mathcal{S}_\varepsilon = \left\{ \{\tilde{\mathbf{u}}_k^\tau\}_{k=1}^{K_{\text{exit}}^\tau} \mid \mathbf{u}_0 \right\}$  be the set of  $\tau$ -parametrized  $\varepsilon$ -precision trajectories generated by expanding  $\mathbf{u}_K$  to first order in  $\varepsilon$ , where  $\tau$  varies with variations in the perturbation sequence  $\{\mathbf{P}_k\}_{k=0}^K$ . Let  $K_{\text{exit}}^\tau$  be the exit time of the  $\tau$ -parametrized trajectory  $\{\tilde{\mathbf{u}}_k^\tau\}_{k=1}^{K_{\text{exit}}^\tau}$  from the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$ , where we have  $K_{\text{exit}}^\tau = \inf_{K \geq 1} \left\{ K \mid \|\tilde{\mathbf{u}}_K^\tau\|^2 > \varepsilon^2 \right\}$ . Let  $K^1$  be defined as

$$K^1 = \inf_{K \geq 1} \left\{ K \mid \inf_{\tau} \left\{ \|\tilde{\mathbf{u}}_K^\tau\|^2 \right\} > \varepsilon^2 \right\}. \quad (6)$$

Then the following inequality holds:

$$K^1 \geq \sup_{\tau} \left\{ K_{\text{exit}}^\tau \right\} = \sup_{\tau} \inf_{K \geq 1} \left\{ K \mid \|\tilde{\mathbf{u}}_K^\tau\|^2 > \varepsilon^2 \right\}.$$

Finally, the linear exit time theorem for the  $\varepsilon$ -precision trajectories (Theorem 3.2 in [10]) states that for gradient descent with  $\alpha = \frac{1}{L}$  where  $\varepsilon < \frac{2\beta}{M}$ , and some **minimum projection**

value  $\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \geq \Delta$  of the initial radial vector  $\mathbf{u}_0$  on  $\mathcal{E}_{US}$  with  $\mathbf{u}_0 = \varepsilon \sum_{i \in \mathcal{N}_S} \theta_i^s \mathbf{v}_i + \varepsilon \sum_{j \in \mathcal{N}_{US}} \theta_j^{us} \mathbf{v}_j$ , there exist  $\varepsilon$ -precision trajectories  $\{\tilde{\mathbf{u}}_K\}_{K=1}^{K_{exit}}$  with linear exit time. Moreover their exit time  $K_{exit}$  from  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  is approximately upper bounded as

$$K_{exit} < K^1 \lesssim \frac{\log \left( \left( 2 + \frac{\varepsilon M}{2L} \right) \log \left( \frac{2 + \frac{\varepsilon M}{2L}}{1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}} \right) \frac{2\delta}{\varepsilon M n} \right)}{2 \log \left( \frac{2 + \frac{\varepsilon M}{2L}}{1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}} \right)}. \quad (7)$$

In [10, Theorem 3.2], we provide a necessary initial condition for the linear exit time bound, which is

$$\Delta > \varepsilon \frac{MLn}{\delta(L + \beta)} = \mathcal{O}(\varepsilon),$$

where it is required that  $\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \geq \Delta$ . In this work we provide the sufficient boundary conditions for linear exit time  $\varepsilon$ -precision trajectories.

Before moving to the next section that details the sufficient conditions, we show that the  $\varepsilon$ -precision trajectory  $\{\tilde{\mathbf{u}}_K\}_{K=0}^{K_{exit}}$  generated by expanding the matrix product in the expression  $\mathbf{u}_K = \prod_{k=0}^{K-1} [\mathbf{A}_k + \varepsilon \mathbf{P}_k] \mathbf{u}_0$  to first order in  $\varepsilon$  has a very small relative error compared to the exact trajectory.

1) *Relative Error Margin in the  $\varepsilon$ -Precision Trajectory:*  
By the definition of the  $\varepsilon$ -precision trajectory, we have that

$$\tilde{\mathbf{u}}_K = \prod_{k=0}^{K-1} \mathbf{A}_k \mathbf{u}_0 + \varepsilon \sum_{r=0}^{K-1} \prod_{k=0}^r \mathbf{A}_k \mathbf{P}_r \prod_{k=r+1}^{K-1} \mathbf{A}_k \mathbf{u}_0, \quad (8)$$

which is obtained by expanding the matrix product  $\prod_{k=0}^{K-1} [\mathbf{A}_k + \varepsilon \mathbf{P}_k]$  to first order in  $\varepsilon$ . Now using the ‘‘Approximation Lemma’’ discussed above for  $K\varepsilon \ll 1$  and  $\varepsilon < \|\mathbf{A}^{-1}\|_2^{-1} \|\mathbf{P}\|_2^{-1}$  where  $\sup_{0 \leq k \leq K-1} \|\mathbf{A}_k\|_2 = \|\mathbf{A}\|_2$ ,  $\sup_{0 \leq k \leq K-1} \|\mathbf{A}_k^{-1}\|_2 = \|\mathbf{A}^{-1}\|_2$ ,  $\sup_{0 \leq k \leq K-1} \|\mathbf{P}_k\|_2 = \|\mathbf{P}\|_2$  for some matrices  $\mathbf{A}$  and  $\mathbf{P}$ , we get that:

$$\mathbf{u}_K = \prod_{k=0}^{K-1} [\mathbf{A}_k + \varepsilon \mathbf{P}_k] \mathbf{u}_0 \quad (9)$$

$$= \prod_{k=0}^{K-1} \mathbf{A}_k \mathbf{u}_0 + \varepsilon \sum_{r=0}^{K-1} \prod_{k=0}^r \mathbf{A}_k \mathbf{P}_r \prod_{k=r+1}^{K-1} \mathbf{A}_k \mathbf{u}_0 + \mathcal{O} \left( \|\mathbf{A}\|_2^K (K\varepsilon)^2 \frac{\|\mathbf{P}\|_2^2}{\|\mathbf{A}\|_2^2} \|\mathbf{u}_0\| \right) \quad (10)$$

$$= \tilde{\mathbf{u}}_K + \mathcal{O} \left( \|\mathbf{A}\|_2^K (K\varepsilon)^2 \varepsilon \right). \quad (11)$$

Next, from the proof of [10, Lemma 3.4] we recall that  $\mathbf{A}_k = \sum_{i \in \mathcal{N}_S} c_i^s(k) \mathbf{v}_i \mathbf{v}_i^T + \sum_{j \in \mathcal{N}_{US}} c_j^{us}(k) \mathbf{v}_j \mathbf{v}_j^T$  where  $c_i^s(k) = 1 - \alpha \lambda_i^s + \mathcal{O}(\varepsilon)$ ,  $c_j^{us}(k) = 1 - \alpha \lambda_j^{us} + \mathcal{O}(\varepsilon)$  and  $\lambda_i^s, \mathbf{v}_i$  and  $\lambda_j^{us}, \mathbf{v}_j$  are the eigenvalue-eigenvector pairs corresponding to the stable and unstable subspaces of  $\nabla^2 f(\mathbf{x}^*)$ , respectively. Also,  $\mathbf{u}_0 = \varepsilon \sum_{i \in \mathcal{N}_S} \theta_i^s \mathbf{v}_i + \varepsilon \sum_{j \in \mathcal{N}_{US}} \theta_j^{us} \mathbf{v}_j$  and for  $\alpha = \frac{1}{L}$  we have

the bounds  $1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \leq c_j^{us}(k) \leq 2 + \frac{\varepsilon M}{2L}$  and  $-\frac{\varepsilon M}{2L} \leq c_i^s(k) \leq 1 - \frac{\beta}{L} + \frac{\varepsilon M}{2L}$  (see [10, Lemma 3.4]). Hence we have that:

$$\|\mathbf{u}_K\| = \left\| \prod_{k=0}^{K-1} [\mathbf{A}_k + \varepsilon \mathbf{P}_k] \mathbf{u}_0 \right\| \quad (12)$$

$$\geq \left\| \prod_{k=0}^{K-1} \mathbf{A}_k \mathbf{u}_0 \right\| - \left\| \varepsilon \sum_{r=0}^{K-1} \prod_{k=0}^r \mathbf{A}_k \mathbf{P}_r \prod_{k=r+1}^{K-1} \mathbf{A}_k \mathbf{u}_0 \right\| - \mathcal{O} \left( \|\mathbf{A}\|_2^K (K\varepsilon)^2 \frac{\|\mathbf{P}\|_2^2}{\|\mathbf{A}\|_2^2} \|\mathbf{u}_0\| \right) \quad (13)$$

$$\geq \left\| \prod_{k=0}^{K-1} \mathbf{A}_k \mathbf{u}_0 \right\| - \mathcal{O} \left( \|\mathbf{A}\|_2^K (K\varepsilon) \frac{\|\mathbf{P}\|_2}{\|\mathbf{A}\|_2} \|\mathbf{u}_0\| \right) \quad (14)$$

$$= \left\| \left( \prod_{k=0}^{K-1} c_i^s(k) \right) \varepsilon \sum_{i \in \mathcal{N}_S} \theta_i^s \mathbf{v}_i + \left( \prod_{k=0}^{K-1} c_j^{us}(k) \right) \varepsilon \sum_{j \in \mathcal{N}_{US}} \theta_j^{us} \mathbf{v}_j \right\| - \mathcal{O} \left( \|\mathbf{A}\|_2^K (K\varepsilon) \varepsilon \right) \quad (15)$$

$$\geq \varepsilon \left( \inf \{c_j^{us}(k)\} \right)^K \sqrt{\left( \frac{\inf \{c_i^s(k)\}}{\inf \{c_j^{us}(k)\}} \right)^{2K} \sum_{i \in \mathcal{N}_S} (\theta_i^s)^2 + \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2} - \mathcal{O} \left( \|\mathbf{A}\|_2^K (K\varepsilon) \varepsilon \right) \quad (16)$$

$$\approx \varepsilon \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right)^K \sqrt{\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2} - \mathcal{O} \left( \|\mathbf{A}\|_2^K (K\varepsilon) \varepsilon \right), \quad (17)$$

where we used  $\inf \{c_j^{us}(k)\} = \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right)$ ,  $\inf \{c_i^s(k)\} = -\frac{\varepsilon M}{2L}$  and  $\varepsilon^{2K} \approx 0$  (here  $\varepsilon \ll 1$  since  $K\varepsilon \ll 1$ ). Simplifying (11) by using the substitution  $\|\mathbf{A}\|_2 = \sup \{\|\mathbf{A}_k\|_2\} = \sup \{c_j^{us}(k)\} = 2 + \frac{\varepsilon M}{2L}$  and taking norm yields

$$\|\mathbf{u}_K - \tilde{\mathbf{u}}_K\| = \mathcal{O} \left( \|\mathbf{A}\|_2^K (K\varepsilon)^2 \varepsilon \right) = \mathcal{O} \left( \left( 2 + \frac{\varepsilon M}{2L} \right)^K (K\varepsilon)^2 \varepsilon \right). \quad (18)$$

Finally, dividing (18) by (17) we get the following bound on the relative error:

$$\frac{\|\mathbf{u}_K - \tilde{\mathbf{u}}_K\|}{\|\mathbf{u}_K\|} \leq \frac{\mathcal{O} \left( \left( 2 + \frac{\varepsilon M}{2L} \right)^K (K\varepsilon)^2 \varepsilon \right)}{\varepsilon \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right)^K \sqrt{\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2} - \mathcal{O} \left( \|\mathbf{A}\|_2^K (K\varepsilon) \varepsilon \right)} \quad (19)$$

$$= \frac{\mathcal{O} \left( \frac{\left( 2 + \frac{\varepsilon M}{2L} \right)^K}{\left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right)^K} (K\varepsilon)^2 \right)}{\sqrt{\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2} - \mathcal{O} \left( \frac{\left( 2 + \frac{\varepsilon M}{2L} \right)^K}{\left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right)^K} (K\varepsilon) \right)} \quad (20)$$

$$\leq \frac{\mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}}\left(\log\left(\frac{1}{\varepsilon}\right)\varepsilon\right)^2\right)}{\sqrt{\sum_{j \in \mathcal{N}_{US}}(\theta_j^{us})^2 - \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}}\left(\log\left(\frac{1}{\varepsilon}\right)\varepsilon\right)^2\right)}}, \quad (21)$$

where we have substituted the upper bound on  $K_{exit}$  from (7) into  $K$ . Now, if  $\sqrt{\sum_{j \in \mathcal{N}_{US}}(\theta_j^{us})^2} > \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}}\left(\log\left(\frac{1}{\varepsilon}\right)\varepsilon\right)^2\right)$  then the relative error is of the order  $\mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}}\left(\log\left(\frac{1}{\varepsilon}\right)\varepsilon\right)^2\right)$ , which goes to 0 as  $\varepsilon \rightarrow 0$ .

### B. Sufficient Conditions for Linear Exit Time

Our first theorem states that the first order approximation of any gradient descent trajectory starting from an  $\varepsilon$  neighborhood of any strict saddle point  $\mathbf{x}^*$  will escape this neighborhood in linear time, i.e.,  $\mathcal{O}(\log(\varepsilon^{-1}))$ , provided the projection value of its initialization on the unstable subspace of  $\nabla^2 f(\mathbf{x}^*)$  is lower bounded.

**Theorem 1.** *The  $\varepsilon$ -precision trajectory  $\{\tilde{\mathbf{u}}_K\}_{K=0}^{K_{exit}}$  generated by the gradient descent method for step size  $\alpha = \frac{1}{L}$  on any function satisfying Assumptions A1–A4 has linear exit time (7) from the strict saddle neighborhood  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  provided the projection value of the initialization  $\mathbf{u}_0$  onto the unstable subspace  $\mathcal{E}_{US}$  of the Hessian  $\nabla^2 f(\mathbf{x}^*)$ , given by  $\sum_{j \in \mathcal{N}_{US}}(\theta_j^{us})^2$ , is lower bounded as:*

$$\sum_{j \in \mathcal{N}_{US}}(\theta_j^{us})^2 \gtrsim \frac{\left(2 + \frac{\varepsilon M}{2L}\right) \left(\frac{2\delta \sqrt{\mu} \log\left(1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}\right)}{Mn}\right)}{\frac{1}{a} \log\left(\frac{2\delta \left(2 + \frac{\varepsilon M}{2L}\right) \log\left(1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}\right) \log\left(\frac{2 + \frac{\varepsilon M}{2L}}{1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}}\right)}{\varepsilon M n \log\left(2 + \frac{\varepsilon M}{2L}\right)}\right) + 1}, \quad (22)$$

where  $\sqrt[4]{\mu} = \frac{Mn \log\left(2 + \frac{\varepsilon M}{2L}\right)}{2\delta \left(2 + \frac{\varepsilon M}{2L}\right) \log\left(1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}\right) \log\left(\frac{2 + \frac{\varepsilon M}{2L}}{1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}}\right)}$ , the parameter  $a = \frac{\log\left(2 + \frac{\varepsilon M}{2L}\right)}{\log\left(2 + \frac{\varepsilon M}{2L}\right) - \log\left(1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}\right)}$  and we require that:

$$\varepsilon < \min \left\{ \inf_{\|\mathbf{u}\|=1} \left( \limsup_{j \rightarrow \infty} \sqrt[j]{r_j(\mathbf{u})} \right)^{-1}, \frac{2L\delta}{M(2Ln^2 - \delta)} + \mathcal{O}(\varepsilon^2), \frac{2\beta}{M} \right\}, \quad (23)$$

where we have  $r_j(\mathbf{u}) = \left\| \left( \frac{d^j}{dw^j} \nabla^2 f(\mathbf{x}^* + w\mathbf{u}) \right) \Big|_{w=0} \right\|_2$ , the vector  $\mathbf{u}_0 = \varepsilon \sum_{i \in \mathcal{N}_S} \theta_i^s \mathbf{v}_i + \varepsilon \sum_{j \in \mathcal{N}_{US}} \theta_j^{us} \mathbf{v}_j$  and  $\mathbf{v}_i, \mathbf{v}_j$  are the eigenvectors of the Hessian  $\nabla^2 f(\mathbf{x}^*)$  and  $\delta$  is as in Proposition 2

In terms of order notation, we require the following lower bound on the projection  $\sum_{j \in \mathcal{N}_{US}}(\theta_j^{us})^2$ :

$$\sum_{j \in \mathcal{N}_{US}}(\theta_j^{us})^2 \gtrsim \Theta\left(\frac{1}{\log(\varepsilon^{-1})}\right). \quad (24)$$

The proof of this theorem is provided in Appendix A. Recall from (21) that for the relative error in the  $\varepsilon$ -precision trajectory to be bounded, we require that  $\sqrt{\sum_{j \in \mathcal{N}_{US}}(\theta_j^{us})^2} > \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}}\left(\log\left(\frac{1}{\varepsilon}\right)\varepsilon\right)^2\right)$ . However, this condition is already satisfied by the sufficient condition  $\sum_{j \in \mathcal{N}_{US}}(\theta_j^{us})^2 \gtrsim \Theta\left(\frac{1}{\log(\varepsilon^{-1})}\right)$  in terms of order since  $\Theta\left(\sqrt{\frac{1}{\log(\varepsilon^{-1})}}\right) > \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}}\left(\log\left(\frac{1}{\varepsilon}\right)\varepsilon\right)^2\right)$  as  $\varepsilon \rightarrow 0$ .

The above result can be interpreted as follows: for any sufficiently small  $\varepsilon$  bounded from (23) if a gradient descent trajectory at the surface of any saddle neighborhood  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  has a projection value of order  $\Theta\left(\frac{1}{\log(\varepsilon^{-1})}\right)$  on the unstable subspace of  $\nabla^2 f(\mathbf{x}^*)$ , then this trajectory is guaranteed to exit the saddle neighborhood in linear time. This result is crucial since it furthers the findings of the state of the art [45] where a non-zero projection value guarantees almost sure escape from the saddle point but does not provide any insights into whether a non-zero projection value could lead to fast escaping trajectories, something which Theorem 1 establishes rigorously. Moreover the projection value bound in Theorem 1 is insightful in the sense that it illustrates the dependency to the quantities like condition number, problem dimension, spectral gap, etc. Since this result ensures that fast escaping gradient trajectories are indeed dense with respect to random initialization on the surface of the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$ , we can safely say that fast escaping trajectories for gradient descent method from small saddle neighborhoods of Morse functions will be a generic phenomenon. In case if the sufficient condition is not satisfied, one can perform a single step perturbation to land on a point which satisfies this condition. Then reverting back to gradient descent update, linear exit time from the saddle neighborhood will be guaranteed. This particular idea will serve as a basis for the development of a single step perturbation based gradient descent method for escaping saddle points faster.

**Remark 2.** Note that there are several key differences between the analytical approach of this work and the one utilized in [22]. First, Lemma 17 in [22] works with gradient trajectories on a local quadratic approximation of the nonconvex function, which does not require an approximation of the Hessian. In contrast, Theorem 1 utilizes an explicit approximation of the Hessian—as opposed to a quadratic approximation of the function—in order to approximate the gradient trajectories around the saddle point. This approach later allows us to establish that the linear exit time guarantees for the approximate trajectories also hold for the exact trajectories (see equation (337) in Appendix H). Second, our analytical approach helps develop a theory that relates the linear exit time of a gradient trajectory from a saddle neighborhood to the angle between the initial radial vector of the trajectory and the unstable subspace of  $\nabla^2 f(\mathbf{x}^*)$ , while such geometric

connections are lacking in [22]. In fact, to the best of our knowledge, such analysis that helps study the relationship between fast escaping trajectories and their initial angles with the unstable subspace of  $\nabla^2 f(\mathbf{x}^*)$  has not been carried out in any of the prior works. Additionally, while the guarantees in our work are deterministic in nature, Lemma 17 in [22] is probabilistic in nature. Finally, and perhaps most importantly, our analytical approach leads to estimates of the exit time / escape rate and boundary conditions that are sharp in the sense that they depend on the spectral gap, condition number and problem dimension also, something that is novel to the existing nonconvex optimization literature.

We now move to the next section which provides a rate analysis in regions outside the small saddle neighborhood  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  where the local analyticity property no longer exists and we are only left with the class of  $\mathcal{C}^2$  gradient and Hessian Lipschitz, Morse functions, i.e., functions satisfying Assumptions A2–A4.

#### IV. SEQUENTIAL MONOTONICITY

The first theorem in this section establishes a monotonicity property of the gradient descent trajectories in a strict saddle neighborhood. This property is termed as “sequential monotonicity” which implies that within some neighborhood of the strict saddle point  $\mathbf{x}^*$  any gradient trajectory, which does not converge to  $\mathbf{x}^*$ , first continuously contracts towards  $\mathbf{x}^*$  up to some point and from there onward expands continuously away from  $\mathbf{x}^*$  until it escapes this neighborhood.

**Theorem 2.** *On the class of  $\mathcal{C}^2$  gradient and Hessian Lipschitz, Morse functions, if a gradient trajectory with respect to some stationary point  $\mathbf{x}^*$  has non-contractive dynamics at any iteration  $k = K$ , then it has expansive dynamics for all iterations  $k > K$  provided  $\|\mathbf{x}_k - \mathbf{x}^*\|$  is bounded above by some  $\xi > 0$  where  $\{\mathbf{x}_k\}$  is the sequence that generates the gradient trajectory. This property of the sequence of radial distances  $\{\|\mathbf{x}_k - \mathbf{x}^*\|\}$  can be termed as the sequential monotonicity.*

Moreover, in the case of  $\mathbf{x}^*$  being a strict saddle point, we have for gradient trajectories with step size  $\alpha = \frac{1}{L}$  that

$\xi < \frac{1}{\zeta M} \frac{\left( (1 + \frac{\beta}{L})^2 + \frac{1}{4(1 + \frac{\beta}{L})^2} - \frac{5}{4} \right)}{6}$  for some  $\zeta > 2$ . Specifically, consider the tuple  $(\mathbf{x}, \mathbf{x}^+, \mathbf{x}^{++})$  that is equivalent to the tuple  $(\mathbf{x}_k, \mathbf{x}_{k+1}, \mathbf{x}_{k+2})$  for any  $k$ . Let  $\|\mathbf{x}^+ - \mathbf{x}^*\| > \|\mathbf{x} - \mathbf{x}^*\|$  and  $\|\mathbf{x} - \mathbf{x}^*\| < \xi$ . Then the following holds:

$$a. \quad \|\mathbf{x}^{++} - \mathbf{x}^*\| \geq \bar{\rho}(\mathbf{x}) \|\mathbf{x}^+ - \mathbf{x}^*\| - \sigma(\mathbf{x}), \quad \text{and} \quad (25)$$

$$b. \quad \|\mathbf{x}^{++} - \mathbf{x}^*\| > \|\mathbf{x}^+ - \mathbf{x}^*\|, \quad (26)$$

where  $\sigma(\mathbf{x}) = \mathcal{O}(\|\mathbf{x} - \mathbf{x}^*\|^2)$  and  $\bar{\rho}(\mathbf{x}) > 1 + \frac{\left( (1 + \frac{\beta}{L})^2 + \frac{1}{4(1 + \frac{\beta}{L})^2} - \frac{5}{4} \right)}{12}$ .

The proof of this theorem is given in Appendix B.

**Remark 3.** The upper bound on  $\xi$  given by the quantity

$\frac{1}{\zeta M} \frac{\left( (1 + \frac{\beta}{L})^2 + \frac{1}{4(1 + \frac{\beta}{L})^2} - \frac{5}{4} \right)}{6}$  for  $\zeta > 2$  is always positive and is

equal to 0 only when  $\beta = 0$ . Moreover, for Morse functions that are well conditioned at their stationary points, i.e.,  $0 < \frac{\beta}{L} < 1$ , this quantity can be treated as a constant. Moreover this bound on  $\xi$  also makes sure that there cannot be any other critical

point within a radius of  $\frac{1}{\zeta M} \frac{\left( (1 + \frac{\beta}{L})^2 + \frac{1}{4(1 + \frac{\beta}{L})^2} - \frac{5}{4} \right)}{6}$  for  $\zeta > 2$  from  $\mathbf{x}^*$ . If another stationary point did exist within this radius of  $\mathbf{x}^*$  say  $\mathbf{x}_1^*$  then  $\|\nabla f(\mathbf{x}_1^*)\| \geq \beta \|\mathbf{x}^* - \mathbf{x}_1^*\| > 0$  from (158) which contradicts the fact that  $\mathbf{x}_1^*$  is a critical point of  $f$ . This seemingly trivial result will be of utility in Proposition 3 where we define separation between critical points.

In words, Theorem 2 states that within any  $\xi$  neighborhood

of the saddle point where  $\xi < \frac{1}{\zeta M} \frac{\left( (1 + \frac{\beta}{L})^2 + \frac{1}{4(1 + \frac{\beta}{L})^2} - \frac{5}{4} \right)}{6}$  for some  $\zeta > 2$ , every gradient descent trajectory first contracts continuously towards  $\mathbf{x}^*$ . The first iteration after the end of contraction phase is either marked by expansion or preservation of radial distance, i.e., no expansion or contraction. In both cases the trajectory from here onward expands continuously till it exits  $\mathcal{B}_\xi(\mathbf{x}^*)$  where in the latter case it is assumed that the trajectory didn't already contract to  $\mathbf{x}^*$ . Furthermore expansion happens at an almost geometric rate as evident from part (a.) of the theorem which can be leveraged to obtain linear rate for the expansion phase of trajectories inside  $\mathcal{B}_\xi(\mathbf{x}^*)$ .

So far we have been able to develop a machinery that will help us in providing linear rate of expansion inside  $\mathcal{B}_\xi(\mathbf{x}^*)$ . It remains to develop a proof technique which can generate linear rates of contraction inside  $\mathcal{B}_\xi(\mathbf{x}^*)$ . In order to do so we introduce certain terms that are required for better understanding the contraction and expansion dynamics of the trajectory. In this regard, let  $K_{\text{exit}}$  be the first exit time of the gradient descent trajectory from the ball  $\mathcal{B}_\xi(\mathbf{x}^*)$ , where we assume that the trajectory starts at the boundary of the ball  $\mathcal{B}_\xi(\mathbf{x}^*)$ , i.e.,  $\mathbf{x}_0 \in \mathcal{B}_\xi(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*)$  and  $\xi$  is bounded from Theorem 2. Next, for any  $\varepsilon < \xi$ , let  $\mathcal{B}_\xi(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*)$  be a compact shell centered at  $\mathbf{x}^*$ . Let  $k = K_c$  be the last iteration for which the gradient trajectory has contractive dynamics inside the shell and  $k = K_e$  be the first iteration for which the gradient trajectory has expansive dynamics inside the shell. Note that  $K_c$  and  $K_e$  are equal iff either the trajectory starts expanding before reaching the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  or the trajectory just touches the surface of the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  and then expands from there onward.

The next lemma provides further insights into the behavior of function sequence  $\{f(\mathbf{x}_k)\}_{k=0}^{K_c}$  associated with iterate sequence  $\{\mathbf{x}_k\}_{k=0}^{K_c}$  where  $0 \leq k \leq K_c$  are the iterations with contraction dynamics.

**Lemma 1.** *On the class of  $\mathcal{C}^2$  gradient and Hessian Lipschitz, Morse functions, the function sequence  $\{f(\mathbf{x}_k)\}_{k=0}^{K_c}$  associated with iterate sequence  $\{\mathbf{x}_k\}_{k=0}^{K_c}$  for  $\|\mathbf{x}_{K_c} - \mathbf{x}^*\| < \frac{3\beta^2}{4ML}$  and  $K_c < K_e$  satisfies the Polyak–Łojasiewicz condition [15] where for any  $0 \leq k \leq K_c$  we have that:*

$$0 < f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{L}{2\beta^2} \|\nabla f(\mathbf{x}_k)\|^2.$$

The proof of this lemma is given in Appendix C. Using this lemma, it can be readily checked that the function

sequence  $\{f(\mathbf{x}_k)\}_{k=0}^{K_c}$  is strongly monotonic in the contraction phase of the trajectory. Formally, for  $0 \leq k \leq K_c$  using Lemma 1 and the gradient Lipschitz condition we will have the inequality  $f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq \left(1 - \frac{\beta^2}{L^2}\right) (f(\mathbf{x}_k) - f(\mathbf{x}^*))$ . Therefore linear rates for the contraction phase of trajectory can be recovered using this result. It should however be noted that the function sequence  $\{f(\mathbf{x}_k)\}_{k=K_c}^{\hat{K}_{exit}}$  associated with the expansion phase of the trajectory does not satisfy the Polyak–Łojasiewicz condition from Lemma 1 and therefore we require Theorem 2 to generate linear rates of expansion for the trajectory in its expansion phase (see discussion within the proof of Lemma 1 for details).

Before stating the final theorem of this section we introduce the term ‘sojourn time’. It is defined as the time the trajectory spends inside the shell  $\tilde{\mathcal{B}}_\xi(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*)$  before leaving this region. The sojourn time will be the sum of contraction time (derived using Lemma 1) and the expansion time (derived using Theorem 1) for any trajectory inside the shell  $\tilde{\mathcal{B}}_\xi(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*)$ . We are now ready to state the theorem.

**Theorem 3.** *The sojourn time  $K_{shell}$  for a gradient trajectory inside the compact shell  $\tilde{\mathcal{B}}_\xi(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*)$  for a strict saddle point  $\mathbf{x}^*$  of any  $\mathcal{C}^2$  gradient and Hessian Lipschitz, Morse function is bounded by*

$$K_{shell} \leq \frac{\log\left(\frac{L}{2}\xi^2\right) - \log\left(\frac{\beta^2}{2L}\varepsilon^2 - \frac{2M}{3}\varepsilon^3\right)}{\log\left(1 - \frac{\beta^2}{L^2}\right)^{-1}} + \frac{\log(\xi) - \log(\varepsilon)}{\log\left(\frac{\inf\{\tilde{\rho}(\mathbf{x}_{k-2})\}}{1+M\xi}\right)} + 3, \quad (27)$$

where the sojourn time  $K_{shell} = \hat{K}_{exit} + K_c - K_e$  with  $K_c \leq \frac{\log\left(\frac{L}{2}\xi^2\right) - \log\left(\frac{\beta^2}{2L}\varepsilon^2 - \frac{2M}{3}\varepsilon^3\right)}{\log\left(1 - \frac{\beta^2}{L^2}\right)^{-1}} + 1$ ,  $\hat{K}_{exit} - K_e \leq$

$\frac{\log(\xi) - \log(\varepsilon)}{\log\left(\frac{\inf\{\tilde{\rho}(\mathbf{x}_{k-2})\}}{1+M\xi}\right)} + 2$ , and infimum in the term  $\inf\{\tilde{\rho}(\mathbf{x}_{k-2})\}$

is taken over the indices  $K_e + 2 \leq k \leq \hat{K}_{exit}$ . Further,  $\hat{K}_{exit} - K_e$  is the time for which the gradient trajectory has expansive dynamics inside the shell  $\tilde{\mathcal{B}}_\xi(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*)$ , and  $K_c$  is the time for which the gradient trajectory has contractive dynamics inside the shell  $\tilde{\mathcal{B}}_\xi(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*)$ . Also,

$$\xi \leq \frac{1}{\zeta M} \frac{\left((1+\frac{\beta}{L})^2 + \frac{1}{4(1+\frac{\beta}{L})^2} - \frac{5}{4}\right)}{6} \quad \text{with } \zeta > 2, \quad \varepsilon < \frac{3\beta^2}{4ML} \quad \text{and} \quad \frac{\inf\{\tilde{\rho}(\mathbf{x}_{k-2})\}}{1+M\xi} > 1.$$

In terms of order notation,  $K_{shell}$  has the following rate:

$$K_{shell} = \mathcal{O}\left(\log\left(\frac{\xi}{\varepsilon}\right)\right) + \mathcal{O}(1), \quad (28)$$

where  $K_c = \mathcal{O}\left(\log\left(\frac{\xi}{\varepsilon}\right)\right)$  and  $\hat{K}_{exit} - K_e = \mathcal{O}\left(\log\left(\frac{\xi}{\varepsilon}\right)\right) + \mathcal{O}(1)$ .

The proof of this theorem is given in Appendix D. Theorem 3 provides an upper bound on the travel time of the trajectory inside the shell  $\tilde{\mathcal{B}}_\xi(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*)$ . The upper bound is linear since it is the sum of rates in the contraction and expansion phase of the trajectory and both these rates are linear by virtue of Lemma 1 and Theorem 2 respectively. In contrast to the linear exit time bound (7) which only holds for very small values of  $\varepsilon$  from Theorem 1, this rate holds for much bigger  $\xi$  neighborhoods and at the same time does not require the function to be analytic. The power of Theorem 3 will become more apparent once we develop a fast algorithm for escaping strict saddle points of Morse functions. This theorem will facilitate in keeping the algorithm very close to the gradient descent method since it proves that any escaping gradient descent trajectory from some small ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  will leave a larger ball  $\mathcal{B}_\xi(\mathbf{x}^*)$  at a linear rate irrespective of its exit point on  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$ . Hence any algorithm, which exits some small ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  using the gradient descent update, can keep on performing gradient descent updates so as to have linear rate of escape from a larger ball  $\mathcal{B}_\xi(\mathbf{x}^*)$ .

## V. ADDITIONAL LEMMAS

We now discuss some additional yet important lemmas instrumental in analysing the gradient trajectory/approximate trajectory behavior in saddle neighborhoods of any strict saddle point  $\mathbf{x}^*$ . Briefly, the first-order perturbations are ignored only in Lemmas 2 and 3 whereas an exact analysis is carried out in the subsequent lemmas. Moreover, the error introduced by the zeroth-order approximations (see (195) in Appendix E) in Lemmas 2 and 3 is sufficiently small by virtue of the initial conditions. Also, Lemmas 2 and 3 are qualitative in nature and therefore not used in proving the convergence of Algorithm 1 proposed in this work. For rest of the lemmas in this section (Lemmas 4-6), exact analysis is carried out since they will play a crucial role in the convergence analysis of the proposed algorithm (Algorithm 1). We also note that Assumptions A2–A4 hold for all the lemmas in this section where Lemmas 2, 3 use the extra assumption of local analyticity around the strict saddle point. The proofs of the lemmas in this section are given in Appendix E.

**Lemma 2.** *The gradient trajectories  $\{\mathbf{u}_k\}_{k=0}^{\hat{K}_{exit}}$  inside the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  with linear exit time and satisfying the initial condition  $\sqrt{\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2} > \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right)\varepsilon\right)\right)$  approximately exhibit hyperbolic behavior in the sense that they first move exponentially fast towards the saddle point  $\mathbf{x}^*$ , reach some point of minimum distance from  $\mathbf{x}^*$ , denoted by  $\mathbf{x}_{critical}$ , and then move exponentially fast away from  $\mathbf{x}^*$  for some iterations so as to escape the saddle region. For the case when  $\mathbf{x}_{critical} \rightarrow \mathbf{x}^*$ , their first-order approximation or the  $\varepsilon$ -precision trajectories can take very large time to exit the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$ , i.e.,  $K^1 \rightarrow \infty$  where  $K^1$  is defined in (6). When  $\mathbf{x}_{critical} = \mathbf{x}^*$ , we have  $K_{exit} = K^1 = \infty$ , which implies that the  $\varepsilon$ -precision trajectories and hence the gradient trajectory can never escape the saddle region.*

**Lemma 3.** *In the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$ , gradient descent trajectories with linear exit time and satisfying the initial condi-*

tion  $\sqrt{\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2} > \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)\right)$  approximately<sup>4</sup> never curve around the stationary point  $\mathbf{x}^*$ . Moreover, all the linear exit time gradient descent trajectories lie approximately inside some orthant of the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$ , i.e., the entry and exit point approximately subtend an angle less than or equal to  $\frac{\pi}{2}$  at the point  $\mathbf{x}^*$ .

**Lemma 4.** The function value at the exit point on the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  for any gradient descent trajectory is strictly less than  $f(\mathbf{x}^*)$  provided  $\varepsilon$  is sufficiently small.

**Lemma 5.** For any  $\varepsilon \ll 2^{-\frac{2}{\kappa^2}}$  where  $\kappa = \frac{\beta}{L}$ , a gradient trajectory having exited the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  can never re-enter this ball.

**Lemma 6.** The gradient descent trajectories exiting the ball  $\mathcal{B}_\xi(\mathbf{x}^*)$ , where  $\xi$  is defined in Theorem 3, can never re-enter this ball provided the gradient magnitudes outside the ball  $\mathcal{B}_\xi(\mathbf{x}^*)$  are sufficiently large with  $\|\nabla f(\mathbf{x})\| \geq \gamma > \frac{1}{\sqrt{2}} L \xi$ .

Note that Lemma 4 is used in our analysis for establishing that the function sequence  $\{f(\mathbf{x}_k)\}_{k=K_e}^{K_{exit}}$  associated with the expansion phase of the trajectory inside  $\mathcal{B}_\xi(\mathbf{x}^*)$  does not satisfy the Polyak–Łojasiewicz condition from Lemma 1. Lemmas 5 and 6 are termed as the “no-return conditions” to  $\varepsilon$  and  $\xi$  radius saddle neighborhoods respectively. Choosing  $\varepsilon$  from Lemma 5 will guarantee that any gradient trajectory can visit the saddle neighborhood  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  at most once. In particular, if the function satisfies the condition of large gradient magnitudes for certain  $\xi$  from Lemma 6 then any gradient trajectory can visit the saddle neighborhood  $\mathcal{B}_\xi(\mathbf{x}^*)$  at most once, and such a function is called a *well-structured* function (see discussion after Proposition 5 for details).

## VI. PROPOSED ALGORITHM

Since we have established the preliminaries on our unstable projection value and the sequential monotonicity property, we propose a method called the Curvature Conditioned Regularized Gradient Descent (CCRGD) (Algorithm 1) that can guarantee escaping saddle points in approximately linear time for Morse functions, by virtue of Theorems 1 and 3, and that is also guaranteed to converge to a local minimum.

We first establish that the proposed algorithm escapes any saddle point of a function satisfying Assumptions A1–A4 at a linear rate and the function values generated by the algorithm decrease monotonically.

**Lemma 7.** The trajectory generated by the CCRGD algorithm (Algorithm 1) in some  $\varepsilon$  neighborhood  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  of any strict saddle point  $\mathbf{x}^*$  of a function satisfying Assumptions A1–A4, where  $\varepsilon$  is bounded by Theorem 1, exits  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  in approximately linear time<sup>5</sup> with the exit time bounded by (7).

The proof of this lemma is given in Appendix F.

<sup>4</sup>When we say this condition holds approximately, we mean that it holds for a first-order approximation of the gradient descent trajectory (see the proof of Lemma 3 for further details).

<sup>5</sup>The term “approximately linear time” implies  $K_{exit} \leq \mathcal{O}(\log(\varepsilon^{-1})) + g(\varepsilon)$  where  $g(\cdot)$  is some absolutely continuous positive function such that  $g(\varepsilon) \rightarrow 0$  as  $\varepsilon \rightarrow 0$ . See the exact expression for  $g(\cdot)$  in (340) within Appendix H.

### Algorithm 1 Curvature Conditioned Regularized Gradient Descent (CCRGD)

**Initialize**  $\{\mathbf{x}_0, \mathbf{y}_0, \mathbf{y}_1\}$  to  $\mathbf{0}$ , a radius  $\varepsilon$  bounded by Theorem 1, constants  $L, M, \beta, \delta$ , minimum unstable projection value  $P_{min}(\varepsilon)$  from the lower bound in (74), condition flag  $\Xi = 0$ ,  $\kappa = \frac{\beta}{L}$ , and step size  $\alpha = \frac{1}{L}$

- 1: **for**  $k = 0, 1, \dots$  **do**
- 2:   Get  $\nabla f(\mathbf{x}_k)$  from first-order oracle
- 3:   **if**  $\|\nabla f(\mathbf{x}_k)\| > L\varepsilon$  **then**
- 4:      $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)$
- 5:     **if**  $\Xi = 1$  **then**
- 6:        $\Xi \leftarrow 0$  ▷ Update condition flag
- 7:   **else if**  $\|\nabla f(\mathbf{x}_k)\| \leq L\varepsilon$  **and**  $\Xi = 1$  **then**
- 8:      $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)$
- 9:   **else if**  $\|\nabla f(\mathbf{x}_k)\| \leq L\varepsilon$  **and**  $\Xi = 0$  **then**
- 10:      $\Xi \leftarrow 1$  ▷ Update condition flag
- 11:      $\mathbf{y}_0 \leftarrow \mathbf{x}_k$
- 12:      $\mathbf{y}_1 \leftarrow \mathbf{y}_0 - \alpha \nabla f(\mathbf{y}_0)$
- 13:      $\mathbf{V}_1 \leftarrow \langle \mathbf{y}_1 - \mathbf{y}_0, \mathbf{y}_1 - \mathbf{y}_0 \rangle$
- 14:      $\mathbf{V}_2 \leftarrow \alpha \langle \mathbf{y}_1 - \mathbf{y}_0, \nabla f(\mathbf{y}_1) - \nabla f(\mathbf{y}_0) \rangle$
- 15:     **if**  $\frac{4\varepsilon^2}{27\kappa^2} < \mathbf{V}_1 - \mathbf{V}_2 < \left(\frac{50P_{min}(\varepsilon)+4}{27}\right) \frac{\varepsilon^2}{\kappa^2}$  **then** ▷ Curvature Check Condition
- 16:       Get  $\mathbf{H} \leftarrow \alpha \nabla^2 f(\mathbf{x}_k)$  from second-order oracle
- 17:        $\mathbf{x}_{k+1} \in \arg \min_{\|\mathbf{x} - \mathbf{x}_k\| = \frac{\|\nabla f(\mathbf{x}_k)\|}{\beta}} \left( \frac{1}{2} (\mathbf{x} - \mathbf{x}_k)^T \mathbf{H} (\mathbf{x} - \mathbf{x}_k) \right)$  ▷ Solve constrained eigenvalue problem
- 18:     **else if**  $0 < \mathbf{V}_1 - \mathbf{V}_2 \leq \frac{4\varepsilon^2}{27\kappa^2}$  **then** ▷ Curvature Check Condition
- 19:       Get  $\mathbf{H} \leftarrow \alpha \nabla^2 f(\mathbf{x}_k)$  from second-order oracle
- 20:       **if**  $\lambda_{min}(\mathbf{H}) < 0$  **then**
- 21:          $\mathbf{x}_{k+1} \in \arg \min_{\|\mathbf{x} - \mathbf{x}_k\| = \frac{\|\nabla f(\mathbf{x}_k)\|}{\beta}} \left( \frac{1}{2} (\mathbf{x} - \mathbf{x}_k)^T \mathbf{H} (\mathbf{x} - \mathbf{x}_k) \right)$  ▷ Solve constrained eigenvalue problem
- 22:       **else** break from the **for** loop
- 23:     **else** continue

**Output:** Second-order stationary solution  $\mathbf{x}_k$

**Lemma 8.** The function value sequence  $\{f(\mathbf{x}_k)\}$  generated by the CCRGD algorithm (Algorithm 1) in some  $\varepsilon$  neighborhood  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  of any strict saddle point  $\mathbf{x}^*$  of a function satisfying Assumptions A1–A4 where  $\varepsilon$  is bounded by Theorem 1 decreases monotonically.

The proof of this lemma is given in Appendix F.

**Remark 4.** Note that the second-order step after the Curvature Check Condition 15 of Algorithm 1 can be replaced by Perturbed Gradient Descent (GD) type of update from [24] since one-step noise injection is known to escape saddle points. However there is no guarantee that such replacement will provably generate trajectories that exit the saddle neighborhood in linear time. The best one can achieve with a Perturbed GD type of update is fast escape with high probability. Since the focus of this work is to develop a deterministic algorithm that generates trajectories with linear

exit time, we refrain from analyzing the class of Perturbed GD type methods, which are designed for saddle escape but not necessarily with a linear rate.

## VII. CONVERGENCE RATES TO A MINIMUM

Now that we have developed an algorithm that escapes saddle neighborhoods in approximately linear time, our goal is to show that it (Algorithm 1) converges to some local minimum and obtain its rate of convergence.

### A. Asymptotic convergence

First, we show that the iterate sequence  $\{\mathbf{x}_k\}$  generated by Algorithm 1 avoids strict saddle points.

**Lemma 9.** *The iterate sequence  $\{\mathbf{x}_k\}$  generated by Algorithm 1 or any of its subsequence on the class of  $\mathcal{C}^2$  gradient and Hessian Lipschitz, Morse functions does not converge to a strict saddle point.*

The proof of this lemma is given in Appendix G.

**Remark 5.** *Note that in the proposed CCRGD algorithm (Algorithm 1) we always implicitly assume that the initialization  $\mathbf{x}_0$  does not belong to the set of strict saddle points of the function  $f(\cdot)$ . In particular since  $f(\cdot)$  is a Morse function, its critical points are isolated [42] and hence are countable. Since the probability that the initialization point, when chosen randomly from  $\mathbb{R}^n$ , falls into this countable set is zero, our assumption is justified.*

The next two lemmas establish that the function sequence  $\{f(\mathbf{x}_k)\}$  converges to a limit within a compact set in  $\mathbb{R}^n$  and the trajectory of  $\{\mathbf{x}_k\}$  generated by Algorithm 1 encounters at most finitely many saddle points. These lemmas will also be instrumental in providing global rates of convergence.

**Lemma 10.** *The sequence  $\{f(\mathbf{x}_k)\}$ , where  $\{\mathbf{x}_k\}$  is the iterate sequence generated by Algorithm 1 on the class of  $\mathcal{C}^2$  gradient and Hessian Lipschitz, coercive functions, converges to a limit value while the iterates  $\mathbf{x}_k$  stay in a compact set in  $\mathbb{R}^n$ .*

The proof of this lemma is given in Appendix G.

**Lemma 11.** *The iterate sequence  $\{\mathbf{x}_k\}$  generated by Algorithm 1 on the class of  $\mathcal{C}^2$  gradient and Hessian Lipschitz, coercive Morse functions stays within a compact subset of  $\mathbb{R}^n$  and encounters at most finitely many saddle points.*

The proof of this lemma is given in Appendix G.

It is needless to state that finite critical points imply isolated critical points<sup>6</sup>. The condition of isolated critical points however holds in general for the class of Morse functions. We now state the Global Convergence Theorem from [39] which is instrumental in establishing the asymptotic convergence of Algorithm 1 to a local minimum. Its proof is detailed in section 7.7 of [39] so we do not present its proof here and directly use this theorem.

<sup>6</sup>The condition of isolated critical points means that there is some separation between the critical points.

**Theorem 4 (Global Convergence Theorem [39]).** *Let  $\mathbf{A}$  be an algorithm on a vector space  $X$ , and suppose that, given  $\mathbf{x}_0$  the sequence  $\{\mathbf{x}_k\}_{k=0}^{\infty}$  is generated satisfying  $\mathbf{x}_{k+1} \in \mathbf{A}(\mathbf{x}_k)$ . Let a solution set  $S \subset X$  be given, and suppose:*

- 1) *all points  $\mathbf{x}_k$  are contained in a compact set  $D \subset X$ ,*
- 2) *there is a continuous function  $Z$  on  $X$  such that:*
  - *if  $\mathbf{x} \notin S$ , then  $Z(\mathbf{y}) < Z(\mathbf{x})$  for all  $\mathbf{y} \in \mathbf{A}(\mathbf{x})$ ,*
  - *if  $\mathbf{x} \in S$ , then  $Z(\mathbf{y}) \leq Z(\mathbf{x})$  for all  $\mathbf{y} \in \mathbf{A}(\mathbf{x})$ ,*
- 3) *the mapping  $\mathbf{A}$  is closed at points outside  $S$ .*

*Then the limit of any convergent subsequence of  $\{\mathbf{x}_k\}$  is a solution. If under the conditions of the Global Convergence Theorem,  $S$  consists of a single point  $\bar{\mathbf{x}}$ , then the sequence  $\{\mathbf{x}_k\}$  converges to  $\bar{\mathbf{x}}$ .*

Using Theorem 4 and Lemmas 9–11 we now establish the asymptotic convergence of the sequence  $\{\mathbf{x}_k\}$  to a local minimum.

**Theorem 5.** *The iterate sequence  $\{\mathbf{x}_k\}$  generated by Algorithm 1 on the class of  $\mathcal{C}^2$  gradient and Hessian Lipschitz, coercive Morse functions has a convergent subsequence that converges to a local minimum. Since the local minimum is a fixed point of Algorithm 1 the sequence  $\{\mathbf{x}_k\}$  also converges to this local minimum.*

The proof of this theorem is given in Appendix G.

### B. Global Rate of Convergence

To develop rate of convergence of the sequence  $\{\mathbf{x}_k\}$  to some local minimum  $\mathbf{x}_{\text{optimal}}^*$  of  $f(\cdot)$  we first introduce certain propositions.

**Proposition 3.** *In some compact domain  $\mathcal{U}$ , let  $\mathcal{S}_*$  be the set of all critical points of a function  $f(\cdot)$  satisfying Assumptions A1–A4, where  $\mathbf{x}_j^* \in \mathcal{S}_*$  denotes the  $j^{\text{th}}$  critical point with  $|\mathcal{S}_*| = l$  and  $l$  is finite. Then the distance between any two critical points of the function  $f(\cdot)$  is lower bounded*

*by some  $R > 0$  where  $R > \frac{1}{\zeta M} \frac{\left( (1 + \frac{\beta}{L})^2 + \frac{1}{4(1 + \frac{\beta}{L})^2} - \frac{5}{4} \right)}{6}$  for  $\zeta > 2$ ,*

*i.e.,  $\|\mathbf{x}_i^* - \mathbf{x}_j^*\| \geq R$  for any  $\mathbf{x}_i^*$  and  $\mathbf{x}_j^*$  in  $\mathcal{S}_*$  and  $\xi$  is chosen such that  $\xi \ll R$  where  $\xi$  is bounded from Theorem 3*

*Proof.* Since a Morse function on a compact manifold has finitely many critical points [42], the compact domain  $\mathcal{U}$  will have finitely many critical points. The lower bound

$R > \frac{1}{\zeta M} \frac{\left( (1 + \frac{\beta}{L})^2 + \frac{1}{4(1 + \frac{\beta}{L})^2} - \frac{5}{4} \right)}{6}$  for  $\zeta > 2$  follows from Remark 3. ■

**Proposition 4.** *Let the sequence  $\{\mathbf{x}_k\}$  generated by Algorithm 1 on a function  $f(\cdot)$  satisfying Assumptions A1–A4 converges to the local minimum  $\mathbf{x}_{\text{optimal}}^* \in \mathcal{S}_*$  from Theorem 5 and we have  $\|\mathbf{x}_0 - \mathbf{x}_{\text{optimal}}^*\| \leq \zeta$  for some  $\zeta > 0$ , where  $\mathbf{x}_0$  is the initialization point for Algorithm 1. Also, without loss of generality we can assume the following condition on the initialization:*

$$\|\mathbf{x}_0 - \mathbf{x}_j^*\| \leq \xi$$

for some strict saddle point  $\mathbf{x}_j^*$ .

*Proof.* From Theorem 5 the sequence  $\{\mathbf{x}_k\}$  generated by Algorithm 1 converges to some local minimum  $\mathbf{x}_{\text{optimal}}^*$  and this local minimum lies in a compact set in  $\mathbb{R}^n$  from Lemma 11. Hence the compact set can be taken to be the compact domain  $\mathcal{U}$  from Proposition 3 where we have  $\mathbf{x}_0 \in \mathcal{U}$  and  $\mathbf{x}_{\text{optimal}}^* \in \mathcal{S}_* \subset \mathcal{U}$ . Finally  $\|\mathbf{x}_0 - \mathbf{x}_{\text{optimal}}^*\| \leq \zeta$  follows from the compactness of  $\mathcal{U}$ . ■

**Proposition 5.** For any Morse function, the gradient magnitude at any  $\mathbf{x} \in \mathcal{U} \setminus \bigcup_{j=1}^l \bar{\mathcal{B}}_\xi(\mathbf{x}_j^*)$  for any sufficiently small  $\xi$  is lower bounded by some  $\gamma$  where we have that:

$$\|\nabla f(\mathbf{x})\| \geq \gamma = \Omega(\xi)$$

and  $\xi$  is bounded from Theorem 3. Further, for any sufficiently small  $\varepsilon$  where  $\varepsilon \ll 1$ , we can write  $\gamma = \Theta(\varepsilon^v)$  where  $v \in [0, 1]$  is a  $\xi$  dependent parameter that controls the function geometry in regions away from its critical points. Hence, very small values of  $v$  imply well-structured functions, i.e., functions whose gradients are almost of constant order in regions away from its critical points whereas  $v \uparrow 1$  implies ill-structured functions, i.e., functions whose gradients are almost of  $\varepsilon$  order in regions away from their critical points.

*Proof.* For any Morse function on a compact domain  $\mathcal{U}$ , the region away from its critical points defined by  $\mathcal{U} \setminus \bigcup_{j=1}^l \bar{\mathcal{B}}_\xi(\mathbf{x}_j^*)$  can be categorized into three sub-regions on the basis of gradient magnitudes in these regions. Expressing the gradient magnitudes as function of  $\varepsilon$  and some  $\xi$  where  $\varepsilon < \xi$  and  $\varepsilon \ll 1$ , we can write  $\|\nabla f(\mathbf{x})\| \geq \gamma = \Theta(\varepsilon^v)$  for any  $\mathbf{x} \in \mathcal{U} \setminus \bigcup_{j=1}^l \bar{\mathcal{B}}_\xi(\mathbf{x}_j^*)$ . The parameter  $v \geq 0$  is a function of  $\xi$  which controls gradient magnitudes in regions away from the function's critical points. Since  $\xi$  is a free variable that is bounded above from Theorem 3, we can choose  $\xi$  such that  $\gamma = \Omega(\xi)$  so as to restrict  $v$  in the interval  $[0, 1]$ . Then based on the values of  $v$  we have:

- regions with “large” gradient magnitudes when  $\gamma = \Theta(\varepsilon^v)$  is a constant for  $v \downarrow 0$ ,
- regions with “moderate to small” gradient magnitudes when  $\gamma = \Theta(\varepsilon^v)$  is moderate or small for  $0 < v < 1$ , and
- regions with sufficiently “small” gradient magnitudes when  $\gamma = \Theta(\varepsilon^v)$  is almost of  $\varepsilon$  order for  $v \uparrow 1$ .

Since only the above three cases or their combinations are possible in regions away from critical points, Proposition 5 captures every possible Morse function. When a function in regions away from its critical points satisfies a combination of two or more of these cases, then  $\gamma$  is automatically the minimum of the occurring cases as  $\|\nabla f(\mathbf{x})\|$  is lower bounded by  $\gamma$ . ■

Note that from Proposition 5 for  $v$  close to 0 the quantity  $\gamma$  is of constant order, i.e.,  $\gamma \approx \Theta(1)$ . Since  $\gamma = \Omega(\xi)$  and  $\gamma$  is of constant order hence we will have that  $\gamma \gg \xi$  which implies  $\gamma > \frac{1}{\sqrt{2}}L\xi$  for moderate values of  $\xi$  and therefore the

<sup>7</sup>The value of  $v$  cannot be greater than or equal to 1 since by definition  $\gamma = \Omega(\xi)$  and  $\xi > \varepsilon$  which implies  $\gamma = \Omega(\varepsilon)$ .

**no-return condition** to such  $\xi$ -saddle neighborhood holds from Lemma 6. For all other choices of  $v$  we have  $\gamma = \Theta(\varepsilon^v)$  and therefore  $\xi = \mathcal{O}(\varepsilon^v)$  where  $\varepsilon \ll 1$  due to which **no-return condition** to a small  $\xi$ -saddle neighborhood holds from Lemma 5.

**Remark 6.** The gradient Lipschitz condition gives an upper bound on the gradient norm within the function landscape, whereas we require a lower bound on the gradient norm. The reason is that in order to obtain rates of convergence for our particular problem in some compact domain  $\mathcal{U}$ , we need rates in the following three regions:

- Region **R1**, given by  $\bigcup_{j=1}^l \bar{\mathcal{B}}_\xi(\mathbf{x}_j^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}_j^*)$ ,
- Region **R2**, given by  $\bigcup_{j=1}^l \bar{\mathcal{B}}_\varepsilon(\mathbf{x}_j^*)$ , and
- Region **R3**, given by  $\mathcal{U} \setminus \bigcup_{j=1}^l \bar{\mathcal{B}}_\xi(\mathbf{x}_j^*)$ .

While one can derive optimal rates in regions **R1** and **R2** using only the Lipschitz assumption, the rate in region **R3** can only be upper bounded by  $\mathcal{O}\left(\left(\inf_{\mathbf{x}_k \in \mathcal{U} \setminus \bigcup_{j=1}^l \bar{\mathcal{B}}_\xi(\mathbf{x}_j^*)} \|\nabla f(\mathbf{x}_k)\|^2\right)^{-1}\right)$  using the gradient Lipschitz condition alone. We afterwards require a lower bound on the gradient norm in the region **R3** so as to upper bound this Big-O term. It is clear that the gradient Lipschitz bound cannot help us here since it gives an upper bound on the gradient norm. Since  $\inf_{\mathbf{x}_k \in \mathcal{U} \setminus \bigcup_{j=1}^l \bar{\mathcal{B}}_\xi(\mathbf{x}_j^*)} \|\nabla f(\mathbf{x}_k)\|$  can be arbitrary small, as the function could be extremely flat in regions away from the critical points, it becomes imperative to introduce the parameter  $v$ , which controls the function geometry in region **R3**. Note that properties like the Polyak-Łojasiewicz condition can also be used to lower bound the gradient norms, but such properties only work with specialized nonconvex functions such as the invex functions that do not have saddle points.

Our next lemma establishes the Lipschitz continuity of  $f(\cdot)$  in the compact domain  $\mathcal{U}$ .

**Lemma 12.** As a consequence of Proposition 4 the function  $f(\cdot)$  is Lipschitz continuous in the compact domain  $\mathcal{U}$ , where the Lipschitz constant is given by  $L\text{diam}(\mathcal{U})$ .

*Proof.* By the gradient Lipschitz continuity of  $f$  for any  $\mathbf{x} \in \mathcal{U}$  where  $\mathcal{U}$  has atleast one critical point  $\mathbf{x}^*$  of  $f$ , we have the following bound:

$$\|\nabla f(\mathbf{x})\| \leq L\|\mathbf{x} - \mathbf{x}^*\| \leq L\text{diam}(\mathcal{U}) \quad (29)$$

$$\implies \sup_{\mathbf{x} \in \mathcal{U}} \|\nabla f(\mathbf{x})\| \leq L\text{diam}(\mathcal{U}). \quad (30)$$

From the Mean value theorem, for any  $\mathbf{x}, \mathbf{y} \in \mathcal{U}$  we have that:

$$f(\mathbf{x}) - f(\mathbf{y}) \leq \sup_{\mathbf{x} \in \mathcal{U}} \|\nabla f(\mathbf{x})\| \|\mathbf{x} - \mathbf{y}\| \leq L\text{diam}(\mathcal{U}) \|\mathbf{x} - \mathbf{y}\|. \quad (31)$$

The above lemma will help us in developing global rates of convergence in terms of the iterate sequence  $\{\mathbf{x}_k\}$ . In the absence of this lemma global rates of convergence can still be obtained however such rates would be in terms of the function value sequence  $\{f(\mathbf{x}_k)\}$ . Since the condition

$\mathbf{x}_k \rightarrow \mathbf{x}_{optimal}^*$  implies strong convergence whereas the condition  $f(\mathbf{x}_k) \rightarrow f(\mathbf{x}_{optimal}^*)$  implies weak convergence, Lemma 12 becomes absolutely necessary for establishing a stronger convergence result.

Now that we are interested in developing convergence rates for the iterate sequence, we need a handle on the largest distance our iterate  $\mathbf{x}_k$  can possibly travel from the initialization  $\mathbf{x}_0$  within some compact domain  $\mathcal{U}$  before converging to a neighborhood of  $\mathbf{x}_{optimal}^*$ . Quantifying this distance is essential since the total number of iterations or the travel time of any trajectory depends on how much distance it travelled before converging to some local minimum neighborhood. In the best case the trajectory could take a bee line path between  $\mathbf{x}_0$  and  $\mathbf{x}_{optimal}^*$  whereas in the worst case a trajectory could possibly travel much farther than  $\mathbf{x}_{optimal}^*$  before turning back and eventually converging. The next theorem provides a precise bound on the farthest distance any worst case trajectory could travel to before returning back for good. In doing so it also provides a handle on the number of saddle point neighborhoods encountered in the path of such trajectory.

**Theorem 6.** *On a function satisfying Assumptions A1–A4, the trajectory generated from the iterate sequence  $\{\mathbf{x}_k\}$  by Algorithm 1 that has escaped some ball  $\mathcal{B}_{R_0}(\mathbf{x}_0^*)$  cannot escape the ball  $\mathcal{B}_{R_\omega}(\mathbf{x}_0^*) \supset \mathcal{B}_{R_0}(\mathbf{x}_0^*)$  if it has to re-enter the ball  $\mathcal{B}_{R_0}(\mathbf{x}_0^*)$  in finite number of iterations, where we have that  $\mathbf{x}_0 \in \mathcal{B}_\xi(\mathbf{x}_0^*)$  and  $\mathbf{x}_0^* \in \mathcal{S}_*$  is a strict saddle point provided that the radius  $R_\omega$  satisfies the condition:*

$$R_\omega \leq R_0 + 2L\text{diam}(\mathcal{U})\frac{R_0}{\gamma} + N_0K_{exit}\left(\frac{1}{\beta} + \frac{L}{2\beta^2}\right)\frac{L^2\varepsilon^2}{\gamma} + N_0(K_{exit} + K_{shell})\xi + \frac{1}{2}(R_0 + \xi) \quad (32)$$

where  $N_0 = \frac{2L\text{diam}(\mathcal{U})\frac{R_0}{\gamma}}{\left(\frac{\gamma}{2} - \left(\frac{1}{\beta} + \frac{L}{2\beta^2}\right)\frac{L^2\varepsilon^2}{\gamma} - \gamma(K_{exit} + K_{shell})\frac{\xi}{R}\right)}$  is an upper

bound on the number of strict saddle neighborhoods of radius  $\xi$  encountered by the trajectory of  $\{\mathbf{x}_k\}$ . Note that here  $K_{exit}$  is upper bounded by (7),  $K_{shell}$  is upper bounded by Theorem 3 and the compact domain  $\mathcal{U}$  contains the ball  $\mathcal{B}_{R_\omega}(\mathbf{x}_0^*)$ , i.e.,  $\mathcal{U} \supset \mathcal{B}_{R_\omega}(\mathbf{x}_0^*)$ .

The proof of this theorem is given in Appendix H.

**Remark 7.** *In order to characterize the convergence rate for Algorithm 1 we need to focus on the worst-case trajectories that can be generated by it. Theorem 6 helps capture the behavior of such worst-case trajectories by finding the radius of the largest possible ball whose boundary can be reached by such trajectories.*

We are now ready to state the final theorem of this work which quantifies the convergence rate of Algorithm 1 to some  $\varepsilon$ -neighborhood of a local minimum.

**Theorem 7.** *On a function satisfying Assumptions A1–A4, the total time  $K_{max}$  for the trajectory of  $\{\mathbf{x}_k\}$  generated from Algorithm 1 to converge to a sufficiently small  $\varepsilon$ -neighborhood*

*of a local minimum  $\mathbf{x}_{optimal}^*$  is bounded by:*

$$K_{max} < T\left(K_{exit} + K_{shell}\right) + 4L\text{diam}(\mathcal{U})\frac{\xi L}{\gamma^2} + 2T\left(\frac{1}{\beta} + \frac{L}{2\beta^2}\right)\frac{\varepsilon^2}{\gamma^2} + \frac{\log\left(\frac{\xi}{\varepsilon}\right)}{\log\left(1 - \frac{\beta}{L}\right)^{-1}}, \quad (33)$$

where  $T < \frac{2L\text{diam}(\mathcal{U})\frac{\xi}{R}}{\left(\frac{\gamma}{2} - \left(\frac{1}{\beta} + \frac{L}{2\beta^2}\right)\frac{L^2\varepsilon^2}{\gamma} - \gamma(K_{exit} + K_{shell})\frac{\xi}{R}\right)}$  is the total number of  $\xi$  radius saddle neighborhoods encountered,  $\varepsilon$  and  $\xi$  are bounded from Theorems 1, 3 and  $\mathbf{x}_0$  is initialized in a  $\xi$ -neighborhood of any strict saddle point.

The proof of this theorem is given in Appendix H. In terms of the order notation, using (7) and (28) followed by choosing some sufficiently small  $\varepsilon$  where  $\varepsilon$  is bounded by theorem 1, some moderately small  $\xi$  from Propositions 3, 5 and substituting  $\gamma = \Theta(\varepsilon^v)$ ,  $K_{max}$  has the following dependency on  $\varepsilon$ :

$$K_{max} = \mathcal{O}\left(T \log\left(\frac{1}{\varepsilon}\right)\right) + \mathcal{O}\left(T \log\left(\frac{\xi}{\varepsilon}\right)\right) + \mathcal{O}\left(\frac{1}{\varepsilon^{2v}}\right) \quad (34)$$

where  $T = \mathcal{O}\left(\frac{1}{\varepsilon^v}\right)$  is the number of saddles encountered and  $v \in [0, 1)$  is a parameter of the function  $f(\cdot)$  defined in Proposition 5 which controls the function geometry in regions away from its critical points. The third term on the right hand side of (34) is  $\mathcal{O}\left(\frac{1}{\varepsilon^{2v}}\right)$  which quantifies the travel time of the trajectory in the region  $\mathcal{U} \setminus \bigcup_{j=1}^U \bar{\mathcal{B}}_\xi(\mathbf{x}_j^*)$  (for details, see proof of Theorem 7 in Appendix H).

Observe that the dominant term in the expression of convergence rate from (34) is  $\mathcal{O}\left(\frac{1}{\varepsilon^{2v}}\right)$  where  $v \in [0, 1)$ . Compared to the state of the art<sup>8</sup> Perturbed GD method [24] which has a convergence rate of order  $\mathcal{O}\left(\frac{1}{\varepsilon^2} \log^4\left(\frac{1}{\varepsilon}\right)\right)$ , there is no poly-logarithmic dependence in our term  $\mathcal{O}\left(\frac{1}{\varepsilon^{2v}}\right)$  and in the worst case this term is still better than  $\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$  provided  $\varepsilon$  and  $\xi$  are chosen to be sufficiently small from Proposition 5. In particular, for well-structured functions which have large gradient magnitudes in regions away from critical points, we will have  $\frac{1}{\varepsilon^{2v}} \ll \frac{1}{\varepsilon^2}$  thereby yielding a superior convergence rate to sufficiently small neighborhood of a local minimum. This improvement over the rate  $\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$  is only possible because of Theorem 3 which gives a linear travel time within  $\xi$  radius saddle neighborhoods. In the absence of Theorem 3, we would not have  $\xi$  radius saddle neighborhoods within which fast travel is possible. Then we only have a much smaller  $\varepsilon$

<sup>8</sup>While Table 1 lists various state-of-the-art algorithms, all those listed works except [24] use either accelerated gradient methods or Newton method as their base algorithm. Hence for sake of fairness, the rate comparison is done only with the Perturbed GD method of [24].

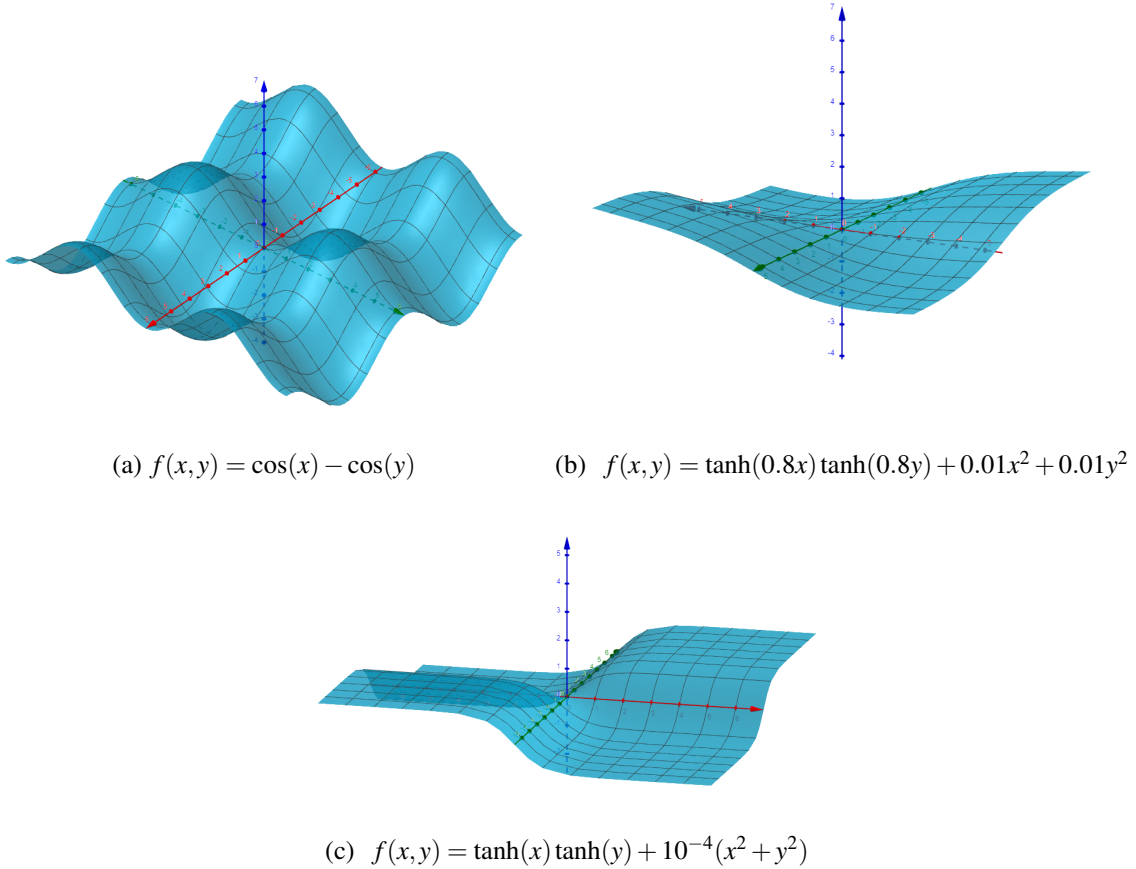


Fig. 1. Landscapes of three smooth nonconvex functions with different values of the parameter  $v$  for  $\varepsilon = 10^{-8}$ . The function in (a) corresponds to  $v = 0$ , the function in (b) has  $v = 3/4$ , while the function in (c) corresponds to  $v = 0.99$ . It is evident, at least from these simple examples, that the functions that have a relatively steeper landscape in regions away from their critical points have a small  $v$ , whereas functions with a relatively flatter landscape in regions away from the critical points have a large  $v$ .

radius saddle neighborhood from Theorem 1 and outside such neighborhood, the travel time of the trajectory will be  $\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$ . Existence of larger saddle neighborhoods from Theorem 3 enables us to invoke Proposition 5 using which we can choose our  $\varepsilon$  sufficiently small and a certain  $\xi$  so that the gradient magnitude in the region  $\mathcal{U} \setminus \bigcup_{j=1}^l \bar{\mathcal{B}}_\xi(\mathbf{x}_j^*)$  is lower bounded by  $\gamma = \Omega(\xi) = \Theta(\varepsilon^v)$  for some  $v \in [0, 1)$ . Then we get the improved rate of  $\mathcal{O}\left(\frac{1}{\varepsilon^{2v}}\right)$  in the region  $\mathcal{U} \setminus \bigcup_{j=1}^l \bar{\mathcal{B}}_\xi(\mathbf{x}_j^*)$  for our trajectory. It should however be noted that the value of parameter  $v$  is not known explicitly since it depends on the function landscape in the region  $\mathcal{U} \setminus \bigcup_{j=1}^l \bar{\mathcal{B}}_\xi(\mathbf{x}_j^*)$ . Specifying certain value for  $v$  would require more assumptions on the function landscape which is beyond the scope of this work. We now state a lemma with three example functions, one with  $v$  equal to 0 (Figure 1(a)), another for which  $v$  is between 0 and 1 (Figure 1(b)), and the last one for which  $v$  is close to 1 (Figure 1(c)).

**Lemma 13.** Suppose  $\xi = C\varepsilon^v$  for some sufficiently small  $\varepsilon > 0$ , where  $v \in [0, 1)$ ,  $C \gg \varepsilon$ , and  $\varepsilon \leq \xi$ . Then the following statements hold:

(a) The function  $f(x,y) = \cos(x) - \cos(y)$  on  $\mathcal{U} = \mathbb{R}^2$  admits

the constant  $v = 0$  for any  $\xi \in (0, 0.1]$  and for any  $\varepsilon \in (0, 0.1]$ .

(b) The function  $f(x,y) = \tanh(0.8x) \tanh(0.8y) + 0.01x^2 + 0.01y^2$  on  $\mathcal{U} = \mathbb{R}^2$  admits the constant  $v = \frac{3}{4}$  for any  $\xi \in (0, 10^{-6}]$  and for any  $\varepsilon \in (0, 10^{-8}]$ .

(c) The function  $f(x,y) = \tanh(x) \tanh(y) + 10^{-4}(x^2 + y^2)$  on  $\mathcal{U} = \mathbb{R}^2$  admits the constant  $v = 0.99$  for any  $\xi \in (0, 10^{-8}]$  and for any  $\varepsilon \in (0, 10^{-8}]$ .

A proof of this lemma is deferred to the supplementary file of this work for compactness and readability purposes. In a nutshell, the first example above is a periodic function<sup>10</sup> with a saddle point at the origin, where it can be seen that the norm of the directional derivatives in the  $x$  and  $y$  variable around zero grow like  $\sin(|x|)$  and  $\sin(|y|)$ , which are roughly

<sup>9</sup>Note that the parameter  $v$  in the third function can be equal to 1 for certain  $\varepsilon, \xi$  pair by setting  $\varepsilon = \xi$ . However even with that choice our convergence rate from Theorem 7 will be  $\mathcal{O}(\varepsilon^{-2})$  which is better than the  $\mathcal{O}(\varepsilon^{-2} \log^4(\varepsilon^{-2}))$  rate in the existing state of the art [24].

<sup>10</sup>It should be noted that the result of  $v = 0$  from Lemma 13 part (a) will hold true for any higher dimensional sinusoids by extending the argument from the proof of Lemma 13 part (a) for dimensions greater than 2. For instance functions of the form  $f(\mathbf{x}) = \sum_{i=1}^n a_i \cos(b_i x_i)$  will have  $v = 0$  provided  $a_i, b_i$  are sufficiently far away from 0.

proportional to  $|x|$  and  $|y|$ , respectively, for  $x$  and  $y$  around zero. This leads to linear growth of the gradient norm. Since the function is perfectly conditioned at its critical points, we can choose  $\xi \gg \varepsilon$  that then corresponds to  $v = 0$ . The second and third examples are coercive functions involving a quadratic term and hyperbolic tangent functions, where we can adjust the gradient growth around the critical points by arranging the magnitude of the quadratic term and the scaling of the hyperbolic tangent.

We now establish that the convergence guarantees from Theorems 5 and 7 will hold even when the function  $f(\cdot)$  is not globally gradient and Hessian Lipschitz continuous. To do so we use a theorem from nonlinear functional analysis [46], namely, the Kirschbraun Theorem [46], [47], which is stated below.

**Theorem 8 ([47]).** *If  $U$  is a subset of some Hilbert space  $H_1$ , and  $H_2$  is another Hilbert space, and*

$$g : U \rightarrow H_2$$

*is a Lipschitz-continuous map, then there is a Lipschitz-continuous map*

$$G : H_1 \rightarrow H_2$$

*that extends  $g$  and has the same Lipschitz constant as  $G$ .*

Using Theorem 8 we now show that the CCRGD algorithm (Algorithm 1) converges to a local minimum and has the same convergence rate as the one in Theorem 7 even without the gradient and Hessian Lipschitz boundedness assumption on the function  $f(\cdot)$ .

**Theorem 9.** *Suppose the function  $f(\cdot)$  satisfies only Assumptions A1 and A4. Then there exists a finite constant  $\bar{L} > 0$  such that if  $\alpha = \frac{1}{\bar{L}}$  the iterate sequence  $\{\mathbf{x}_k\}$  generated by the proposed CCRGD algorithm 1 always stays within the compact set  $\{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$  and converges to a local minimum in this compact set with the convergence rate given by Theorem 7*

The proof of this theorem is given in Appendix H. This theorem is of particular significance due to the fact that it generalizes our theoretical developments to a larger class of learning problems where the objective function may not be globally gradient Lipschitz continuous, one such problem being the low-rank matrix factorization.

### C. Computational Complexity of the CCRGD Algorithm

Essentially, the Curvature Conditioned Regularized Gradient Descent (CCRGD) algorithm follows the Gradient Descent (GD) update and switches to a second-order step at most once in any  $\varepsilon$ -strict saddle neighborhood and only if the curvature check condition (15) from CCRGD is not satisfied. The GD update can be computed in time  $t_{GD} = \Theta(n)$  and checking the condition (15) incurs a maximum time of  $t_{misc} = \Theta(n)$  per iteration of the CCRGD algorithm. The second-order step from CCRGD algorithm involves normalization of the gradient with  $\beta$ , which takes  $\Theta(n)$  time as well as the computation of one of the dominant eigenvectors. The latter

can be computed easily using power iteration-type methods, which incur a computation time of  $\Theta(n^2 \log_{\lambda_1/\lambda_2}(\varepsilon^{-1}))$ ; here,  $\lambda_1, \lambda_2$  are the top-two dominant eigenvalues of  $\mathbf{I} - \alpha \nabla^2 f(\mathbf{x}_k)$  with  $\lambda_1 > \lambda_2$  and  $\varepsilon$  is the desired accuracy for the power iteration (see, e.g., [48]–[50]). Therefore, the second-order constraint (SOC) step (Step 17 or 21) can incur a maximum computation time of  $t_{SOC} = \Theta(n^2 \log_{\lambda_1/\lambda_2}(\varepsilon^{-1}))$  in any strict-saddle neighborhood, where the maximum number of such strict-saddle neighborhoods is bounded by

$$T < \frac{2L \text{diam}(\mathcal{U})^{\frac{\xi}{R}}}{\left(\frac{\gamma}{2} - \left(\frac{1}{\beta} + \frac{L}{2\beta^2}\right) \frac{L^2 \varepsilon^2}{R} - \gamma(K_{exit} + K_{shell}) \frac{\xi}{R}\right)}$$

from Theorem 7. Next, we have also  $K_{max} = K_{GD} + K_{CCRGD}$  from Theorem 7 where  $K_{max}$  is the total iteration complexity of the CCRGD algorithm,  $K_{CCRGD} \leq T$  is the number of second-order steps and  $K_{GD}$  is the number of the GD steps. Then, using the bound on  $K_{max}$  from Theorem 7, the total computational complexity  $t_{CCRGD}$  of the proposed CCRGD algorithm is bounded by:

$$t_{CCRGD} \leq t_{GD}K_{GD} + t_{SOC}K_{CCRGD} + t_{misc}K_{max} \quad (35)$$

$$\leq (t_{GD} + t_{misc})K_{max} + t_{SOC}T \quad (36)$$

$$= \Theta\left(nT \log\left(\frac{1}{\varepsilon}\right)\right) + \Theta\left(nT \log\left(\frac{\xi}{\varepsilon}\right)\right) + \Theta\left(\frac{n}{\varepsilon^{2v}}\right) + \Theta\left((n^2 \log_{\lambda_1/\lambda_2}(\varepsilon^{-1}))T\right). \quad (37)$$

To summarize, we can make use of existing off-the-shelf eigenvalue/eigenvector solvers for the constrained problems, and our complexity benefits from the fact that we do not need to solve these constrained problems frequently.

## VIII. NUMERICAL RESULTS

To test the efficacy of the proposed method, we simulate Algorithm 1 on two different problems, a modified Rastrigin function and a low-rank matrix factorization problem.

### A. Modified Rastrigin Function

The Rastrigin function is a nonconvex function that was first proposed in [51] and the generalized versions appeared in [52], [53]. The function is given by

$$f(\mathbf{x}) = A n + \sum_{i=1}^n (x_i^2 - \cos(2\pi x_i)), \quad (38)$$

where  $A = 10$  and  $x_i \in [-5.12, 5.12]$ , and  $f(\cdot)$  has a global minimum at  $\mathbf{x} = \mathbf{0}$ . In this section, we use a modified version of (38) given by:

$$f(\mathbf{x}) = \sum_{i=1}^n a_i \cos(b_i x_i), \quad (39)$$

where (39) differs from (38) in the sense that (39) does not have a quadratic term added to it (hence possibly some local minima are global minima). The modified formulation of the Rastrigin function is analytic and locally Morse at its critical point  $\mathbf{x}^* = \mathbf{0}$  for the choice of parameters given below. It satisfies all the listed Assumptions A1–A4 in this work except

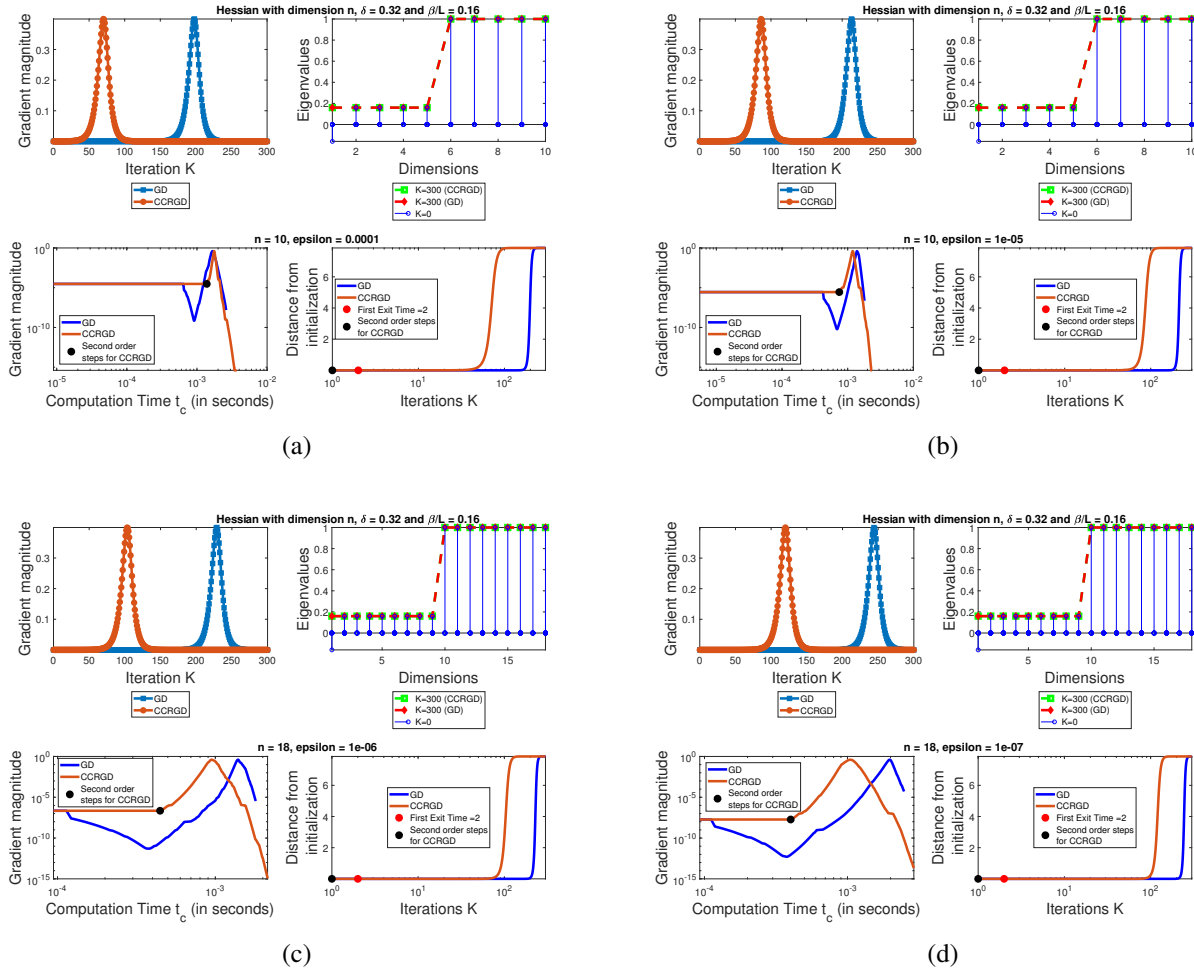


Fig. 2. Simulation results for gradient descent (GD) and the proposed CCRGD algorithm on the modified Rastrigin function for various values of  $n$  and  $\varepsilon$ .

coercivity due to the fact that we removed the quadratic growth term from it. In particular, for the formulation (39) we will have  $L \leq \sum_i |a_i b_i|$ ,  $M \leq \sum_i |a_i b_i^2|$  and  $\beta, \delta$  are evaluated from the simulations. This particular example highlights the fact that convergence to a local minimum is possible even without the coercivity assumption.

For simulations, we set  $a_i = 1$  for  $i = 1$  and  $a_i = -1$  elsewhere,  $b_i = 1$  for  $1 \leq i \leq \lfloor \frac{n}{2} \rfloor$  and  $b_i = 0.4$  for  $\lfloor \frac{n}{2} \rfloor + 1 \leq i \leq n$ . The point  $\mathbf{x}^* = \mathbf{0}$  is a strict saddle point in our case and the initialization of the proposed CCRGD algorithm (Algorithm 1) and the gradient descent (GD) method is done in an  $\varepsilon$  neighborhood of  $\mathbf{x}^*$ . Specifically, the iterate  $\mathbf{x}_0$  is initialized in an  $\varepsilon$  neighborhood of the strict saddle point  $\mathbf{x}^*$  with a very small unstable subspace projection value, i.e.,  $\left\| \pi_{\mathcal{E}_{US}}(\mathbf{x}_0 - \mathbf{x}^*) \right\| < 10^{-8}$  where  $\mathcal{E}_{US}$  is the unstable subspace of  $\nabla^2 f(\mathbf{x}^*)$  and the initialization point is same for both methods. In addition, the step size for both methods is set to  $\alpha = \frac{1}{L}$ , where  $L$  is the maximum absolute eigenvalue of the Hessian we estimated in the saddle neighborhood.

The results of our simulations are reported in Figures 2(a)–(d), where each subfigure has a total of four plots for a different combination of  $(n, \varepsilon)$ . In each of the subfigures, the top-left plot shows that the gradient norm of the proposed CCRGD

method first increases and then decreases while the GD method struggles to increase its gradient norm for many iterations. The top-right plot in each subfigure shows the initial and final eigenvalues of the Hessian at an iterate generated by the two methods, while the blue stem subplot in there shows the eigenvalue spectrum at the initialization (which is the same for both methods). It can be seen from the two plots in each subfigure that the GD method takes much longer to converge to a second-order stationary point in the given number of iterations, while the CCRGD method easily converges to a local minimum.

Finally in the bottom plots of each subfigure, the first plot is the gradient magnitude vs computation time and the second plot shows the evolution of distance of the iterate from the initialization for the two methods. The first bottom plot marks those times where the CCRGD method invoked the second-order Step. Similarly, the second bottom plot marks the iteration where the CCRGD method first exited the initial saddle neighborhood (this iteration index is the “First Exit Time”) and also marks those iteration indices where the CCRGD method invoked the second-order Step 15 in Algorithm 1. From these figures it is clear that even though the second order step of CCRGD method is computationally more time consuming, yet the overall performance in time of CCRGD is better with

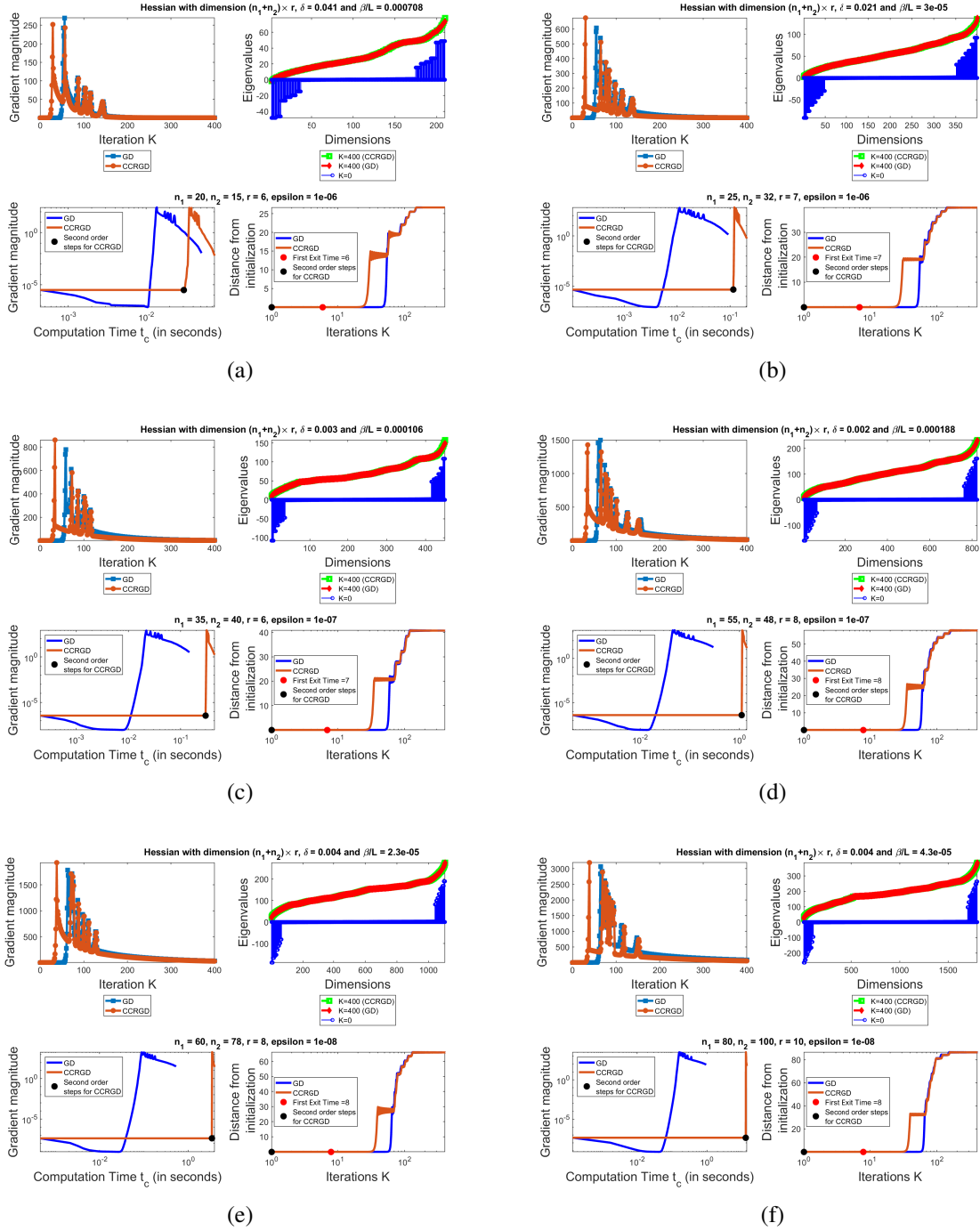


Fig. 3. Simulation results for gradient descent (GD) and the proposed CCRGD algorithm on a low-rank matrix factorization problem for various values of  $n_1$ ,  $n_2$ ,  $r$ , and  $\epsilon$ .

respect to GD for the given cost function [39]

### B. Low-Rank Matrix Factorization

The objective function for the problem in consideration is as follows:

$$f(\mathbf{X}_1, \mathbf{X}_2) = \frac{1}{4} \|\mathbf{M} - \mathbf{X}_1 \mathbf{X}_2^T\|_F^2 + \omega_1 \|\mathbf{X}_1\|_F^2 + \omega_2 \|\mathbf{X}_2\|_F^2, \quad (40)$$

where  $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$ ,  $\mathbf{X}_1 \in \mathbb{R}^{n_1 \times r}$  and  $\mathbf{X}_2 \in \mathbb{R}^{n_2 \times r}$  such that  $r \leq \min\{n_1, n_2\}$  is the rank of matrix  $\mathbf{M}$ .

To simplify the problem structure so as to make (40) some function of a single variable  $\mathbf{X}$ , let  $\mathbf{X}_1$  and  $\mathbf{X}_2$  be blocks of the variable  $\mathbf{X}$  such that

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix},$$

where we have  $\mathbf{X}_1 = \mathbf{B}_1 \mathbf{X}$  and  $\mathbf{X}_2 = \mathbf{B}_2 \mathbf{X}$  with  $\mathbf{B}_1 = \begin{bmatrix} \mathbf{I}_{n_1 \times n_1} & \mathbf{0}_{n_1 \times n_2} \end{bmatrix}$  and  $\mathbf{B}_2 = \begin{bmatrix} \mathbf{0}_{n_2 \times n_1} & \mathbf{I}_{n_2 \times n_2} \end{bmatrix}$ . Here  $\mathbf{I}_{n_1 \times n_1}$ ,  $\mathbf{I}_{n_2 \times n_2}$  represent the identity matrices and  $\mathbf{0}_{n_1 \times n_2}$ ,  $\mathbf{0}_{n_2 \times n_1}$  represent the null rectangular matrices of appropriate

dimensions. Using this change of variable, (40) can be written as a function of  $\mathbf{X}$ :

$$f(\mathbf{X}) = \frac{1}{4} \|\mathbf{M} - \mathbf{B}_1 \mathbf{X} \mathbf{X}^T \mathbf{B}_2^T\|_F^2 + \varpi_1 \|\mathbf{B}_1 \mathbf{X}\|_F^2 + \varpi_2 \|\mathbf{B}_2 \mathbf{X}\|_F^2. \quad (41)$$

Next,  $\nabla f(\mathbf{X})$  can be given as follows:

$$\begin{aligned} \nabla f(\mathbf{X}) &= \frac{1}{2} (\mathbf{B}_1^T \mathbf{B}_1 \mathbf{X} \mathbf{X}^T \mathbf{B}_2^T \mathbf{B}_2 + \mathbf{B}_2^T \mathbf{B}_2 \mathbf{X} \mathbf{X}^T \mathbf{B}_1^T \mathbf{B}_1) \mathbf{X} - \\ &\quad \frac{1}{2} (\mathbf{B}_2^T \mathbf{M}^T \mathbf{B}_1 + \mathbf{B}_1^T \mathbf{M} \mathbf{B}_2) \mathbf{X} + 2\varpi_1 \mathbf{B}_1^T \mathbf{B}_1 \mathbf{X} + 2\varpi_2 \mathbf{B}_2^T \mathbf{B}_2 \mathbf{X}. \end{aligned} \quad (42)$$

Since the gradient in (42) is a matrix, hence the corresponding Hessian will be a tensor, whereas our analysis assumes the Hessian to be a matrix. To circumvent this problem, we make use of [54, Theorem 9] by vectorizing matrix  $\mathbf{X}$  so that  $\nabla^2 f(\text{vec}(\mathbf{X}))$  is a Jacobian matrix.

The closed form expression for the Jacobian is as follows:

$$\begin{aligned} \nabla^2 f(\text{vec}(\mathbf{X})) &= \frac{1}{2} \left( ((\mathbf{X}^T \mathbf{B}_2^T \mathbf{B}_2) \otimes \mathbf{I}_{n \times n}) ((\mathbf{X} \otimes \mathbf{I}_{n \times n}) (\mathbf{I}_{r \times r} \otimes (\mathbf{B}_1^T \mathbf{B}_1))) + \right. \\ &\quad \left. (\mathbf{I}_{n \times n} \otimes (\mathbf{B}_1^T \mathbf{B}_1 \mathbf{X})) + (\mathbf{I}_{r \times r} \otimes (\mathbf{B}_1^T \mathbf{B}_1 \mathbf{X} \mathbf{X}^T)) (\mathbf{I}_{r \times r} \otimes (\mathbf{B}_2^T \mathbf{B}_2)) \right) \\ &\quad + \frac{1}{2} \left( ((\mathbf{X}^T \mathbf{B}_1^T \mathbf{B}_1) \otimes \mathbf{I}_{n \times n}) ((\mathbf{X} \otimes \mathbf{I}_{n \times n}) (\mathbf{I}_{r \times r} \otimes (\mathbf{B}_2^T \mathbf{B}_2))) \right. \\ &\quad \left. + (\mathbf{I}_{n \times n} \otimes (\mathbf{B}_2^T \mathbf{B}_2 \mathbf{X})) + (\mathbf{I}_{r \times r} \otimes (\mathbf{B}_2^T \mathbf{B}_2 \mathbf{X} \mathbf{X}^T)) (\mathbf{I}_{r \times r} \otimes (\mathbf{B}_1^T \mathbf{B}_1)) \right) \\ &\quad - \frac{1}{2} \left( \mathbf{I}_{r \times r} \otimes (\mathbf{B}_2^T \mathbf{M}^T \mathbf{B}_1 + \mathbf{B}_1^T \mathbf{M} \mathbf{B}_2) \right) \\ &\quad + 2 \left( \mathbf{I}_{r \times r} \otimes (\varpi_1 \mathbf{B}_1^T \mathbf{B}_1 + \varpi_1 \mathbf{B}_2^T \mathbf{B}_2) \right), \end{aligned} \quad (43)$$

where  $n = n_1 + n_2$ . For simulations, matrix  $\mathbf{M}$  was generated randomly using the relation

$$\mathbf{M} = \mathbf{U}_1 \mathbf{U}_2^T + \rho^2 \mathbf{N},$$

where  $\mathbf{U}_1 \in \mathbb{R}^{n_1 \times r}$ ,  $\mathbf{U}_2 \in \mathbb{R}^{n_2 \times r}$  and the entries of these matrices were independently sampled from a standard normal distribution. Matrix  $\mathbf{N} \in \mathbb{R}^{n_1 \times n_2}$  is the additive noise generated from a normal distribution whose variance is scaled by  $\rho$ . The formulation (40) is coercive, analytic (so Assumption A1 gets satisfied) and the Hessian at the critical point  $\mathbf{X} = \mathbf{0}$  is invertible (Assumption A4 gets satisfied locally at the critical point  $\mathbf{X} = \mathbf{0}$ ) but the function at  $\mathbf{X} = \mathbf{0}$  has a poor condition number which will be evident from the simulations. Since the function in (40) is analytic and hence  $\mathcal{C}^\infty$  smooth, it is gradient and Hessian Lipschitz in every compact set and therefore satisfies the Assumptions A2, A3 locally. Moreover, from Theorem 9 the CCRGD algorithm for this particular problem still converges with a convergence rate given by Theorem 7. This particular example highlights the fact that the proposed algorithm 1 can exhibit convergence to a local minimum even when the function's gradient and Hessian are not globally Lipschitz. The highly ill conditioned nature of the problem however could possibly make the function non-Morse at other critical points. Since the closed form expression of the Hessian

in (43) is very complex, we steer away from the computation of its eigenvalues at critical points other than  $\mathbf{X} = \mathbf{0}$ .

For the experiments, we use  $\varpi_1 = \varpi_2 = 0.5$ ,  $\rho = 0.5$ , and step size  $\alpha = \frac{1}{L}$  where  $L = \lambda_{\max}(\nabla^2 f(\text{vec}(\mathbf{X})))$ . Also, for the particular selection of parameters,  $\mathbf{X} = \mathbf{0}$  is a strict saddle point. Hence,  $\mathbf{X}$  is initialized on the boundary of ball  $\mathcal{B}_\varepsilon(\mathbf{0})$  and  $\varepsilon$  is varied in the simulations along with  $n_1, n_2, r$ . Finally, the proposed method is plotted against the standard gradient descent method where the metric is  $\|\mathbf{X}_k - \mathbf{X}_{\text{init}}\|_F$  with  $\mathbf{X}_{\text{init}}$  being the common initialization for the two methods.

The simulation results for Algorithm 1 are presented in Figures 3(a)–(f) and comparisons are made with the GD method. For the sake of uniformity, the plots within each subfigure of Figure 3 follow the same convention as the plots within each subfigure of Figure 2. From the plots, it is evident that the functions are not well-conditioned for different cases and both GD and CCRGD encounter cascaded saddles. Though CCRGD is computationally slightly more expensive than GD, yet still it is able to match the performance of GD in terms of convergence to a local minimum in iterations, which is evident from the eigenvalues of the Hessian at final iterate. Also from the first subplot of each figure, it is clear that CCRGD reduces gradient magnitude faster than GD eventually with respect to iterations (tail of orange CCRGD curve stays below the tail of blue GD curve in every first subplot). Moreover in every case CCRGD is able to escape the first saddle neighborhood much more faster than GD in iterations due to a single second order step which is invoked only once over all iterations.

Briefly, it can be seen from the simulations that when the objective is the modified Rastrigin function, our method (CCRGD) can improve upon GD in terms of the computational time whereas when the objective is low-rank matrix factorization, CCRGD is slower than GD when the problem dimension  $n$ , where  $n = (n_1 + n_2) \times r$ , is large due to the computationally expensive operation of eigenvector computations, which scales asymptotically as  $n^2$ . These results highlight the fact that the superiority of CCRGD over GD in terms of the computational time starts diminishing in high-dimensional problem settings. And while faster algorithms for eigenvalue/eigenvector solvers or a faster alternative to the second-order step in CCRGD can lead to further computational improvements over GD, we leave such developments for future work.

### C. Simulations for Nonconvex functions with Different Values of Parameter $v$

Recall that we had provided three examples (discussed in Lemma 13); one with  $v$  equal to 0 (Figure 1(a)), another for which  $v$  is between 0 and 1 (Figure 1(b)), and the last one for which  $v$  is close to 1 (Figure 1(c)). Other values of  $v \in (0, 1)$  can also be obtained by adjusting the constants in these examples. Our results suggest that when  $v$  is closer to 1, the convergence of both GD and the proposed CCRGD method (algorithm 1) can be slower as compared to the case when  $v$  is closer to 0. We have provided simulations in Figure 4 for illustrating this point where we compared three examples with  $v = 0$  (Figures 4(a) and 4(c)) and  $v = 0.99$  (Figure 4(b)).

In each of the subfigures in Figure 4, the top-left plot shows the evolution of gradient norms of the proposed CCRGD

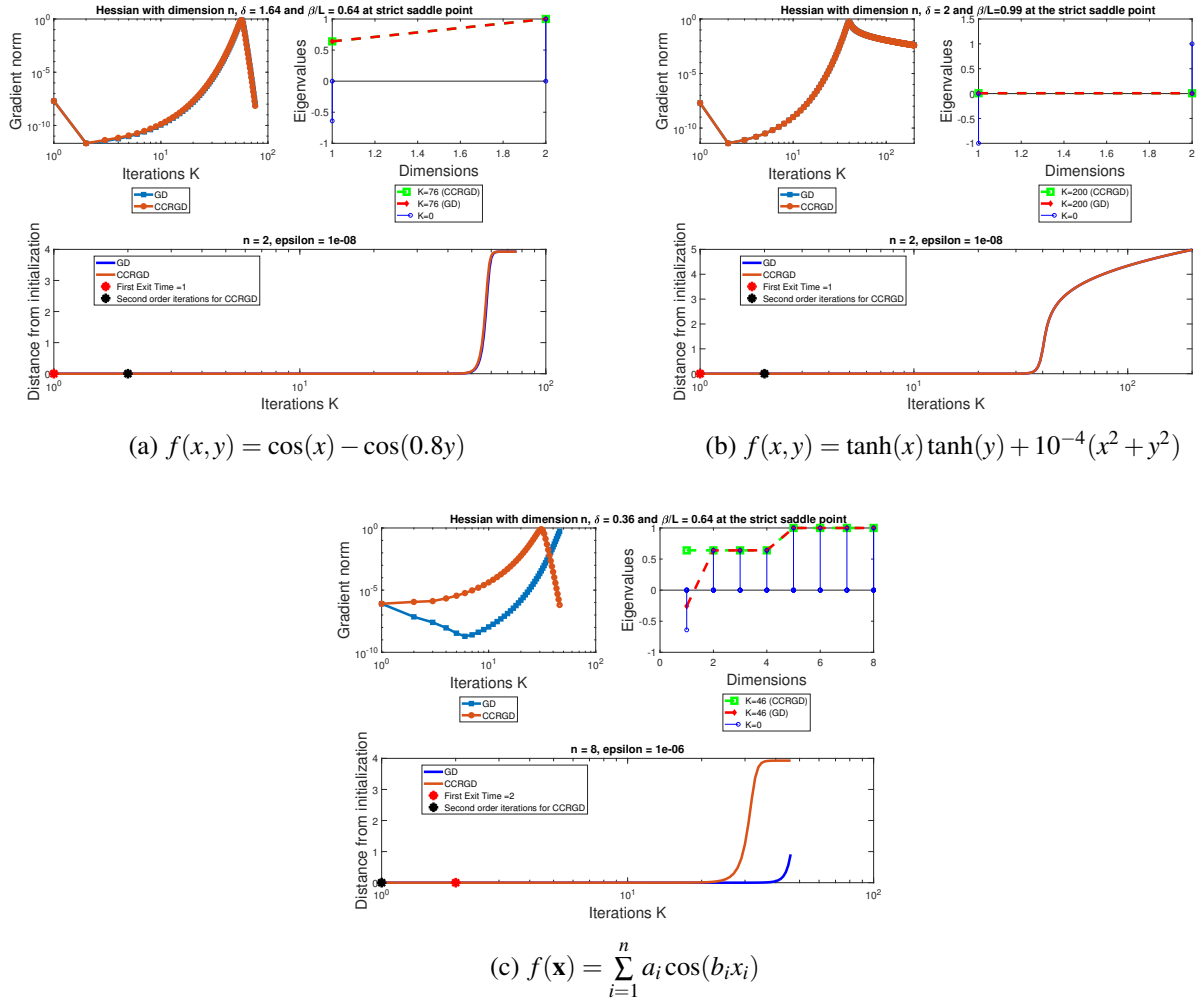


Fig. 4. Comparison between the performance of GD and CCRGD on smooth nonconvex function landscapes with different estimated values of parameter  $\nu$  for  $\epsilon = 10^{-8}$ . Functions in figures (a), (c) correspond to  $\nu = 0$ , while the function in figure (b) corresponds to  $\nu = 0.99$ . In Figure (c), we have  $b_i = 1$  for  $1 \leq i \leq 4$  and  $b_i = 0.8$  otherwise,  $a_i = -1$  for  $1 \leq i \leq 7$  and  $a_i = 1$  otherwise. Figure (c) is used to show the deviation in the performance of CCRGD and GD, which otherwise is not apparent for the lower-dimension case of  $n = 2$  in Figure (a). Also observe that in Figure (a) when  $\nu = 0$ , both algorithms converge to  $\epsilon$  accuracy in less than 100 iterations, whereas in Figure (b) when  $\nu = 0.99$ , both algorithms cannot even converge to  $10^3 \epsilon$  accuracy in 200 iterations. It is evident at least from these simple examples that these algorithms have fast convergence on functions which have a small  $\nu$ , whereas for functions with a large  $\nu$  such speedup does not happen.

method and the GD method as a function of iterations. The top-right plot in each subfigure shows the initial and final eigenvalues of the Hessian at an iterate generated by the two methods, while the blue stem subplot in there shows the eigenvalue spectrum at the initialization (which is the same for both methods). Finally, the bottom plot in each subfigure marks the iteration where the CCRGD method first exited the initial saddle neighborhood (this iteration index is the “First Exit Time”) and also marks those iteration indices where the CCRGD method invoked the second-order step (Step 17 or 21) in Algorithm 1.

It is clear from Figure 4(c) that the claimed speedup is indeed there when the parameter  $\nu$  is equal to 0. In particular, it can be observed from Figure 4(c) that while CCRGD is able to first rapidly increase and then decrease its gradient norm, thus converging to a second-order stationary point, its counterpart GD still remains in the phase of increasing its gradient norm within the number of iterations under consid-

eration. Note that this stark contrast between the performance of CCRGD and GD is not observable in Figure 4(a) because of the lower-dimensional nature of the function ( $n = 2$ ) in Figure 4(a). Indeed, both methods in this case converge extremely fast to  $\epsilon$  accuracy (in less than 100 iterations). In contrast, we see from Figure 4(b)—which corresponds to  $\nu = 0.99$ —that the gradient norm for both CCRGD and GD methods first increases rapidly, which implies fast escape from a well-conditioned saddle neighborhood, and then decreases very slowly due to the almost flat geometry of the function away from the saddle point. Hence, even with a fast saddle escape, convergence to a local minimum remains slow for both methods when  $\nu$  is close to 1.

## IX. CONCLUSION

This work focuses on the global analysis of gradient trajectories for a class of nonconvex functions that have strict saddle points in their geometry. Building on top of the results from

$$\begin{aligned} \Psi(K) \approx & \left( \left[ \left( 1 - \alpha L - \frac{\alpha \varepsilon M}{2} \right)^{2K} - 2K \left( 1 - \alpha \beta + \frac{\alpha \varepsilon M}{2} \right)^{2K-1} \frac{\alpha \varepsilon M L n}{2\delta} \right] \sum_{i \in \mathcal{N}_S} (\theta_i^s)^2 + \right. \\ & \left[ \left( 1 + \alpha \beta - \frac{\alpha \varepsilon M}{2} \right)^{2K} - 2K \left( 1 + \alpha L + \frac{\alpha \varepsilon M}{2} \right)^{2K-1} \frac{\alpha \varepsilon M L n}{2\delta} \right] \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \\ & - \frac{\alpha \varepsilon M L n}{2\delta(\alpha L + \alpha \beta)} \left( 1 + \alpha L + \frac{\alpha \varepsilon M}{2} \right)^K \left( 1 - \alpha \beta + \frac{\alpha \varepsilon M}{2} \right)^K \left( \sum_{i \in \mathcal{N}_S} (\theta_i^s)^2 + \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \right) \\ & \left. - \frac{\alpha \varepsilon M L n}{2\delta(\alpha L + \alpha \beta)} \left( 1 + \alpha L + \frac{\alpha \varepsilon M}{2} \right)^{2K} \right) \left( \sum_{i \in \mathcal{N}_S} (\theta_i^s)^2 + \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \right) \end{aligned} \quad (46)$$

$$\begin{aligned} \Psi(K) \gtrsim & \left( \left[ \left( 1 - \alpha L - \frac{\alpha \varepsilon M}{2} \right)^{2K} - 2K \left( 1 - \alpha \beta + \frac{\alpha \varepsilon M}{2} \right)^{2K-1} \frac{\alpha \varepsilon M L n}{2\delta} \right] \sum_{i \in \mathcal{N}_S} (\theta_i^s)^2 + \right. \\ & \left[ \left( 1 + \alpha \beta - \frac{\alpha \varepsilon M}{2} \right)^{2K} - 2K \left( 1 + \alpha L + \frac{\alpha \varepsilon M}{2} \right)^{2K-1} \frac{\alpha \varepsilon M L n}{2\delta} \right] \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 - \varepsilon M L n \frac{\left( 1 + \alpha L + \frac{\alpha \varepsilon M}{2} \right)^{2K}}{\delta(L + \beta)} \end{aligned} \quad (47)$$

$$\begin{aligned} \Psi(K) \gtrsim & \left( \left[ \left( -\frac{\varepsilon M}{2L} \right)^{2K} - 2K \left( 1 - \frac{\beta}{L} + \frac{\varepsilon M}{2L} \right)^{2K-1} \frac{\varepsilon M n}{2\delta} \right] \sum_{i \in \mathcal{N}_S} (\theta_i^s)^2 + \right. \\ & \left[ \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right)^{2K} - 2K \left( 2 + \frac{\varepsilon M}{2L} \right)^{2K-1} \frac{\varepsilon M n}{2\delta} \right] \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 - \varepsilon M L n \frac{\left( 2 + \frac{\varepsilon M}{2L} \right)^{2K}}{\delta(L + \beta)} \end{aligned} \quad (48)$$

$$\begin{aligned} \Psi(K) \gtrsim & \left( \left[ -2K \left( 1 - \frac{\beta}{L} + \frac{\varepsilon M}{2L} \right)^{2K-1} \frac{\varepsilon M n}{2\delta} \right] \sum_{i \in \mathcal{N}_S} (\theta_i^s)^2 + \right. \\ & \left[ \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right)^{2K} - 2K \left( 2 + \frac{\varepsilon M}{2L} \right)^{2K-1} \frac{\varepsilon M n}{2\delta} \right] \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 - \varepsilon M L n \frac{\left( 2 + \frac{\varepsilon M}{2L} \right)^{2K}}{\delta(L + \beta)} \end{aligned} \quad (49)$$

our earlier work [10], sufficient boundary conditions are developed here that guarantee approximate linear exit time of gradient trajectories from saddle neighborhoods. Further, the gradient trajectories are analyzed in an augmented saddle neighborhood and it is proved that the trajectories exhibit sequential monotonicity. Using this result, bounds on the total travel time are given for trajectories in this region. A robust algorithm is also developed in this work that uses the sufficient boundary conditions to check whether a given trajectory will exit saddle neighborhood in linear time and invokes a second-order step otherwise. Several intuitive yet important lemmas are proved, characterizing the behaviour of gradient trajectories in saddle neighborhoods and two theorems are proved that provide rate of convergence of the algorithm to a local minimum.

#### APPENDIX A

In order to prove Theorem 1 we first establish 3 supporting lemmas.

**Lemma 14.** *The smooth extension of the lower bound on the trajectory function  $\Psi(K)$  (Theorem 3.1, [10]) given by the function  $\underline{\Psi}(K)$  for  $\alpha = \frac{1}{L}$  slopes upward for some small positive values of  $K$  and then it slopes downward for very*

*large values of  $K$ , i.e.,  $\underline{\Psi}(K)$  becomes a decreasing function for large values of  $K$  ( $\underline{\Psi}(K) \rightarrow -\infty$  as  $K \rightarrow \infty$ ) provided the initial unstable projection value satisfies the necessary condition  $\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 > \Delta$  where  $\Delta > \frac{\varepsilon M L n}{\delta(L + \beta)}$ .*

*Proof.* From Theorem 3.1 in [10], for every value of parameter  $\tau$ , there exists a lower bound on the squared radial distance  $\|\tilde{\mathbf{u}}_K^\tau\|^2$  for all  $K$  in the range  $1 \leq K \leq \sup_\tau \left\{ K_{exit}^\tau \right\}$  provided  $K\varepsilon \ll 1$ . Moreover, this lower bound can be expressed using a function of  $K$  called the trajectory function  $\Psi(K)$ . Formally, we have that:

$$\varepsilon^2 \geq \inf_\tau \|\tilde{\mathbf{u}}_K^\tau\|^2 > \varepsilon^2 \Psi(K), \quad (44)$$

where the the trajectory function  $\Psi(K)$  is given by:

$$\begin{aligned} \Psi(K) = & \left( c_1^{2K} - 2K c_2^{2K-1} b_1 - b_2 c_3^K c_2^K - b_2 c_3^{2K} \right) \sum_{i \in \mathcal{N}_S} (\theta_i^s)^2 + \\ & \left( c_4^{2K} - 2K c_3^{2K-1} b_1 - b_2 c_3^K c_2^K - b_2 c_3^{2K} \right) \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \end{aligned} \quad (45)$$

$$\begin{aligned} \Psi(K) = & \left( \left[ -2K \left( 1 - \frac{\beta}{L} + \frac{\varepsilon M}{2L} \right)^{2K-1} \frac{\varepsilon M n}{2\delta} \right] \sum_{i \in \mathcal{N}_S} (\theta_i^s)^2 + \right. \\ & \left. \left[ \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right)^{2K} - 2K \left( 2 + \frac{\varepsilon M}{2L} \right)^{2K-1} \frac{\varepsilon M n}{2\delta} \right] \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 - \varepsilon M L n \frac{\left( 2 + \frac{\varepsilon M}{2L} \right)^{2K}}{\delta(2\beta - \varepsilon M)} \right) \end{aligned} \quad (50)$$

$$\begin{aligned} \frac{d\Psi(K)}{dK} = & \left( \left[ -4K \log \left( 1 - \frac{\beta}{L} + \frac{\varepsilon M}{2L} \right) - 2 \right] \left( 1 - \frac{\beta}{L} + \frac{\varepsilon M}{2L} \right)^{2K-1} \frac{\varepsilon M n}{2\delta} \sum_{i \in \mathcal{N}_S} (\theta_i^s)^2 + \right. \\ & \left[ 2 \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right)^{2K} \log \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right) - 2 \left( 2 + \frac{\varepsilon M}{2L} \right)^{2K-1} \frac{\varepsilon M n}{2\delta} - \right. \\ & \left. \left. 4K \left( 2 + \frac{\varepsilon M}{2L} \right)^{2K-1} \frac{\varepsilon M n}{2\delta} \log \left( 2 + \frac{\varepsilon M}{2L} \right) \right] \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 - 2\varepsilon M L n \frac{\left( 2 + \frac{\varepsilon M}{2L} \right)^{2K}}{\delta(2\beta - \varepsilon M)} \log \left( 2 + \frac{\varepsilon M}{2L} \right) \right) \end{aligned} \quad (51)$$

with  $c_1 = \left( 1 - \alpha L - \frac{\alpha \varepsilon M}{2} - \mathcal{O}(\varepsilon^2) \right)$ ,  $c_2 = \left( 1 - \alpha \beta + \frac{\alpha \varepsilon M}{2} + \mathcal{O}(\varepsilon^2) \right)$ ,  $c_3 = \left( 1 + \alpha L + \frac{\alpha \varepsilon M}{2} + \mathcal{O}(\varepsilon^2) \right)$ ,  $c_4 = \left( 1 + \alpha \beta - \frac{\alpha \varepsilon M}{2} - \mathcal{O}(\varepsilon^2) \right)$ ,  $b_1 = \left( \frac{\alpha \varepsilon M L n}{2\delta} + \mathcal{O}(\varepsilon^2) \right)$ , and  $b_2 = \frac{\left( \frac{\alpha \varepsilon M L n}{2\delta} + \mathcal{O}(\varepsilon^2) \right) \left( 1 + \mathcal{O}(K\varepsilon) \right)}{\left( \alpha L + \alpha \beta + \mathcal{O}(\varepsilon^2) \right)}$ .

Substituting these coefficients in the expression for  $\Psi(K)$  followed by dropping order  $\mathcal{O}(\varepsilon^2)$  and  $\mathcal{O}(K\varepsilon)$  terms (for  $K\varepsilon \ll 1$ ) appearing on its right hand side, we get an approximate expression for  $\Psi(K)$  from (46) where in (47) we used the relation  $\left( \sum_{i \in \mathcal{N}_S} (\theta_i^s)^2 + \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \right) = 1$  and the inequality  $\left( 1 - \alpha \beta + \frac{\alpha \varepsilon M}{2} \right) < \left( 1 + \alpha L + \frac{\alpha \varepsilon M}{2} \right)$ . Now for  $\alpha = \frac{1}{L}$ , (47) becomes the approximate inequality (48), which can be simplified further as (49).

We next first assume that the approximate lower bound on  $\Psi(K)$  from (49) is a continuous function of  $K$  so as to allow differentiation of this lower bound with respect to variable  $K$ . This continuous extension is possible since the approximate lower bound on  $\Psi(K)$  from (49) is a well-defined function of  $K$ . Note that we do not use the lower bound from (48) since we are looking for values of  $K$  greater than 1 and the derivative of  $\left( -\frac{\varepsilon M}{2L} \right)^{2K}$  is of at most order  $\mathcal{O}(\varepsilon^{2K-1})$  for  $K > 1$  with small  $\varepsilon$ . Representing this approximate lower bound in (49) as  $\underline{\Psi}(K)$  where we have that  $\Psi(K) \gtrsim \underline{\Psi}(K)$ , followed by differentiating it with respect to  $K$  yields the expression for  $\frac{d\Psi(K)}{dK}$  in (51).

It can be inferred from (51) that for  $\varepsilon < \frac{2\beta}{M}$  and  $\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 > \Delta$  where  $\Delta > \frac{\varepsilon M L n}{\delta(L+\beta)}$ , the function  $\underline{\Psi}(K)$  slopes upward for some small positive values of  $K$  and then it slopes downward for very large values of  $K$ , i.e.,  $\underline{\Psi}(K)$  becomes a decreasing function for large values of  $K$  ( $\underline{\Psi}(K) \rightarrow -\infty$  as  $K \rightarrow \infty$ ). ■

**Lemma 15.** *The sufficient (though not necessary) condition that guarantees the escape of the approximate lower bound  $\underline{\Psi}(K)$  on the trajectory function  $\Psi(K)$  from the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  is as follows:*

$$1 \leq \sup_{K \in G_\Psi} \left\{ \underline{\Psi}(K) \right\} \quad (52)$$

where  $G_\Psi = \left\{ K \mid K \in (0, K^1], \frac{d^2\Psi(K)}{dK^2} < 0, \frac{d\Psi(K)}{dK} = 0 \right\}$  and  $K^1 = \mathcal{O}(\log(\varepsilon^{-1}))$ . Moreover, there exists some  $K_0 = \mathcal{O}(\log(\varepsilon^{-1}))$  in the set  $G_\Psi$  implying that the set  $G_\Psi$  is non empty.

*Proof.* Recall that from the condition (44), the exit time is obtained by evaluating the first  $K$  where  $\Psi(K) > 1$ . From the inequality (49), by setting the right hand side greater than equal to 1 for some given  $K$  of order  $\mathcal{O}(\log(\varepsilon^{-1}))$ , we will have  $\Psi(K) \gtrsim 1$ . Hence the sufficient condition on the unstable projection value  $\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2$  for escaping saddle with linear rate can be obtained from (49) by setting its right hand side greater than equal to 1. Notice that for very large  $K$ , the right hand side of (49) is always less than 1. Moreover, there exists some  $K_{min} \geq 1$  and  $K_{max} > 1$  such that the approximate lower bound of (49) can become greater than 1 only in the interval  $(K_{min}, K_{max})$ . Therefore we only need to find some  $K_0 \in (K_{min}, K_{max})$  where the function  $\underline{\Psi}(K)$  has zero slope and the value  $\underline{\Psi}(K_0)$  is greater than or equal to 1 for guaranteeing escape. The condition  $\underline{\Psi}(K_0) \geq 1$  would imply  $\Psi(K_0) \gtrsim 1$  thereby approximately guaranteeing escape from the condition (44) which gets reversed for  $K = K_0$ .

The above condition can be achieved in many different ways. However, to ensure that the so-called sufficient conditions have minimal restrictions, we must have  $K_0$  to be the local maximum of the function  $\underline{\Psi}(K)$  on the interval  $K \in (0, C]$  where  $C$  is some arbitrary positive finite value with  $C \leq K_{max}$ . Note that  $K_0$  is a root of the equation  $\frac{d\Psi(K)}{dK} = 0$ . The condition that  $K_0$  is the local maximum of  $\underline{\Psi}(K)$  on the interval  $K \in (0, C]$  ensures existence of at least one value of  $K_0$  such that  $\underline{\Psi}(K_0) \geq 1$  and hence  $\Psi(K_0) \gtrsim \underline{\Psi}(K_0) \geq 1$ .

Next, recall that from Theorem 3.2 in [10] we have the condition of  $K_{exit} < K^1 \lesssim \mathcal{O}(\log(\varepsilon^{-1}))$  for  $\varepsilon$ -precision tra-

$$0 = \frac{d\Psi}{dK} \Big|_{K=K_0} = \left( \left[ -4K_0 \log \left( 1 - \frac{\beta}{L} + \frac{\varepsilon M}{2L} \right) - 2 \right] \left( 1 - \frac{\beta}{L} + \frac{\varepsilon M}{2L} \right)^{2K_0-1} \frac{\varepsilon M n}{2\delta} \sum_{i \in \mathcal{N}_S} (\theta_i^s)^2 + \right. \\ \left. \left[ 2 \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right)^{2K_0} \log \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right) - 2 \left( 2 + \frac{\varepsilon M}{2L} \right)^{2K_0-1} \frac{\varepsilon M n}{2\delta} - \right. \right. \\ \left. \left. 4K_0 \left( 2 + \frac{\varepsilon M}{2L} \right)^{2K_0-1} \frac{\varepsilon M n}{2\delta} \log \left( 2 + \frac{\varepsilon M}{2L} \right) \right] \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 - 2\varepsilon M L n \frac{\left( 2 + \frac{\varepsilon M}{2L} \right)^{2K_0}}{\delta(2\beta - \varepsilon M)} \log \left( 2 + \frac{\varepsilon M}{2L} \right) \right) \quad (54)$$

$$0 = \left( \left[ -4K_0 \log \left( 1 - \frac{\beta}{L} + \frac{\varepsilon M}{2L} \right) - 2 \right] \left( \frac{1 - \frac{\beta}{L} + \frac{\varepsilon M}{2L}}{2 + \frac{\varepsilon M}{2L}} \right)^{2K_0} \left( 1 - \frac{\beta}{L} + \frac{\varepsilon M}{2L} \right)^{-1} \frac{\varepsilon M n}{2\delta} \sum_{i \in \mathcal{N}_S} (\theta_i^s)^2 + \right. \\ \left[ 2 \left( \frac{1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}}{2 + \frac{\varepsilon M}{2L}} \right)^{2K_0} \log \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right) - 2 \left( 2 + \frac{\varepsilon M}{2L} \right)^{-1} \frac{\varepsilon M n}{2\delta} - \right. \\ \left. 4K_0 \left( 2 + \frac{\varepsilon M}{2L} \right)^{-1} \frac{\varepsilon M n}{2\delta} \log \left( 2 + \frac{\varepsilon M}{2L} \right) \right] \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 - 2\varepsilon M L n \frac{1}{\delta(2\beta - \varepsilon M)} \log \left( 2 + \frac{\varepsilon M}{2L} \right) \right) \quad (55)$$

$$0 = \underbrace{\left( \left[ -4K_0 \log \left( 1 - \frac{\beta}{L} + \frac{\varepsilon M}{2L} \right) - 2 \right] \left( 1 - \frac{\beta}{L} + \frac{\varepsilon M}{2L} \right)^{-1} \frac{\mu \eta \varepsilon^{(1+a+b)} M n}{2\delta} \sum_{i \in \mathcal{N}_S} (\theta_i^s)^2 + \right.}_{F_1} \\ \left[ 2 \left( \frac{1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}}{2 + \frac{\varepsilon M}{2L}} \right)^{2K_0} \log \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right) - 2 \left( 2 + \frac{\varepsilon M}{2L} \right)^{-1} \frac{\varepsilon M n}{2\delta} - \right. \\ \left. 4K_0 \left( 2 + \frac{\varepsilon M}{2L} \right)^{-1} \frac{\varepsilon M n}{2\delta} \log \left( 2 + \frac{\varepsilon M}{2L} \right) \right] \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 - 2\varepsilon M L n \frac{1}{\delta(2\beta - \varepsilon M)} \log \left( 2 + \frac{\varepsilon M}{2L} \right) \right) \quad (56)$$

$$\left( \frac{1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}}{2 + \frac{\varepsilon M}{2L}} \right)^{2K_0} \approx \left( 2 + \frac{\varepsilon M}{2L} \right)^{-1} \frac{\varepsilon M n}{2\delta} \frac{\log \left( 2 + \frac{\varepsilon M}{2L} \right)}{\log \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right)} 2K_0 + \left( 2 + \frac{\varepsilon M}{2L} \right)^{-1} \frac{\varepsilon M n}{2\delta \log \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right)} + \\ \frac{\varepsilon M L n}{\delta(2\beta - \varepsilon M) \log \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right) \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2} \log \left( 2 + \frac{\varepsilon M}{2L} \right) \quad (57)$$

jectories with linear exit time. Note that the linear exit time was obtained explicitly by solving for the roots of equation  $\Psi(K) = 1$ . Now  $K_0$  is the local maximum of the function  $\Psi(K)$  on the interval  $K \in (0, C]$  and we have  $\Psi(K_0) \geq 1$  hence we can set  $C = K^\dagger$  which is valid since  $C$  was arbitrary with  $K_{\text{exit}} < C \leq K_{\text{max}}$ . Similarly,  $K_{\text{max}}$  was arbitrary hence we can set  $K_{\text{max}} = 2K^\dagger$ . Therefore we will have  $\|\tilde{\mathbf{u}}_{K_0}^\tau\|^2 > \varepsilon^2$  for all values of  $\tau$  where  $\{\tilde{\mathbf{u}}_K^\tau\}_{K=0}^{K_{\text{exit}}}$  was the  $\varepsilon$ -precision trajectory defined in [10].

Then the sufficient (though not necessary) condition that guarantees the escape of the approximate lower bound  $\Psi(K)$  on the trajectory function  $\Psi(K)$  from the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  is as follows:

$$1 \leq \sup_{K \in G_\Psi} \left\{ \Psi(K) \right\} \quad (53)$$

where  $G_\Psi = \left\{ K \mid K \in (0, K^\dagger], \frac{d^2\Psi(K)}{dK^2} < 0, \frac{d\Psi(K)}{dK} = 0 \right\}$ .

The condition (53) can be relaxed to obtain  $\Psi(K_0) \geq 1$  for some  $K_0 \in G_\Psi$ . Note that the set  $G_\Psi$  is non-empty since the function  $\Psi(K)$  slopes upwards for small positive  $K$  whereas  $\Psi(K) \rightarrow -\infty$  as  $K \rightarrow \infty$ . Simplifying the derivative condition (51) by setting it to 0 we get (55). Observe that the roots of (55) cannot be explicitly computed due to the transcendental nature of this equation. However, the roots can be obtained if the order of  $K_0$  is known with respect to  $\varepsilon$ . Since  $K_0 \in G_\Psi$ , we will have  $K_0 < K^\dagger \lesssim \mathcal{O}(\log(\varepsilon^{-1}))$ . Therefore, we compute only those values of  $K_0$  which are linear, i.e.,  $K_0 = \mathcal{O}(\log(\varepsilon^{-1}))$ . For such a  $K_0$ , setting  $\frac{1}{\left( 2 + \frac{\varepsilon M}{2L} \right)^{2K_0}} = \mu \varepsilon^a$  where  $\mu > 0$ ,  $a > 0$  and

$\left( 1 - \frac{\beta}{L} + \frac{\varepsilon M}{2L} \right)^{2K_0} = \eta \varepsilon^b$  where  $\eta > 0$ ,  $b > 0$  provided  $\varepsilon < \frac{2\beta}{M}$ , the equality (55) becomes (56), which further simplifies to (57).

Note that in the step (57), we dropped the term  $F_1$  (since this term  $F_1 = \mathcal{O}(K_0 \varepsilon^{(1+a+b)}) = \mathcal{O}(\varepsilon^{(1+a+b)} \log(\varepsilon^{-1}))$ ) to obtain

$$K_0 \lesssim \underbrace{\frac{1}{2} \log \left( \frac{2 + \frac{\varepsilon M}{2L}}{1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}} \right) \left( \frac{2\delta \left( 2 + \frac{\varepsilon M}{2L} \right) \log \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right) \log \left( \frac{2 + \frac{\varepsilon M}{2L}}{1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}} \right)}{\varepsilon M n \log \left( 2 + \frac{\varepsilon M}{2L} \right)}}_{\hat{K}_0} \quad (61)$$

the approximate equality (57). The approximate solution for (57) can be obtained using a transcendental equation of the form  $q^x = cx + d$  where  $x = 2K_0$  and the coefficients are as follows:

$$q = \left( \frac{1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}}{2 + \frac{\varepsilon M}{2L}} \right), c = \left( 2 + \frac{\varepsilon M}{2L} \right)^{-1} \frac{\varepsilon M n}{2\delta} \frac{\log \left( 2 + \frac{\varepsilon M}{2L} \right)}{\log \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right)} \quad (58)$$

$$d = \left( 2 + \frac{\varepsilon M}{2L} \right)^{-1} \frac{\varepsilon M n}{2\delta \log \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right)} + \frac{\varepsilon M L n \log \left( 2 + \frac{\varepsilon M}{2L} \right)}{\delta (2\beta - \varepsilon M) \log \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right) \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2} \quad (59)$$

The solution for this equation is given by the following relation:

$$x = -\frac{W\left(-\frac{\log q}{c} q^{-\frac{d}{c}}\right)}{\log q} - \frac{d}{c} \leq \frac{\log\left(-\frac{\log q}{c} q^{-\frac{d}{c}}\right)}{\log q^{-1}} - \frac{d}{c} = \frac{\log\left(-\frac{\log q}{c}\right)}{\log q^{-1}}, \quad (60)$$

where  $W(\cdot)$  is the Lambert W function and we have that  $W(y) \leq \log(y)$  for large  $y$ . Substituting these coefficients in (57), we obtain the approximate condition (61) where  $\hat{K}_0$  is the approximate upper bound on  $K_0$ .

However, for the condition  $K_0 \in G_{\Psi}$  to hold, we also require  $\left. \frac{d^2 \Psi}{dK^2} \right|_{K=K_0} < 0$  condition to hold. It can be readily checked that  $\left. \frac{d\Psi}{dK} \right|_{K=\hat{K}_0} < 0$  whereas  $\frac{d\Psi}{dK}$  is positive for very small values of  $K$ . Hence, there must exist a local maximum at some  $K_0 < \hat{K}_0$  which would imply  $\left. \frac{d^2 \Psi}{dK^2} \right|_{K=K_0} < 0$ . Hence, it is not required to explicitly solve the condition  $\left. \frac{d^2 \Psi}{dK^2} \right|_{K=K_0} < 0$ .

It is worth mentioning that dropping the term  $F_1$  to obtain the approximate equality (57) is justified. Observe that in the two approximate transcendental equations (56) and (57) with  $K_0$  as the variable, the right-hand sides will be greater than their left-hand sides respectively at the value  $K_0 = \hat{K}_0$ . Also, for small values of  $K_0$  the respective left-hand sides of (56) and (57) dominate, hence the approximate equality occurs for some  $K_0 < \hat{K}_0$ . Now, we are only left to prove that the approximations (56) and (57) are almost equal at

$K_0 = \hat{K}_0$ . This can be established by proving that the term  $F_1 = \mathcal{O}(\hat{K}_0 \varepsilon^{(1+a+b)}) = \mathcal{O}(\varepsilon^{(1+a+b)} \log(\varepsilon^{-1}))$  is negligible w.r.t. other terms in (56) at  $K_0 = \hat{K}_0$ . From the particular approximate upper bound in (61), it can be verified that  $a > 1$ . Using the substitution  $\frac{1}{\left( 2 + \frac{\varepsilon M}{2L} \right)^{2\hat{K}_0}} = \mu \varepsilon^a$  where  $\mu > 0$ ,  $a > 0$ , taking

log both sides followed by substituting the approximate upper bound  $\hat{K}_0$  from (61) yields:

$$a \log \left( \frac{1}{\sqrt[a]{\mu \varepsilon}} \right) = 2\hat{K}_0 \log \left( 2 + \frac{\varepsilon M}{2L} \right) \quad (62)$$

$$a \log \left( \frac{1}{\sqrt[a]{\mu \varepsilon}} \right) = \frac{\log \left( \frac{2\delta \left( 2 + \frac{\varepsilon M}{2L} \right) \log \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right) \log \left( \frac{2 + \frac{\varepsilon M}{2L}}{1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}} \right)}{\varepsilon M n \log \left( 2 + \frac{\varepsilon M}{2L} \right)} \right)}{\log \left( \frac{2 + \frac{\varepsilon M}{2L}}{1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}} \right)} \log \left( 2 + \frac{\varepsilon M}{2L} \right) \quad (63)$$

$$a = \frac{\log \left( 2 + \frac{\varepsilon M}{2L} \right)}{\log \left( 2 + \frac{\varepsilon M}{2L} \right) - \log \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right)} > 1, \quad (64)$$

where in the last step we have that  $\frac{1}{\sqrt[a]{\mu \varepsilon}} = 2\delta \left( 2 + \frac{\varepsilon M}{2L} \right) \log \left( 1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \right) \log \left( \frac{2 + \frac{\varepsilon M}{2L}}{1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}} \right) / \varepsilon M n \log \left( 2 + \frac{\varepsilon M}{2L} \right)$ . Now with  $a > 1$

we have the following condition for any  $b > 0$ :

$$\lim_{\varepsilon \rightarrow 0^+} \frac{\varepsilon^{(1+a+b)} \log(\varepsilon^{-1})}{\varepsilon^2} = 0. \quad (65)$$

Hence, for sufficiently small  $\varepsilon$ , term  $F_1$  can be of at most order  $\mathcal{O}(\varepsilon^2)$ . ■

**Lemma 16.** *There exists some  $K_0 = \mathcal{O}(\log(\varepsilon^{-1}))$  in the set  $G_{\Psi}$  such that  $\Psi(K_0) \geq 1$  provided the lower bound on the unstable projection value  $\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2$  has the following order:*

$$\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \gtrsim \Theta \left( \frac{1}{\log(\varepsilon^{-1})} \right). \quad (66)$$

*Proof.* Recall that from the relaxation of condition (53), we require  $\Psi(K_0) \geq 1$ . Since  $K_0$  is not explicitly available and we only have the approximate upper bound  $\hat{K}_0$  from (61), hence we use the substitution  $K_0 = \chi \hat{K}_0$  for some  $0 < \chi \leq 1$  and set the value of function  $\Psi$  at this point greater than equal to 1.

$$\left( \left[ \left( \frac{1 + \frac{\beta}{L} - \frac{\epsilon M}{2L}}{2 + \frac{\epsilon M}{2L}} \right)^{2K_0} - 2K_0 \left( 2 + \frac{\epsilon M}{2L} \right)^{-1} \frac{\epsilon M n}{2\delta} \right] \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 - \frac{\epsilon M L n}{\delta(2\beta - \epsilon M)} \right) \gtrsim \frac{1}{\left( 2 + \frac{\epsilon M}{2L} \right)^{2K_0}} \quad (67)$$

$$\begin{aligned} & \left( \left( 2 + \frac{\epsilon M}{2L} \right)^{-1} \frac{\epsilon M n}{2\delta} \frac{\log \left( 2 + \frac{\epsilon M}{2L} \right)}{\log \left( 1 + \frac{\beta}{L} - \frac{\epsilon M}{2L} \right)} 2K_0 - \left( 2 + \frac{\epsilon M}{2L} \right)^{-1} \frac{\epsilon M n}{2\delta} 2K_0 + \left( 2 + \frac{\epsilon M}{2L} \right)^{-1} \frac{\epsilon M n}{2\delta \log \left( 1 + \frac{\beta}{L} - \frac{\epsilon M}{2L} \right)} \right) \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \gtrsim \\ & \mu^\chi \epsilon^{\chi a} + \frac{\epsilon M L n}{\delta(2\beta - \epsilon M)} - \frac{\epsilon M L n}{\delta(2\beta - \epsilon M) \log \left( 1 + \frac{\beta}{L} - \frac{\epsilon M}{2L} \right)} \log \left( 2 + \frac{\epsilon M}{2L} \right) \end{aligned} \quad (68)$$

$$\begin{aligned} & \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \gtrsim \frac{\left( 2 + \frac{\epsilon M}{2L} \right) \left( \mu^\chi \epsilon^{\chi a} + \frac{\epsilon M L n}{\delta(2\beta - \epsilon M)} - \frac{\epsilon M L n}{\delta(2\beta - \epsilon M) \log \left( 1 + \frac{\beta}{L} - \frac{\epsilon M}{2L} \right)} \log \left( 2 + \frac{\epsilon M}{2L} \right) \right)}{\left( \frac{\epsilon M n}{2\delta} \frac{\log \left( 2 + \frac{\epsilon M}{2L} \right)}{\log \left( 1 + \frac{\beta}{L} - \frac{\epsilon M}{2L} \right)} 2K_0 - \frac{\epsilon M n}{2\delta} 2K_0 + \frac{\epsilon M n}{2\delta \log \left( 1 + \frac{\beta}{L} - \frac{\epsilon M}{2L} \right)} \right)} \end{aligned} \quad (69)$$

$$\begin{aligned} & \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \gtrsim \frac{\left( 2 + \frac{\epsilon M}{2L} \right) \left( \frac{2\delta \mu^\chi \epsilon^{(\chi a - 1)}}{M n} + \frac{1}{\left( \frac{\beta}{L} - \frac{\epsilon M}{2L} \right)} - \frac{\log \left( 2 + \frac{\epsilon M}{2L} \right)}{\left( \frac{\beta}{L} - \frac{\epsilon M}{2L} \right) \log \left( 1 + \frac{\beta}{L} - \frac{\epsilon M}{2L} \right)} \right)}{2\chi \hat{K}_0 \left( \frac{\log \left( 2 + \frac{\epsilon M}{2L} \right)}{\log \left( 1 + \frac{\beta}{L} - \frac{\epsilon M}{2L} \right)} - 1 \right) + \frac{1}{\log \left( 1 + \frac{\beta}{L} - \frac{\epsilon M}{2L} \right)}}. \end{aligned} \quad (70)$$

Substituting  $K_0 = \chi \hat{K}_0$  from (61) into the condition  $\Psi(K_0) \geq 1$ , dropping the first term on the right hand side of (50) (it is of order  $\mathcal{O}(\chi \hat{K}_0 \epsilon^{(1+a+b)}) = \mathcal{O}(\epsilon^{(1+a+b)} \log(\epsilon^{-1}))$  as before, substituting  $\frac{1}{\left( 2 + \frac{\epsilon M}{2L} \right)^{2K_0}} = \mu^\chi \epsilon^{\chi a}$  for  $\mu > 0, \epsilon > 0$ , using (57)

for  $K_0 = \chi \hat{K}_0$  we get (67) which after rearranging yields (68).

Since the term given by the expression  $\left( \frac{\epsilon M n}{2\delta} \frac{\log \left( 2 + \frac{\epsilon M}{2L} \right)}{\log \left( 1 + \frac{\beta}{L} - \frac{\epsilon M}{2L} \right)} 2K_0 - \frac{\epsilon M n}{2\delta} 2K_0 + \frac{\epsilon M n}{2\delta \log \left( 1 + \frac{\beta}{L} - \frac{\epsilon M}{2L} \right)} \right)$  is positive, dividing both sides of (68) by this term yields (69) or equivalently (70) which is a sufficient condition on the unstable projection value  $\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2$ .

Now, recall that from (64) we have  $a > 1$  and we also know that  $\hat{K}_0 \gtrsim K_0 = \chi \hat{K}_0$ . Since  $K_0$  is not explicitly known we can choose a surrogate for  $\chi$  to obtain the sufficient condition. Notice that  $\chi$  is a quantity between 0 and 1. Choosing a large value for  $\chi$  say close to 1 will yield the following order bound  $\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \gtrsim \Theta \left( \frac{\epsilon^{a-1}}{\log(\epsilon^{-1})} \right)$ . Recall that from

(21) we require  $\sqrt{\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2} > \mathcal{O} \left( \frac{1}{\sqrt{\epsilon}} \left( \log \left( \frac{1}{\epsilon} \right) \epsilon \right) \right)$ . However this bound may then contradict the sufficient con-

dition  $\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \gtrsim \Theta \left( \frac{\epsilon^{a-1}}{\log(\epsilon^{-1})} \right)$  if  $a > 2$ , i.e., we have  $\Theta \left( \frac{1}{\epsilon} \left( \log \left( \frac{1}{\epsilon} \right) \epsilon \right)^2 \right) > \Theta \left( \frac{\epsilon^{a-1}}{\log(\epsilon^{-1})} \right)$  as  $\epsilon \rightarrow 0$  (for well conditioned problems, i.e.,  $\frac{\beta}{L}$  close to 1, it can be checked using (64) that  $a$  becomes arbitrarily large). Next, choosing very small values of  $\chi$  say close to 0 will cause the approximation (57) to fail since the term  $F_1$  in (56) can no longer be dropped (this term is of order  $\mathcal{O}(\epsilon)$  for  $\chi = 0$ ).

However, the choice  $\chi = \frac{1}{a}$  is able to strike a balance between both the requirements (dropping of the term  $F_1$  in (56) and satisfying the bound on  $\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2$  from (21)). Observe

that by setting  $\chi = \frac{1}{a}$ , we can get rid of the  $\epsilon$  dependency in the numerator of (70) which generates the order bound  $\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \gtrsim \Theta \left( \frac{1}{\log(\epsilon^{-1})} \right)$  that agrees with the condition

$\sqrt{\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2} > \mathcal{O} \left( \frac{1}{\sqrt{\epsilon}} \left( \log \left( \frac{1}{\epsilon} \right) \epsilon \right) \right)$  from (21) for any  $a > 0$ . Also, it can be easily checked that the term  $F_1$  from (56) for  $K_0 = \chi \hat{K}_0 = \frac{1}{a} \hat{K}_0$  has the order  $\mathcal{O}(\epsilon^{(2+b)} \log(\epsilon^{-1}))$  for some  $b > 0$  hence the term  $F_1$  can be dropped to get the approximation (57). Substituting  $\hat{K}_0$  from (61) and  $\chi = \frac{1}{a}$  in (70) followed by further simplification gives the following

result:

$$\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \gtrsim \frac{\left(2 + \frac{\varepsilon M}{2L}\right) \left( \frac{2\delta \sqrt{\mu} \log\left(1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}\right)}{Mn} + \frac{\log\left(1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}\right)}{\left(\frac{\beta}{L} - \frac{\varepsilon M}{2L}\right)} - \frac{\log\left(2 + \frac{\varepsilon M}{2L}\right)}{\left(\frac{\beta}{L} - \frac{\varepsilon M}{2L}\right)} \right)}{\frac{1}{a} \log\left( \frac{2\delta \left(2 + \frac{\varepsilon M}{2L}\right) \log\left(1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}\right) \log\left(\frac{2 + \frac{\varepsilon M}{2L}}{1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}}\right)}{\varepsilon M n \log\left(2 + \frac{\varepsilon M}{2L}\right)} \right) + 1} \quad (71)$$

Finally, for  $\varepsilon < \frac{2\beta}{M}$ , dropping the negative term  $\frac{\log\left(1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}\right)}{\left(\frac{\beta}{L} - \frac{\varepsilon M}{2L}\right)} - \frac{\log\left(2 + \frac{\varepsilon M}{2L}\right)}{\left(\frac{\beta}{L} - \frac{\varepsilon M}{2L}\right)}$  from the numerator of (71) and setting the condition:

$$\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \gtrsim \frac{\left(2 + \frac{\varepsilon M}{2L}\right) \left( \frac{2\delta \sqrt{\mu} \log\left(1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}\right)}{Mn} \right)}{\frac{1}{a} \log\left( \frac{2\delta \left(2 + \frac{\varepsilon M}{2L}\right) \log\left(1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}\right) \log\left(\frac{2 + \frac{\varepsilon M}{2L}}{1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}}\right)}{\varepsilon M n \log\left(2 + \frac{\varepsilon M}{2L}\right)} \right) + 1}, \quad (72)$$

the approximate lower bound in (71) is guaranteed. Now using the upper bound on  $K_0$  from (61) in the expression  $\mu \varepsilon^a = \frac{1}{\left(2 + \frac{\varepsilon M}{2L}\right)^{2\hat{K}_0}}$ , we have that:

$$\sqrt[a]{\mu} = \frac{Mn \log\left(2 + \frac{\varepsilon M}{2L}\right)}{2\delta \left(2 + \frac{\varepsilon M}{2L}\right) \log\left(1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}\right) \log\left(\frac{2 + \frac{\varepsilon M}{2L}}{1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}}\right)} \quad (73)$$

where  $a = \frac{\log\left(2 + \frac{\varepsilon M}{2L}\right)}{\log\left(2 + \frac{\varepsilon M}{2L}\right) - \log\left(1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}\right)} > 1$ . Hence, the approximate lower bound on the unstable projection value  $\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2$  has the following order:

$$\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \gtrsim \Theta\left(\frac{1}{\log(\varepsilon^{-1})}\right). \quad (74)$$

It is also worth mentioning that the lower bound on the unstable projection value  $\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2$  from (74) is an increasing function of  $\varepsilon$ . ■

#### Proof of Theorem 1

Using Lemmas 14, 15 and 16 we have established that there exists some  $K_0 = \mathcal{O}(\log(\varepsilon^{-1}))$  in the set  $G_{\Psi}$  such that  $\underline{\Psi}(K_0) \geq 1$  provided the initial condition of (74) holds. Since  $K_0 \in G_{\Psi}$  we will have  $K_0 \leq K^l$  where  $K^l$  is upper bounded by the linear exit time bound from (7). Then using the fact that  $\underline{\Psi}(K_0) \geq$

1 we get that  $\Psi(K_0) > \underline{\Psi}(K_0) \geq 1$  implying  $\inf_{\tau} \|\tilde{\mathbf{u}}_{K_0}^{\tau}\| > \varepsilon^2 \Psi(K_0) > \varepsilon^2$  from (44). Hence the approximate trajectories  $\{\tilde{\mathbf{u}}_K^{\tau}\}$  exit  $\mathcal{B}_{\varepsilon}(\mathbf{x}^*)$  at  $K < K_0 < K^l$  under the sufficient initial condition of (74). This completes the proof of Theorem 1. ■

Finally, using the fact that  $\varepsilon < \frac{2\beta}{M}$  and Theorem 3.2 of [10], we can upper bound  $\varepsilon$  as follows:

$$\varepsilon < \min \left\{ \inf_{\|\mathbf{u}\|=1} \left( \limsup_{j \rightarrow \infty} \sqrt[j]{\frac{r_j(\mathbf{u})}{j!}} \right)^{-1}, \frac{2L\delta}{M(2Ln^2 - \delta)} + \mathcal{O}(\varepsilon^2), \frac{2\beta}{M} \right\} \quad (75)$$

where  $r_j(\mathbf{u}) = \left\| \left( \frac{d^j}{dw^j} \nabla^2 f(\mathbf{x}^* + w\mathbf{u}) \right) \Big|_{w=0} \right\|_2$ . ■

#### APPENDIX B

We prove Theorem 2 by first proving a sequence of lemmas.

**Lemma 17.** For an iterative gradient mapping given by  $\mathbf{x}^+ = \mathbf{x} - \alpha \nabla f(\mathbf{x})$  in some neighborhood of  $\mathbf{x}^*$ , if  $\|\mathbf{x}^+ - \mathbf{x}^*\| > \|\mathbf{x} - \mathbf{x}^*\|$  then the following holds:

$$\mathbf{a.} \quad \|\mathbf{x}^{++} - \mathbf{x}^*\| \geq \bar{\rho}(\mathbf{x}) \|\mathbf{x}^+ - \mathbf{x}^*\| - \sigma(\mathbf{x}) \quad (76)$$

$$\mathbf{b.} \quad \|\mathbf{x}^{++} - \mathbf{x}^*\| > \|\mathbf{x}^+ - \mathbf{x}^*\| \quad (77)$$

where  $\sigma(\mathbf{x}) = \mathcal{O}(\|\mathbf{x} - \mathbf{x}^*\|^2)$ ,  $\bar{\rho}(\mathbf{x}) > 1$  and (77) is termed as the sequential monotonicity property.

*Proof.* For an iterative gradient mapping given by  $\mathbf{x}^+ = \mathbf{x} - \alpha \nabla f(\mathbf{x})$  in any neighborhood of  $\mathbf{x}^*$ , we have:

$$\nabla f(\mathbf{x}) = \left( \nabla f(\mathbf{x}^*) + \int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) (\mathbf{x} - \mathbf{x}^*) dp \right). \quad (78)$$

provided function  $f(\cdot)$  is twice continuously differentiable. Using this substitution in the iterative gradient mapping, we have the following result:

$$\|\mathbf{x}^+ - \mathbf{x}^*\| = \|\mathbf{x} - \alpha \nabla f(\mathbf{x}) - \mathbf{x}^*\| \quad (79)$$

$$= \left\| (\mathbf{x} - \mathbf{x}^*) - \alpha \left( \nabla f(\mathbf{x}^*) + \int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) (\mathbf{x} - \mathbf{x}^*) dp \right) \right\| \quad (80)$$

$$= \left\| (\mathbf{x} - \mathbf{x}^*) - \alpha \int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) dp (\mathbf{x} - \mathbf{x}^*) \right\| \quad (81)$$

$$= \left\| \left( \mathbf{I} - \alpha \int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) dp \right) (\mathbf{x} - \mathbf{x}^*) \right\| \quad (82)$$

$$= \sqrt{\left( \sum_{j \in \mathcal{J}_{US}} (v_j^{us} \langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} (v_i^s \langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2 \right) \|\mathbf{x} - \mathbf{x}^*\|^2} \quad (83)$$

where  $\mathbf{u} = \mathbf{x} - \mathbf{x}^*$ ,  $\hat{\mathbf{u}} = \frac{\mathbf{u}}{\|\mathbf{u}\|}$ ,  $\mathbf{x} - \mathbf{x}^* = \|\mathbf{u}\| \left( \sum_{j \in \mathcal{J}_{US}} \langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle \mathbf{e}_j^{us} + \sum_{i \in \mathcal{J}_S} \langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle \mathbf{e}_i^s \right)$  and  $(\mathbf{e}_j^{us}, v_j^{us})$ ,  $(\mathbf{e}_i^s, v_i^s)$  are the eigenvector-eigenvalue pair of the matrix  $\mathbf{D}(\mathbf{x})$  where  $\mathbf{D}(\mathbf{x}) = \left( \mathbf{I} - \alpha \int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) dp \right)$  with  $v_i^s < 1$  for all  $i \in \mathcal{J}_S$ ,

$$\|\mathbf{x}^{++} - \mathbf{x}^*\| = \|\mathbf{x}^+ - \alpha \nabla f(\mathbf{x}^+) - \mathbf{x}^*\| \quad (86)$$

$$= \left\| (\mathbf{x}^+ - \mathbf{x}^*) - \alpha \left( \nabla f(\mathbf{x}^*) + \int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x}^+ - \mathbf{x}^*)) (\mathbf{x}^+ - \mathbf{x}^*) dp \right) \right\| \quad (87)$$

$$= \left\| (\mathbf{x}^+ - \mathbf{x}^*) - \alpha \int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x}^+ - \mathbf{x}^*)) dp (\mathbf{x}^+ - \mathbf{x}^*) \right\| \quad (88)$$

$$= \left\| \left( \mathbf{I} - \alpha \int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x}^+ - \mathbf{x}^*)) dp \right) (\mathbf{x}^+ - \mathbf{x}^*) \right\| \quad (89)$$

$$= \left\| \left( \mathbf{I} - \alpha \int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x}^+ - \mathbf{x}^*)) dp \right) \left( \mathbf{I} - \alpha \int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) dp \right) (\mathbf{x} - \mathbf{x}^*) \right\| \quad (90)$$

$$= \left\| \left( \mathbf{I} - \alpha \int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) dp - \alpha \mathbf{P}(\mathbf{x}) \right) \left( \mathbf{I} - \alpha \int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) dp \right) (\mathbf{x} - \mathbf{x}^*) \right\| \quad (91)$$

$v_j^{us} \geq 1$  for all  $j \in \mathcal{J}_{US}$  and  $\mathcal{J}_{US}, \mathcal{J}_S$  are the index sets associated respectively with these subspaces respectively.

We consider the case of strict expansive dynamics in the current iteration. Given:  $\|\mathbf{x}^+ - \mathbf{x}^*\| > \|\mathbf{x} - \mathbf{x}^*\|$  or equivalently

$$\begin{aligned} \|\mathbf{x}^+ - \mathbf{x}^*\| &= \sqrt{\left( \sum_{j \in \mathcal{J}_{US}} (v_j^{us} \langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} (v_i^s \langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2 \right)} \|\mathbf{u}\| \\ &> \|\mathbf{u}\|. \end{aligned} \quad (84)$$

This implies:

$$\sqrt{\left( \sum_{j \in \mathcal{J}_{US}} (v_j^{us} \langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} (v_i^s \langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2 \right)} > 1. \quad (85)$$

We now show that the claim in (76) holds.

Since  $\mathbf{x}^{++} = \mathbf{x}^+ - \alpha \nabla f(\mathbf{x}^+)$ , we have the equations (86)–(91), where in the last step (91) we used the following result:

$$\int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x}^+ - \mathbf{x}^*)) dp = \int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) dp + \mathbf{P}(\mathbf{x}). \quad (92)$$

and we have that  $\|\mathbf{P}(\mathbf{x})\| = \mathcal{O}(\|\nabla f(\mathbf{x})\|)$  which can be verified from Assumption A3. Rearranging (92) and taking norm both sides we get:

$$\begin{aligned} \|\mathbf{P}(\mathbf{x})\|_2 &= \left\| \int_{p=0}^{p=1} \left( \nabla^2 f(\mathbf{x}^* + p(\mathbf{x}^+ - \mathbf{x}^*)) - \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) \right) dp \right\|_2 \\ &\leq \int_{p=0}^{p=1} \left\| \left( \nabla^2 f(\mathbf{x}^* + p(\mathbf{x}^+ - \mathbf{x}^*)) - \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) \right) \right\|_2 dp \end{aligned} \quad (93)$$

$$\leq \int_{p=0}^{p=1} M \|p(\mathbf{x}^+ - \mathbf{x})\| dp \quad (94)$$

$$\leq \int_{p=0}^{p=1} M \|p(\mathbf{x}^+ - \mathbf{x})\| dp \quad (95)$$

$$= M \|\mathbf{x}^+ - \mathbf{x}\| \int_{p=0}^{p=1} p dp = \frac{M\alpha \|\nabla f(\mathbf{x})\|}{2}. \quad (96)$$

Now recall that  $\mathbf{D}(\mathbf{x}) = \left( \mathbf{I} - \alpha \int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) dp \right)$  hence further simplifying (91) yields the following:

$$\|\mathbf{x}^{++} - \mathbf{x}^*\| = \left\| \left( \mathbf{D}(\mathbf{x}) \right)^2 (\mathbf{x} - \mathbf{x}^*) - \alpha \left( \mathbf{D}(\mathbf{x}) \mathbf{P}(\mathbf{x}) (\mathbf{x} - \mathbf{x}^*) \right) \right\| \quad (97)$$

$$\begin{aligned} &\geq \sqrt{\left( \sum_{j \in \mathcal{J}_{US}} (v_j^{us})^4 (\langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} (v_i^s)^4 (\langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2 \right)} \|\mathbf{x} - \mathbf{x}^*\| \\ &\quad - \alpha \|\mathbf{D}(\mathbf{x})\|_2 \|\mathbf{P}(\mathbf{x})\|_2 \|\mathbf{x} - \mathbf{x}^*\| \end{aligned} \quad (98)$$

$$\begin{aligned} &\geq \sqrt{\left( \sum_{j \in \mathcal{J}_{US}} (v_j^{us})^4 (\langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} (v_i^s)^4 (\langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2 \right)} \|\mathbf{x} - \mathbf{x}^*\| \\ &\quad - \frac{\sup_j \{v_j^{us}\} M\alpha \|\nabla f(\mathbf{x})\| \|\mathbf{x} - \mathbf{x}^*\|}{2} \end{aligned} \quad (99)$$

$$\begin{aligned} &\geq \sqrt{\left( \sum_{j \in \mathcal{J}_{US}} (v_j^{us})^4 (\langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} (v_i^s)^4 (\langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2 \right)} \|\mathbf{x} - \mathbf{x}^*\| \\ &\quad - \frac{\sup_j \{v_j^{us}\} M L \alpha \|\mathbf{x} - \mathbf{x}^*\|^2}{2} \end{aligned} \quad (100)$$

where we used the fact that  $\|\nabla f(\mathbf{x})\| \leq L \|\mathbf{x} - \mathbf{x}^*\|$  by Lipschitz continuity of  $\nabla f(\mathbf{x})$ . Now with  $\sigma(\mathbf{x}) = \frac{\sup_j \{v_j^{us}\} M L \alpha \|\mathbf{x} - \mathbf{x}^*\|^2}{2} = \mathcal{O}(\|\mathbf{x} - \mathbf{x}^*\|^2)$  we are left to prove:

$$\begin{aligned} &\sqrt{\left( \sum_{j \in \mathcal{J}_{US}} (v_j^{us})^4 (\langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} (v_i^s)^4 (\langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2 \right)} \|\mathbf{x} - \mathbf{x}^*\| \\ &> \|\mathbf{x}^+ - \mathbf{x}^*\| \end{aligned} \quad (101)$$

or equivalently the following result:

$$\begin{aligned} &\sqrt{\left( \sum_{j \in \mathcal{J}_{US}} (v_j^{us})^4 (\langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} (v_i^s)^4 (\langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2 \right)} \|\mathbf{u}\| > \\ &\sqrt{\left( \sum_{j \in \mathcal{J}_{US}} (v_j^{us} \langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} (v_i^s \langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2 \right)} \|\mathbf{u}\| \end{aligned} \quad (102)$$

$$\begin{aligned} &\sqrt{\left( \sum_{j \in \mathcal{J}_{US}} (v_j^{us})^4 (\langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} (v_i^s)^4 (\langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2 \right)} > \\ &\sqrt{\left( \sum_{j \in \mathcal{J}_{US}} (v_j^{us} \langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} (v_i^s \langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2 \right)}. \end{aligned} \quad (103)$$

This will hold true if:

$$\sqrt{\left(\sum_{j \in \mathcal{J}_{US}} (v_j^{us} \langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} (v_i^s \langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2\right)} > 1. \quad (104)$$

Recall that  $(\mathbf{v}_j^{us}, \mathbf{e}_j^{us})$ ,  $(\mathbf{e}_i^s, \mathbf{v}_i^s)$  are respectively the eigenvector-eigenvalue pair of the matrix  $\mathbf{D}(\mathbf{x}) = \left(\mathbf{I} - \alpha \int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) dp\right)$  with  $v_i^s < 1$  for all  $i \in \mathcal{J}_S$ ,  $v_j^{us} \geq 1$  for all  $j \in \mathcal{J}_{US}$ . Then the condition (103) can be written as:

$$\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} > \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle} \quad (105)$$

$$\implies \langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle > \langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle \quad (106)$$

where  $\hat{\mathbf{u}}$  is a unit vector. Also we are given (85) that can be written as:

$$\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle} > 1 = \sqrt{\langle \hat{\mathbf{u}}, \hat{\mathbf{u}} \rangle} \quad (107)$$

$$\implies \langle \hat{\mathbf{u}}, ((\mathbf{D}(\mathbf{x}))^2 - \mathbf{I}) \hat{\mathbf{u}} \rangle > 0. \quad (108)$$

Now consider the following difference:

$$\begin{aligned} \langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle - \langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle &= \underbrace{\langle \hat{\mathbf{u}}, ((\mathbf{D}(\mathbf{x}))^2 - \mathbf{I})^2 \hat{\mathbf{u}} \rangle}_{\geq 0} + \\ &\quad \underbrace{\langle \hat{\mathbf{u}}, ((\mathbf{D}(\mathbf{x}))^2 - \mathbf{I}) \hat{\mathbf{u}} \rangle}_{> 0} > 0 \end{aligned} \quad (109)$$

$$\implies \langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle > \langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle \quad (110)$$

which completes the proof for (103). We are now ready to prove the result  $\|\mathbf{x}^{++} - \mathbf{x}^*\| \geq \bar{\rho}(\mathbf{x}) \|\mathbf{x}^+ - \mathbf{x}^*\| - \sigma(\mathbf{x})$ . Recall that from (83) we have that:

$$\|\mathbf{x}^+ - \mathbf{x}^*\| = \sqrt{\left(\sum_{j \in \mathcal{J}_{US}} (v_j^{us} \langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} (v_i^s \langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2\right)} \|\mathbf{x} - \mathbf{x}^*\| \quad (111)$$

$$= \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle} \|\mathbf{x} - \mathbf{x}^*\| \quad (112)$$

Now from (100) we have the following:

$$\begin{aligned} \|\mathbf{x}^{++} - \mathbf{x}^*\| &\geq \\ &\sqrt{\left(\sum_{j \in \mathcal{J}_{US}} (v_j^{us})^4 (\langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} (v_i^s)^4 (\langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2\right)} \|\mathbf{x} - \mathbf{x}^*\| \\ &\quad - \mathcal{O}(\|\mathbf{x} - \mathbf{x}^*\|^2) \end{aligned} \quad (113)$$

$$= \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} \|\mathbf{x} - \mathbf{x}^*\| - \sigma(\mathbf{x}) \quad (114)$$

$$= \frac{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle}}{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}} \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle} \|\mathbf{x} - \mathbf{x}^*\| - \sigma(\mathbf{x}) \quad (115)$$

$$= \frac{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle}}{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}} \|\mathbf{x}^+ - \mathbf{x}^*\| - \sigma(\mathbf{x}) \quad (116)$$

where in the last step we used the substitution from (112). Next, note that  $\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle > \langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle > 1$  and hence we can set  $\bar{\rho}(\mathbf{x}) = \frac{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle}}{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}} > 1$  to complete the proof.

Next, we show that the claim in (77) holds, i.e.,  $\|\mathbf{x}^{++} - \mathbf{x}^*\| > \|\mathbf{x}^+ - \mathbf{x}^*\|$  provided  $\|\mathbf{x} - \mathbf{x}^*\|$  is bounded above.

It can be done using (76) of the result where we lower bound the right hand side of (76) with  $\|\mathbf{x}^+ - \mathbf{x}^*\|$  to get:

$$\|\mathbf{x}^{++} - \mathbf{x}^*\| \geq \bar{\rho}(\mathbf{x}) \|\mathbf{x}^+ - \mathbf{x}^*\| - \sigma(\mathbf{x}) > \|\mathbf{x}^+ - \mathbf{x}^*\| \quad (117)$$

$$\implies (\bar{\rho}(\mathbf{x}) - 1) \|\mathbf{x}^+ - \mathbf{x}^*\| > \sigma(\mathbf{x}). \quad (118)$$

Since  $\sigma(\mathbf{x}) = \mathcal{O}(\|\mathbf{x} - \mathbf{x}^*\|^2)$ , hence  $\|\mathbf{x} - \mathbf{x}^*\|$  should be sufficiently small for (118) to hold. Now, if (118) condition holds true, then we will have the condition

$$\|\mathbf{x}^{++} - \mathbf{x}^*\| \geq \bar{\rho}(\mathbf{x}) \|\mathbf{x}^+ - \mathbf{x}^*\| - \sigma(\mathbf{x}) > \|\mathbf{x}^+ - \mathbf{x}^*\|$$

or equivalently  $\|\mathbf{x}^{++} - \mathbf{x}^*\| > \|\mathbf{x}^+ - \mathbf{x}^*\|$ . Next, for some  $k = K$  let  $\mathbf{x} = \mathbf{x}_K$ ,  $\mathbf{x}^+ = \mathbf{x}_{K+1}$ ,  $\mathbf{x}^{++} = \mathbf{x}_{K+2}$  and we have  $\|\mathbf{x}_{K+1} - \mathbf{x}^*\| > \|\mathbf{x}_K - \mathbf{x}^*\|$  with the condition (118) satisfied, then we also have  $\|\mathbf{x}_{K+2} - \mathbf{x}^*\| > \|\mathbf{x}_{K+1} - \mathbf{x}^*\|$ . Using induction, we then get  $\|\mathbf{x}_{k+1} - \mathbf{x}^*\| > \|\mathbf{x}_k - \mathbf{x}^*\|$  for all  $k \geq K + 1$  provided (118) holds true with  $\mathbf{x} = \mathbf{x}_k$ . ■

Hence, the claim of sequential monotonicity has been proved partially, i.e., if a gradient trajectory has expansive dynamics w.r.t. stationary point  $\mathbf{x}^*$  at some  $k = K$ , then it has expansive dynamics for all iterations  $k > K$  provided  $\|\mathbf{x}_k - \mathbf{x}^*\|$  remains bounded above.<sup>11</sup> Now, we are only left with proving the complete claim, i.e., sequential monotonicity holds even if the gradient trajectory has non-contraction dynamics w.r.t. stationary point  $\mathbf{x}^*$  at some  $k = K$ . Before completing the proof of this claim, we need to do provide a bound on the expansion factor  $\bar{\rho}(\mathbf{x})$ .

**Lemma 18.** The expansion factor  $\bar{\rho}(\mathbf{x})$  in (76) is bounded as

$$\bar{\rho}(\mathbf{x}) > 1 + \frac{\left((1 + \frac{\beta}{L})^2 + \frac{1}{4(1 + \frac{\beta}{L})^2} - \frac{5}{4}\right)}{12}.$$

*Proof.* From the condition (118), we require  $\sigma(\mathbf{x})$  to be upper bounded. Notice that the upper bound on  $\sigma(\mathbf{x})$  goes to 0 as  $\bar{\rho}(\mathbf{x})$  approaches 1. Then, the particular theorem cannot be applied recursively since  $\sigma(\mathbf{x})$  is a positive quantity that comes from (100) and (118) would then fail to hold. Hence, in order to exploit the property (118), we require  $\bar{\rho}(\mathbf{x})$  to be bounded away from 1. Using (112) in (118) and simplifying  $\bar{\rho}(\mathbf{x})$ , we get that:

$$(\bar{\rho}(\mathbf{x}) - 1) \|\mathbf{x}^+ - \mathbf{x}^*\| > \sigma(\mathbf{x}) \quad (119)$$

$$\implies \left(\frac{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle}}{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}} - 1\right) \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle} \|\mathbf{x} - \mathbf{x}^*\| > \sigma(\mathbf{x}) \quad (120)$$

$$\implies \left(\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} - \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}\right) \|\mathbf{x} - \mathbf{x}^*\| > \sigma(\mathbf{x}) \quad (121)$$

<sup>11</sup>Notice that  $\mathbf{x}^*$  can be any stationary point and not just the strict saddle point. Since the stationary points of the function are non-degenerate from our assumptions, the extension of this proof to other types of stationary points is left as an easy exercise to the reader.

where we require the term  $\left(\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} - \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}\right)$  to be bounded away from 0. This will hold true due to the following fact:

$$\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} = \sqrt{\left(\sum_{j \in \mathcal{J}_{US}} (\mathbf{v}_j^{us})^4 (\langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} (\mathbf{v}_i^s)^4 (\langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2\right)} \quad (122)$$

$$\geq \left(\sum_{j \in \mathcal{J}_{US}} \sqrt{(\mathbf{v}_j^{us})^4 (\langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2} + \sum_{i \in \mathcal{J}_S} \sqrt{(\mathbf{v}_i^s)^4 (\langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2}\right) \quad (123)$$

$$= \langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle > \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle} \quad (124)$$

where we used Jensen's inequality for square root function followed by the fact that  $\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle > 1$ . But in order to develop a bound on the radius of ball inside which sequential monotonicity holds, we require something more. Notice that if we plug in the naive lower bound just obtained into (121), all we can get is a projection dependent term which does not generalize to the class of functions being studied. The goal here is to obtain some bound that is independent of  $\hat{\mathbf{u}}$  and solely depends on the function parameters like condition number, etc. The next steps develop a generalized lower bound for  $\left(\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} - \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}\right)$  independent of  $\hat{\mathbf{u}}$ .

Since we have  $\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle} > 1$ , we can write:

$$\begin{aligned} & \left(\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} - \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}\right) \\ &= \frac{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle - \langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} + \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}} \end{aligned} \quad (125)$$

where we require  $\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle > \langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle$ . Next, substituting  $\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle = \left(\sum_{j \in \mathcal{J}_{US}} (\mathbf{v}_j^{us})^4 (\langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} (\mathbf{v}_i^s)^4 (\langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2\right)$  and  $\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle = \left(\sum_{j \in \mathcal{J}_{US}} (\mathbf{v}_j^{us})^2 (\langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} (\mathbf{v}_i^s)^2 (\langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2\right)$  in the left-hand side of (125) followed by simplification yields:

$$\begin{aligned} & \frac{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle - \langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} + \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}} = \\ & \frac{\sum_{j \in \mathcal{J}_{US}} \left((\mathbf{v}_j^{us})^4 - (\mathbf{v}_j^{us})^2\right) (\langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2}{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} + \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}} \\ & + \frac{\sum_{i \in \mathcal{J}_S} \left((\mathbf{v}_i^s)^4 - (\mathbf{v}_i^s)^2\right) (\langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2}{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} + \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}}. \end{aligned} \quad (126)$$

Recall that  $\mathbf{D}(\mathbf{x}) = \left(\mathbf{I} - \alpha \int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) dp\right)$ , hence for any eigenvalue  $v_l$  of the matrix  $\mathbf{D}(\mathbf{x})$  where  $v_l = 1 - \alpha \lambda_l \left(\int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) dp\right)$  and  $1 \leq l \leq n$  with  $v_l \geq 0$

and  $\lambda_l$  is the corresponding eigenvalue of  $\int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) dp$ , we have that:

$$\left\| \left( \int_{p=0}^{p=1} \left( \mathbf{I} - \alpha \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) \right) dp \right)^{-1} \right\|_2^{-1} \leq v_l \leq \left\| \int_{p=0}^{p=1} \left( \mathbf{I} - \alpha \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) \right) dp \right\|_2 \quad (127)$$

$$\int_{p=0}^{p=1} \left\| \left( \mathbf{I} - \alpha \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) \right)^{-1} \right\|_2^{-1} dp \leq v_l \leq \int_{p=0}^{p=1} \left\| \left( \mathbf{I} - \alpha \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) \right) \right\|_2 dp \quad (128)$$

$$\begin{aligned} & 1 - \alpha \int_{p=0}^{p=1} \sup_l \lambda_l (\nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*))) dp \leq v_l \leq \\ & 1 - \alpha \int_{p=0}^{p=1} \inf_l \lambda_l (\nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*))) dp. \end{aligned} \quad (129)$$

Therefore, the bounds on  $v_i^s$  and  $v_j^{us}$  for  $\alpha = \frac{1}{L}$  can be given by:

$$\begin{aligned} & 1 - \alpha \int_{p=0}^{p=1} \sup_{\lambda_l > 0} \lambda_l (\nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*))) dp \leq v_i^s \leq \\ & 1 - \alpha \int_{p=0}^{p=1} \inf_{\lambda_l > 0} \lambda_l (\nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*))) dp \end{aligned} \quad (130)$$

$$1 - \alpha \int_{p=0}^{p=1} L dp \leq v_i^s \leq 1 - \alpha \int_{p=0}^{p=1} \beta dp \quad (131)$$

$$0 \leq v_i^s \leq 1 - \frac{\beta}{L} \quad (132)$$

$$\begin{aligned} & 1 - \alpha \int_{p=0}^{p=1} \sup_{\lambda_l < 0} \lambda_l (\nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*))) dp \leq v_j^{us} \leq \\ & 1 - \alpha \int_{p=0}^{p=1} \inf_{\lambda_l < 0} \lambda_l (\nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*))) dp \end{aligned} \quad (133)$$

$$1 - \alpha \int_{p=0}^{p=1} -\beta dp \leq v_j^{us} \leq 1 - \alpha \int_{p=0}^{p=1} -L dp \quad (134)$$

$$1 + \frac{\beta}{L} \leq v_j^{us} \leq 2 \quad (135)$$

where we used the fact that  $\inf_l |\lambda_l (\nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)))| > \beta$ , i.e., the minimum absolute eigenvalue of the function  $f(\cdot)$  in a neighborhood of  $\mathbf{x}^*$  is greater than  $\beta$  from Assumption A4. Also, we used  $\sup_l |\lambda_l (\nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)))| \leq L$ , from Assumption A2.

Hence, the right-hand-side (R.H.S.) in (126) can be lower bounded as:

$$\begin{aligned} & \frac{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle - \langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} + \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}} = \\ & \frac{\sum_{j \in \mathcal{J}_{US}} \left((\mathbf{v}_j^{us})^4 - (\mathbf{v}_j^{us})^2\right) (\langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2}{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} + \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}} \\ & + \frac{\sum_{i \in \mathcal{J}_S} \left((\mathbf{v}_i^s)^4 - (\mathbf{v}_i^s)^2\right) (\langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2}{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} + \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}} \geq \end{aligned} \quad (136)$$

$$\frac{\left(\sum_{j \in \mathcal{J}_{US}} \left( (1 + \frac{\beta}{L})^4 - (1 + \frac{\beta}{L})^2 \right) (\langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 - \frac{1}{4} \sum_{i \in \mathcal{J}_S} (\langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2 \right)}{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} + \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}} \quad (137)$$

where in the last step we used the fact that the maximum eigenvalue of  $(\mathbf{D}(\mathbf{x}))^2$  is 4 which implies  $\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle} < 2$  and  $\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} < 4$ .

Now, it can be verified that for values of  $\frac{\beta}{L} > 0$ , the right-hand side of (140) is bounded away from 0. Since  $\bar{\rho}(\mathbf{x}) = \frac{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle}}{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}}$ , then using (140) and  $\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle} < 2$  we can write

$$\bar{\rho}(\mathbf{x}) = 1 + \frac{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} - \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}}{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}} > 1 + \frac{\left( (1 + \frac{\beta}{L})^2 + \frac{1}{4(1 + \frac{\beta}{L})^2} - \frac{5}{4} \right)}{12} \quad (141)$$

which is an expansion factor for any  $\frac{\beta}{L} > 0$ . ■

where we used the fact that  $v_j^{us} \geq (1 + \frac{\beta}{L})$  and  $\left( (v_i^s)^4 - (v_i^s)^2 \right) \geq -\frac{1}{4}$  for  $v_i^s < 1$  (minimum of  $h(y) = y^4 - y^2$  for  $0 \leq y < 1$  is  $-\frac{1}{4}$ ).

$$\sum_{j \in \mathcal{J}_{US}} (\langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} (\langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2 = 1$$

and

$$\begin{aligned} \sum_{j \in \mathcal{J}_{US}} (v_j^{us} \langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} (v_i^s \langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2 &> 1 \iff \\ \sum_{j \in \mathcal{J}_{US}} ((v_j^{us})^2 - 1) (\langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} ((v_i^s)^2 - 1) (\langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2 &> 0 \end{aligned} \quad (138)$$

Next we minimize the numerator of the R.H.S. in (137) in a way so as to get rid of the dependency on  $\hat{\mathbf{u}}$ . Recall that the minimization of  $\left( \sum_{j \in \mathcal{J}_{US}} \left( (1 + \frac{\beta}{L})^4 - (1 + \frac{\beta}{L})^2 \right) (\langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 - \frac{1}{4} \sum_{i \in \mathcal{J}_S} (\langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2 \right)$  is constrained by

We now extend the claim of Lemma 17 to the case of non-contraction, i.e.,  $\|\mathbf{x}^+ - \mathbf{x}^*\| = \|\mathbf{x} - \mathbf{x}^*\|$ . In words, we show that sequential monotonicity property from (77) holds even if the gradient trajectory has non-contraction dynamics w.r.t. stationary point  $\mathbf{x}^*$  at some  $k = K$ .

**Lemma 19.** *For an iterative gradient mapping given by  $\mathbf{x}^+ = \mathbf{x} - \alpha \nabla f(\mathbf{x})$  in some neighborhood of  $\mathbf{x}^*$ , if  $\|\mathbf{x}^+ - \mathbf{x}^*\| = \|\mathbf{x} - \mathbf{x}^*\|$  then the following holds:*

$$a. \quad \|\mathbf{x}^{++} - \mathbf{x}^*\| \geq \bar{\rho}(\mathbf{x}) \|\mathbf{x}^+ - \mathbf{x}^*\| - \sigma(\mathbf{x}) \quad (142)$$

$$b. \quad \|\mathbf{x}^{++} - \mathbf{x}^*\| > \|\mathbf{x}^+ - \mathbf{x}^*\| \quad (143)$$

where  $\sigma(\mathbf{x}) = \mathcal{O}(\|\mathbf{x} - \mathbf{x}^*\|^2)$  and  $\bar{\rho}(\mathbf{x}) > 1$ .

*Proof.* Notice that while obtaining (140) from (137), we utilized the given condition of (104) according to which we have:

$$\sum_{j \in \mathcal{J}_{US}} (v_j^{us} \langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} (v_i^s \langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2 > 1.$$

This condition implies that we have  $\|\mathbf{x}^+ - \mathbf{x}^*\| > \|\mathbf{x} - \mathbf{x}^*\|$ . However, it could be the case that we have  $\|\mathbf{x}^+ - \mathbf{x}^*\| = \|\mathbf{x} - \mathbf{x}^*\|$  which would imply

$$\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle = \sum_{j \in \mathcal{J}_{US}} (v_j^{us} \langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 + \sum_{i \in \mathcal{J}_S} (v_i^s \langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2 = 1.$$

Using this condition, it can be readily checked that (140) will still hold but only with a non-strict inequality, i.e., we will have:

$$\left( \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} - \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle} \right) \geq \frac{\left( (1 + \frac{\beta}{L})^2 + \frac{1}{4(1 + \frac{\beta}{L})^2} - \frac{5}{4} \right)}{6}.$$

$$\begin{aligned} &\frac{\left( \sum_{j \in \mathcal{J}_{US}} \left( (1 + \frac{\beta}{L})^4 - (1 + \frac{\beta}{L})^2 \right) (\langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 - \frac{1}{4} \sum_{i \in \mathcal{J}_S} (\langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2 \right)}{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} + \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}} > \\ &\frac{\left( (1 + \frac{\beta}{L})^2 + \frac{1}{4(1 + \frac{\beta}{L})^2} - \frac{5}{4} \right)}{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} + \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}} \\ &> \frac{\left( (1 + \frac{\beta}{L})^2 + \frac{1}{4(1 + \frac{\beta}{L})^2} - \frac{5}{4} \right)}{6} \end{aligned} \quad (139)$$

where the second constraint comes from (104). Relaxing the second constraint by using the bounds  $v_i^s \geq 0$ ,  $v_j^{us} \geq (1 + \frac{\beta}{L})$  we get:

$$((1 + \beta/L)^2 - 1) \sum_{j \in \mathcal{J}_{US}} (\langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2 - \sum_{i \in \mathcal{J}_S} (\langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2 > 0.$$

Let  $a = \sum_{j \in \mathcal{J}_{US}} (\langle \hat{\mathbf{u}}, \mathbf{e}_j^{us} \rangle)^2$ ,  $b = \sum_{i \in \mathcal{J}_S} (\langle \hat{\mathbf{u}}, \mathbf{e}_i^s \rangle)^2$  then from the two constraints we have the following minimization problem for the numerator term in (137):

$$\begin{aligned} \min_{a, b \geq 0} &\left( \left( (1 + \frac{\beta}{L})^4 - (1 + \frac{\beta}{L})^2 \right) a - \frac{1}{4} b \right) \quad \text{such that} \\ &a + b = 1, \text{ and} \\ &((1 + \beta/L)^2 - 1)a - b > 0. \end{aligned}$$

Solving this geometrically we obtain that the minimum is attained at the intersection of lines  $a + b = 1$  and  $((1 + \beta/L)^2 - 1)a - b = 0$  which gives  $a = \frac{1}{(1 + \beta/L)^2}$  and  $b = 1 - \frac{1}{(1 + \beta/L)^2}$ .

Substituting  $a, b$  in our function  $\left( \left( (1 + \frac{\beta}{L})^4 - (1 + \frac{\beta}{L})^2 \right) a - \frac{1}{4} b \right)$  yields the following lower bound in (137):

Now since  $\bar{\rho}(\mathbf{x}) = \frac{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle}}{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}} = \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle}$ , we will have that:

$$\begin{aligned} \bar{\rho}(\mathbf{x}) &\geq \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle} + \frac{\left( (1 + \frac{\beta}{L})^2 + \frac{1}{4(1 + \frac{\beta}{L})^2} - \frac{5}{4} \right)}{6} \\ &= 1 + \frac{\left( (1 + \frac{\beta}{L})^2 + \frac{1}{4(1 + \frac{\beta}{L})^2} - \frac{5}{4} \right)}{6} \\ &> 1 + \frac{\left( (1 + \frac{\beta}{L})^2 + \frac{1}{4(1 + \frac{\beta}{L})^2} - \frac{5}{4} \right)}{12}. \end{aligned} \quad (144)$$

Now if  $\sigma(\mathbf{x})$  satisfies the condition (118) for this  $\bar{\rho}(\mathbf{x})$  then we are guaranteed to have  $\|\mathbf{x}^{++} - \mathbf{x}^*\| > \|\mathbf{x}^+ - \mathbf{x}^*\|$  even when  $\|\mathbf{x}^+ - \mathbf{x}^*\| = \|\mathbf{x} - \mathbf{x}^*\|$ . This completes the proof of the claim. ■

Now that we have established the result that if  $\|\mathbf{x}^+ - \mathbf{x}^*\| \geq \|\mathbf{x} - \mathbf{x}^*\|$ , then we are guaranteed to have  $\|\mathbf{x}^{++} - \mathbf{x}^*\| > \|\mathbf{x}^+ - \mathbf{x}^*\|$  provided  $\sigma(\mathbf{x})$  satisfies the condition (118), we can apply this result recursively for any gradient trajectory generated by the sequence  $\{\mathbf{x}_k\}$  in some neighborhood of  $\mathbf{x}^*$ . The next lemma provides a handle on the radius of this neighborhood inside which the sequential monotonicity property holds.

**Lemma 20.** *The sequential monotonicity property from Lemma 17 and 19 holds for the tuple  $\{\mathbf{x}, \mathbf{x}^+, \mathbf{x}^{++}\}$  whenever*

$$\|\mathbf{x} - \mathbf{x}^*\| \leq \frac{1}{\zeta M} \frac{\left( (1 + \frac{\beta}{L})^2 + \frac{1}{4(1 + \frac{\beta}{L})^2} - \frac{5}{4} \right)}{6} \text{ for some } \zeta > 2.$$

*Proof.* To identify the radius of this neighborhood, we use (118) where we substitute  $\sigma(\mathbf{x})$  from (100) and  $\bar{\rho}(\mathbf{x}) = \frac{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle}}{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}}$  to get the condition:

$$\begin{aligned} (\bar{\rho}(\mathbf{x}) - 1) \|\mathbf{x}^+ - \mathbf{x}^*\| &> \sigma(\mathbf{x}) \\ &= \frac{\sup_j \{v_j^{us}\} ML\alpha \|\mathbf{x} - \mathbf{x}^*\|^2}{2} \\ \Rightarrow \left( \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} - \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle} \right) \|\mathbf{x} - \mathbf{x}^*\| &> \sigma(\mathbf{x}) \\ &= \frac{\sup_j \{v_j^{us}\} ML\alpha \|\mathbf{x} - \mathbf{x}^*\|^2}{2}. \end{aligned} \quad (146)$$

Now, in order to guarantee the condition (147), for some  $\zeta > 2$ , we set  $\left( \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle} - \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle} \right)$  equal to  $\frac{1}{\zeta}$  times its lower bound from (140) and set  $\sigma(\mathbf{x})$  to its upper bound in (147) to get the condition:

$$\begin{aligned} \frac{1}{\zeta} \frac{\left( (1 + \frac{\beta}{L})^2 + \frac{1}{4(1 + \frac{\beta}{L})^2} - \frac{5}{4} \right)}{6} \|\mathbf{x} - \mathbf{x}^*\| &\geq \frac{2ML\alpha \|\mathbf{x} - \mathbf{x}^*\|^2}{2} \\ &\geq \frac{\sup_j \{v_j^{us}\} ML\alpha \|\mathbf{x} - \mathbf{x}^*\|^2}{2} = \sigma(\mathbf{x}) \end{aligned} \quad (148)$$

$$\frac{1}{\zeta M} \frac{\left( (1 + \frac{\beta}{L})^2 + \frac{1}{4(1 + \frac{\beta}{L})^2} - \frac{5}{4} \right)}{6} \geq \|\mathbf{x} - \mathbf{x}^*\| \quad (149)$$

where we used  $\alpha = \frac{1}{L}$  and the bound  $\sup_j \{v_j^{us}\} = 1 + \alpha L < 2$ . Now for  $\frac{\beta}{L} > 0$ , if (149) is satisfied then the condition (147) will hold true. Hence any gradient descent trajectory with  $\alpha = \frac{1}{L}$  inside the ball  $\mathcal{B}_\xi(\mathbf{x}^*)$  will exhibit strictly monotonic expansive dynamics once it has a non-contractive dynamics at any instant. ■

Finally combining Lemmas 17, 18, 19 and 20, Theorem 2 is established.

## APPENDIX C PROOF OF LEMMA 1

Before starting the proof of Lemma 1 we first show that unlike the expansion phase of the trajectory where the iterates satisfy strong monotonicity property from (76), the iterates belonging to the contraction phase of the trajectory may not necessarily satisfy such property. From theorem 2 it was established that a gradient trajectory  $\{\mathbf{x}_k\}$  with  $\mathbf{x}_k \in \mathcal{B}_\xi(\mathbf{x}^*)$  has expansive dynamics for all  $k > K$  if at  $k = K$ , the gradient trajectory has non-contraction dynamics<sup>12</sup>. Let there be some  $k = K_\tau$  such that the sequence  $\{\|\mathbf{x}_k - \mathbf{x}^*\|\}$  is strictly decreasing for all  $k \leq K_\tau$  and is non-decreasing for  $k = K_\tau$ . Then from Theorem 2 we have that  $\{\|\mathbf{x}_k - \mathbf{x}^*\|\}$  is strictly increasing for all  $k > K_\tau$  provided  $\mathbf{x}_k \in \mathcal{B}_\xi(\mathbf{x}^*)$ . Since  $\|\mathbf{x}_{K_\tau} - \mathbf{x}^*\|$  is the minimum of the sequence  $\{\|\mathbf{x}_k - \mathbf{x}^*\|\}$  with  $\mathbf{x}_k \in \mathcal{B}_\xi(\mathbf{x}^*)$ , let  $k = K_c$  and  $k = K_e$  be the indices with  $K_c \leq K_\tau \leq K_e$  defined as follows:

$$K_c = \sup \left\{ k \leq K_\tau \mid \mathbf{x}_k \in \bar{\mathcal{B}}_\xi(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*) \right\} \quad (150)$$

$$K_e = \inf \left\{ k \geq K_\tau \mid \mathbf{x}_k \in \bar{\mathcal{B}}_\xi(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*) \right\}. \quad (151)$$

Let the gradient trajectory exit the ball  $\mathcal{B}_\xi(\mathbf{x}^*)$  at some iteration  $\hat{K}_{exit}$ . Then the total sojourn time for the gradient trajectory inside the compact shell  $\bar{\mathcal{B}}_\xi(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*)$  is  $K_c + (\hat{K}_{exit} - K_e)$ .

Since  $K_c \leq K_\tau$ , we have the condition that  $\|\mathbf{x}_k - \mathbf{x}^*\|$  is monotonically decreasing for all  $0 < k \leq K_c$ . However, even with the monotonically decreasing sequence, it cannot be guaranteed that  $\|\mathbf{x}_k - \mathbf{x}^*\|$  will decrease with a geometric rate. This can be checked very easily from (109) in the proof of theorem 2. From that condition, we are guaranteed geometric expansion since the factor  $\bar{\rho}(\mathbf{x}) = \frac{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle}}{\sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle}} > 1$  from the inequality:

$$\begin{aligned} \langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle - \langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle &= \underbrace{\langle \hat{\mathbf{u}}, ((\mathbf{D}(\mathbf{x}))^2 - \mathbf{I})^2 \hat{\mathbf{u}} \rangle}_{\geq 0} + \\ &\quad \underbrace{\langle \hat{\mathbf{u}}, ((\mathbf{D}(\mathbf{x}))^2 - \mathbf{I}) \hat{\mathbf{u}} \rangle}_{> 0} > 0 \end{aligned} \quad (152)$$

provided we have  $\|\mathbf{x}^+ - \mathbf{x}^*\| > \|\mathbf{x} - \mathbf{x}^*\|$  or equivalently  $\langle \hat{\mathbf{u}}, ((\mathbf{D}(\mathbf{x}))^2 - \mathbf{I}) \hat{\mathbf{u}} \rangle > 0$ . Recall that  $\|\mathbf{x}^+ - \mathbf{x}^*\| = \sqrt{\langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle} \|\mathbf{x} - \mathbf{x}^*\|$  from (112). However, when we have

<sup>12</sup>Note that we assume here  $\mathbf{x}_0 \in \bar{\mathcal{B}}_\xi(\mathbf{x}^*) \setminus \mathcal{B}_\xi(\mathbf{x}^*)$ .

$\|\mathbf{x}^+ - \mathbf{x}^*\| < \|\mathbf{x} - \mathbf{x}^*\|$  or equivalently  $\langle \hat{\mathbf{u}}, ((\mathbf{D}(\mathbf{x}))^2 - \mathbf{I})\hat{\mathbf{u}} \rangle < 0$  then (109) becomes:

$$\begin{aligned} \langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^4 \hat{\mathbf{u}} \rangle - \langle \hat{\mathbf{u}}, (\mathbf{D}(\mathbf{x}))^2 \hat{\mathbf{u}} \rangle &= \underbrace{\langle \hat{\mathbf{u}}, ((\mathbf{D}(\mathbf{x}))^2 - \mathbf{I})^2 \hat{\mathbf{u}} \rangle}_{\geq 0} + \\ &\quad \underbrace{\langle \hat{\mathbf{u}}, ((\mathbf{D}(\mathbf{x}))^2 - \mathbf{I}) \hat{\mathbf{u}} \rangle}_{< 0} \leq 0 \end{aligned} \quad (153)$$

and therefore it cannot be stated with certainty that  $\bar{\rho}(\mathbf{x}) < 1$  when we have  $\|\mathbf{x}^+ - \mathbf{x}^*\| < \|\mathbf{x} - \mathbf{x}^*\|$ . Hence, we work with the function value sequence  $\{f(\mathbf{x}_k)\}$  instead of the iterate sequence  $\{\mathbf{x}_k\}$  in order to develop best possible rate of contraction.

We now prove Lemma 1. Taking norm on (78), using the substitution  $\mathbf{G} = \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*))$  followed by taking the lower bound yields:

$$\|\nabla f(\mathbf{x})\| = \left\| \left( \int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) dp \right) (\mathbf{x} - \mathbf{x}^*) \right\| \quad (154)$$

$$\Rightarrow \|\nabla f(\mathbf{x})\| \geq \left\| \left( \int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) dp \right)^{-1} \right\|_2^{-1} \|\mathbf{x} - \mathbf{x}^*\| \quad (155)$$

$$\geq \left( \int_{p=0}^{p=1} \left\| \left( \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) \right)^{-1} \right\|_2^{-1} dp \right) \|\mathbf{x} - \mathbf{x}^*\| \quad (156)$$

$$\Rightarrow \|\nabla f(\mathbf{x})\| \geq \left( \int_{p=0}^{p=1} \lambda_{\min}(\sqrt{\mathbf{G}\mathbf{G}^T}) dp \right) \|\mathbf{x} - \mathbf{x}^*\| \quad (157)$$

$$\Rightarrow \|\nabla f(\mathbf{x})\| \geq \beta \|\mathbf{x} - \mathbf{x}^*\| \quad (158)$$

where we used the fact that  $\lambda_{\min}(\sqrt{\mathbf{G}\mathbf{G}^T}) = \beta$  since  $\lambda_{\min}(\nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*))) = \beta$  for any  $\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*) \in \mathcal{W}$  from Assumption A4.

Next, using gradient Lipschitz condition on  $f(\cdot)$  for  $\mathbf{x}_k$  and  $\mathbf{x}^*$  along with (158) we get:

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{L}{2} \|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq \frac{L}{2\beta^2} \|\nabla f(\mathbf{x}_k)\|^2 \quad (159)$$

where (159) holds for any  $\mathbf{x}_k \in \mathcal{W}$ .

It is important to note that though (159) holds in general for any  $\mathbf{x}_k \in \mathcal{W}$ , yet it cannot be called the Polyak–Łojasiewicz condition [15] when  $\{\mathbf{x}_k\}$  has expansive dynamics locally w.r.t.  $\mathbf{x}^*$  because then  $f(\mathbf{x}_k) - f(\mathbf{x}^*)$  may not be positive. In particular Lemma 4 shows that  $f(\mathbf{x}_{K_{exit}}) < f(\mathbf{x}^*)$  where  $K_{exit}$  is the exit time from the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  so  $f(\mathbf{x}_k) < f(\mathbf{x}^*)$  for all  $k > K_{exit}$  by monotonicity of  $\{f(\mathbf{x}_k)\}$ . Hence (159) becomes trivial in the expansion phase of the trajectory inside the shell  $\mathcal{B}_\varepsilon(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*)$  due to the fact that  $f(\mathbf{x}_k) - f(\mathbf{x}^*) < 0$  for  $k > K_{exit}$ .

Finally it remains to show that  $f(\mathbf{x}_k) - f(\mathbf{x}^*) > 0$  for the contraction phase provided  $K_c < K_e$  so that (159) is indeed the Polyak–Łojasiewicz condition in this case. We accomplish this by lower bounding the term  $f(\mathbf{x}_{K_c}) - f(\mathbf{x}^*)$ . Then  $f(\mathbf{x}_k) - f(\mathbf{x}^*) > 0$  for  $k < K_c$  will follow immediately from

the monotonicity of the sequence  $\{f(\mathbf{x}_k)\}$ . Observe that the trajectory  $\{\mathbf{x}_k\}$  will enter the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  when  $K_c < K_e$  and to do so it has to contract at  $k = K_c$  since  $K_c$  is the last iteration for which the trajectory contracts inside the shell  $\mathcal{B}_\varepsilon(\mathbf{x}^*) \setminus \mathcal{B}_\varepsilon(\mathbf{x}^*)$ . Therefore we have that  $\|\mathbf{x}_{K_c} - \mathbf{x}^*\| > \|\mathbf{x}_{K_c+1} - \mathbf{x}^*\|$ . Further simplifying this condition we get:

$$\|\mathbf{x}_{K_c} - \mathbf{x}^*\|^2 > \|\mathbf{x}_{K_c+1} - \mathbf{x}^*\|^2 \quad (160)$$

$$\Rightarrow \|\mathbf{x}_{K_c} - \mathbf{x}^*\|^2 > \|\mathbf{x}_{K_c} - \alpha \nabla f(\mathbf{x}_{K_c}) - \mathbf{x}^*\|^2 \quad (161)$$

$$\Rightarrow \|\mathbf{x}_{K_c} - \mathbf{x}^*\|^2 > \|\mathbf{x}_{K_c} - \mathbf{x}^*\|^2 + \|\alpha \nabla f(\mathbf{x}_{K_c})\|^2 - 2\langle \alpha \nabla f(\mathbf{x}_{K_c}), \mathbf{x}_{K_c} - \mathbf{x}^* \rangle \quad (162)$$

$$\Rightarrow \langle \mathbf{x}_{K_c} - \mathbf{x}^*, \nabla f(\mathbf{x}_{K_c}) \rangle > \frac{\alpha}{2} \|\nabla f(\mathbf{x}_{K_c})\|^2 \quad (163)$$

$$\Rightarrow \left\langle \mathbf{x}_{K_c} - \mathbf{x}^*, \left( \int_{p=0}^1 \nabla^2 f(\mathbf{x}^* + p(\mathbf{x}_{K_c} - \mathbf{x}^*)) dp \right) (\mathbf{x}_{K_c} - \mathbf{x}^*) \right\rangle > \frac{\alpha}{2} \|\nabla f(\mathbf{x}_{K_c})\|^2 \quad (164)$$

$$\Rightarrow \left\langle \mathbf{x}_{K_c} - \mathbf{x}^*, \nabla^2 f(\mathbf{x}^*)(\mathbf{x}_{K_c} - \mathbf{x}^*) \right\rangle + \frac{M}{2} \|\mathbf{x}_{K_c} - \mathbf{x}^*\|^3 > \frac{\alpha}{2} \|\nabla f(\mathbf{x}_{K_c})\|^2 \quad (165)$$

where we used the substitution  $\nabla f(\mathbf{x}_{K_c}) = \left( \int_{p=0}^1 \nabla^2 f(\mathbf{x}^* + p(\mathbf{x}_{K_c} - \mathbf{x}^*)) dp \right) (\mathbf{x}_{K_c} - \mathbf{x}^*)$  and the following bound:

$$\left\| \left( \int_{p=0}^1 \nabla^2 f(\mathbf{x}^* + p(\mathbf{x}_{K_c} - \mathbf{x}^*)) dp \right) - \nabla^2 f(\mathbf{x}^*) \right\| \leq \int_{p=0}^1 \|\nabla^2 f(\mathbf{x}^* + p(\mathbf{x}_{K_c} - \mathbf{x}^*)) - \nabla^2 f(\mathbf{x}^*)\| dp \quad (166)$$

$$\leq \int_{p=0}^1 M p \|\mathbf{x}_{K_c} - \mathbf{x}^*\| dp = \frac{M}{2} \|\mathbf{x}_{K_c} - \mathbf{x}^*\| \quad (167)$$

in the last step. Using Hessian Lipschitz condition on  $\mathbf{x}_{K_c}$  and  $\mathbf{x}^*$  followed by substituting the bound (165) we have that:

$$f(\mathbf{x}_{K_c}) \geq f(\mathbf{x}^*) + \left\langle \mathbf{x}_{K_c} - \mathbf{x}^*, \nabla^2 f(\mathbf{x}^*)(\mathbf{x}_{K_c} - \mathbf{x}^*) \right\rangle - \frac{M}{6} \|\mathbf{x}_{K_c} - \mathbf{x}^*\|^3 \quad (168)$$

$$\Rightarrow f(\mathbf{x}_{K_c}) - f(\mathbf{x}^*) \geq \frac{\alpha}{2} \|\nabla f(\mathbf{x}_{K_c})\|^2 - \frac{M}{2} \|\mathbf{x}_{K_c} - \mathbf{x}^*\|^3 - \frac{M}{6} \|\mathbf{x}_{K_c} - \mathbf{x}^*\|^3 \quad (169)$$

$$\Rightarrow f(\mathbf{x}_{K_c}) - f(\mathbf{x}^*) \geq \frac{\beta^2}{2L} \|\mathbf{x}_{K_c} - \mathbf{x}^*\|^2 - \frac{2M}{3} \|\mathbf{x}_{K_c} - \mathbf{x}^*\|^3 \quad (170)$$

where in the last step we used (158) and  $\alpha = \frac{1}{L}$ . Hence for  $\|\mathbf{x}_{K_c} - \mathbf{x}^*\| < \frac{3\beta^2}{4ML}$  we will have  $f(\mathbf{x}_{K_c}) - f(\mathbf{x}^*) > 0$ . ■

## APPENDIX D PROOF OF THEOREM 3

We prove Theorem 3 by first upper bounding  $K_c$  and  $\hat{K}_{exit} - K_e$ .

### Bound on $K_c$

Using gradient Lipschitz condition on  $f(\cdot)$  for  $\mathbf{x}_k$  and  $\mathbf{x}_{k+1}$  where  $\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)$  followed by Lemma 1 and inducting from  $k = 0$  to  $k = K_c$  gives:

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq -\frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|^2 \leq -\frac{\beta^2}{L^2}(f(\mathbf{x}_k) - f(\mathbf{x}^*)) \quad (171)$$

$$\Rightarrow f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq \left(1 - \frac{\beta^2}{L^2}\right)(f(\mathbf{x}_k) - f(\mathbf{x}^*)) \quad (172)$$

$$\Rightarrow f(\mathbf{x}_{K_c}) - f(\mathbf{x}^*) \leq \left(1 - \frac{\beta^2}{L^2}\right)^{K_c}(f(\mathbf{x}_0) - f(\mathbf{x}^*)) \quad (173)$$

$$\Rightarrow K_c \leq \frac{\log(f(\mathbf{x}_{K_c}) - f(\mathbf{x}^*)) - \log(f(\mathbf{x}_0) - f(\mathbf{x}^*))}{\log\left(1 - \frac{\beta^2}{L^2}\right)}. \quad (174)$$

By gradient Lipschitz condition for  $\mathbf{x}_0$  and  $\mathbf{x}^*$ , we have the condition:

$$f(\mathbf{x}_0) - f(\mathbf{x}^*) \leq \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^*\| = \frac{L}{2}\xi^2 \quad (175)$$

where we used the fact that the iterate  $\mathbf{x}_0$  sits on the boundary of the ball  $\mathcal{B}_\xi(\mathbf{x}^*)$ . Finally substituting the bounds (175), (170) into (174) yields the following contraction rate:

$$K_c \leq \frac{\log\left(\frac{L}{2}\xi^2\right) - \log\left(\frac{\beta^2}{2L}\|\mathbf{x}_{K_c} - \mathbf{x}^*\|^2 - \frac{2M}{3}\|\mathbf{x}_{K_c} - \mathbf{x}^*\|^3\right)}{\log\left(1 - \frac{\beta^2}{L^2}\right)^{-1}}. \quad (176)$$

Since  $\varepsilon \leq \|\mathbf{x}_{K_c} - \mathbf{x}^*\| < \xi$ , we can further upper bound  $K_c$  as:

$$K_c \leq \frac{\log\left(\frac{L}{2}\xi^2\right) - \log\left(\frac{\beta^2}{2L}\varepsilon^2 - \frac{2M}{3}\varepsilon^3\right)}{\log\left(1 - \frac{\beta^2}{L^2}\right)^{-1}}. \quad (177)$$

Notice that while developing (177) we used Lemma 1 which requires  $K_c < K_e$ . For the case when  $K_c = K_e$  the trajectory never enters the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  and Lemma 1 no longer holds true. However in that case one can repeat the argument from (160) onward in the proof of Lemma 1 by considering  $K_c - 1$  instead of  $K_c$  and get the same upper bound (177) on  $K_c - 1$ . Therefore combining the two cases we effectively get:

$$K_c \leq \frac{\log\left(\frac{L}{2}\xi^2\right) - \log\left(\frac{\beta^2}{2L}\varepsilon^2 - \frac{2M}{3}\varepsilon^3\right)}{\log\left(1 - \frac{\beta^2}{L^2}\right)^{-1}} + 1. \quad (178)$$

The bound on  $\varepsilon$  given by  $\varepsilon < \frac{3\beta^2}{4ML}$  follows from Lemma 1 and the fact that  $\varepsilon \leq \|\mathbf{x}_{K_c} - \mathbf{x}^*\|$ .

### Bound on $\hat{K}_{exit} - K_e$

Recall that from (76) in theorem 2 we have  $\|\mathbf{x}^{++} - \mathbf{x}^*\| > \bar{\rho}(\mathbf{x})\|\mathbf{x}^+ - \mathbf{x}^*\| - \sigma(\mathbf{x})$  whenever  $\|\mathbf{x}^+ - \mathbf{x}^*\| \geq \|\mathbf{x} - \mathbf{x}^*\|$ . Now for  $K_e \leq k \leq \hat{K}_{exit}$ , the sequence  $\{\|\mathbf{x}_k - \mathbf{x}^*\|\}$  is non-decreasing

from the definition of  $K_e$ . Hence, (76) holds for all such  $\mathbf{x}_k$  which have  $K_e \leq k \leq \hat{K}_{exit}$ . Using (76) with  $\mathbf{x}^+ = \mathbf{x}_{k-1}$  and  $\mathbf{x}^{++} = \mathbf{x}_k$  for  $K_e + 1 \leq k \leq \hat{K}_{exit}$  yields:

$$\|\mathbf{x}_k - \mathbf{x}^*\| > \bar{\rho}(\mathbf{x}_{k-2})\|\mathbf{x}_{k-1} - \mathbf{x}^*\| - \sigma(\mathbf{x}_{k-2}) \quad (179)$$

$$\|\mathbf{x}_k - \mathbf{x}^*\| > \bar{\rho}(\mathbf{x}_{k-2})\|\mathbf{x}_{k-1} - \mathbf{x}^*\| - M\|\mathbf{x}_{k-2} - \mathbf{x}^*\|^2 \quad (180)$$

$$\|\mathbf{x}_k - \mathbf{x}^*\| + M\|\mathbf{x}_{k-2} - \mathbf{x}^*\|^2 > \bar{\rho}(\mathbf{x}_{k-2})\|\mathbf{x}_{k-1} - \mathbf{x}^*\| \quad (181)$$

$$\|\mathbf{x}_k - \mathbf{x}^*\| + M\|\mathbf{x}_k - \mathbf{x}^*\|^2 > \bar{\rho}(\mathbf{x}_{k-2})\|\mathbf{x}_{k-1} - \mathbf{x}^*\| \quad (182)$$

$$\|\mathbf{x}_k - \mathbf{x}^*\| > \frac{\bar{\rho}(\mathbf{x}_{k-2})}{1 + M\|\mathbf{x}_k - \mathbf{x}^*\|}\|\mathbf{x}_{k-1} - \mathbf{x}^*\| > \frac{\bar{\rho}(\mathbf{x}_{k-2})}{1 + M\xi}\|\mathbf{x}_{k-1} - \mathbf{x}^*\| \quad (183)$$

where we used the bound on  $\sigma(\mathbf{x})$  from (100) given by  $\sigma(\mathbf{x}_{k-2}) = M\|\mathbf{x}_{k-2} - \mathbf{x}^*\|^2 \leq M(\xi)^2$  followed by the condition  $\|\mathbf{x}_k - \mathbf{x}^*\| > \|\mathbf{x}_{k-2} - \mathbf{x}^*\|$  arising from the fact that  $\{\|\mathbf{x}_k - \mathbf{x}^*\|\}$  is a monotonically increasing sequence for  $K_e + 2 \leq k \leq \hat{K}_{exit}$  and finally the substitution  $\|\mathbf{x}_{\hat{K}_{exit}} - \mathbf{x}^*\| = \xi$ . Now applying the bound (183) recursively for  $K_e + 2 \leq k \leq \hat{K}_{exit}$  yields:

$$\|\mathbf{x}_{\hat{K}_{exit}} - \mathbf{x}^*\| > \prod_{k=K_e+2}^{\hat{K}_{exit}-1} \frac{\bar{\rho}(\mathbf{x}_{k-2})}{1 + M\xi} \|\mathbf{x}_{K_e+1} - \mathbf{x}^*\| \quad (184)$$

$$\|\mathbf{x}_{\hat{K}_{exit}} - \mathbf{x}^*\| > \left(\frac{\inf_{K_e+2 \leq k \leq \hat{K}_{exit}} \{\bar{\rho}(\mathbf{x}_{k-2})\}}{1 + M\xi}\right)^{\hat{K}_{exit} - K_e - 2} \|\mathbf{x}_{K_e+1} - \mathbf{x}^*\| \quad (185)$$

$$\hat{K}_{exit} - K_e - 2 < \frac{\log\left(\|\mathbf{x}_{\hat{K}_{exit}} - \mathbf{x}^*\|\right) - \log\left(\|\mathbf{x}_{K_e+1} - \mathbf{x}^*\|\right)}{\log\left(\frac{\inf\{\bar{\rho}(\mathbf{x}_{k-2})\}}{1 + M\xi}\right)} \quad (186)$$

$$< \frac{\log(\xi) - \log(\varepsilon)}{\log\left(\frac{\inf\{\bar{\rho}(\mathbf{x}_{k-2})\}}{1 + M\xi}\right)} \quad (187)$$

$$\hat{K}_{exit} - K_e < \frac{\log(\xi) - \log(\varepsilon)}{\log\left(\frac{\inf\{\bar{\rho}(\mathbf{x}_{k-2})\}}{1 + M\xi}\right)} + 2 \quad (187)$$

where in the last step we used  $\|\mathbf{x}_{\hat{K}_{exit}} - \mathbf{x}^*\| = \xi$ ,  $\|\mathbf{x}_{K_e+1} - \mathbf{x}^*\| \geq \varepsilon$  and the range of infimum is omitted after second step. Note that we require the condition  $\left(\frac{\inf\{\bar{\rho}(\mathbf{x}_{k-2})\}}{1 + M\xi}\right) > 1$ , however this is trivially satisfied which can be easily checked from (140) and (149).

For  $\xi \leq \frac{1}{\zeta M} \left( \frac{(1 + \frac{\beta}{L})^2 + \frac{1}{4(1 + \frac{\beta}{L})^2} - \frac{5}{4}}{6} \right)$  where  $\zeta > 2$ , we get the condition:

$$\frac{\bar{\rho}(\mathbf{x})}{1 + M\xi} > \frac{1 + \frac{\left( (1 + \frac{\beta}{L})^2 + \frac{1}{4(1 + \frac{\beta}{L})^2} - \frac{5}{4} \right)}{12}}{1 + \frac{\left( (1 + \frac{\beta}{L})^2 + \frac{1}{4(1 + \frac{\beta}{L})^2} - \frac{5}{4} \right)}{6\zeta}} > 1. \quad (188)$$

Finally adding (177) and (187), we get the following bound:

$$K_{shell} \leq \frac{\log\left(\frac{L}{2}\xi^2\right) - \log\left(\frac{\beta^2}{2L}\varepsilon^2 - \frac{2M}{3}\varepsilon^3\right)}{\log\left(1 - \frac{\beta^2}{L^2}\right)^{-1}} + \frac{\log(\xi) - \log(\varepsilon)}{\log\left(\frac{\inf\{\tilde{p}(\mathbf{x}_{k-2})\}}{1+M\xi}\right)} + 2 \quad (189)$$

where  $K_{shell} = K_c + \hat{K}_{exit} - K_e$ . ■

## APPENDIX E PROOF OF LEMMAS 2-6

Before proving Lemma 2 and 3 we need the relative error bound on zero-th order approximation of the gradient trajectory. Expanding the expression  $\mathbf{u}_K = \prod_{k=0}^{K-1} [\mathbf{A}_k + \varepsilon \mathbf{P}_k] \mathbf{u}_0$  from section III-A1 to zero-th order we get the following bound on tail error:

$$\mathbf{u}_K = \prod_{k=0}^{K-1} [\mathbf{A}_k + \varepsilon \mathbf{P}_k] \mathbf{u}_0 \quad (190)$$

$$= \prod_{k=0}^{K-1} \mathbf{A}_k \mathbf{u}_0 + \mathcal{O}\left(\|\mathbf{A}\|_2^K (K\varepsilon) \frac{\|\mathbf{P}\|_2}{\|\mathbf{A}\|_2} \|\mathbf{u}_0\|\right) \quad (191)$$

$$\Rightarrow \left\| \mathbf{u}_K - \prod_{k=0}^{K-1} \mathbf{A}_k \mathbf{u}_0 \right\| = \mathcal{O}\left(\|\mathbf{A}\|_2^K (K\varepsilon) \varepsilon\right) \quad (192)$$

where the above bound is obtained by following steps similar to (11). Then using this tail error bound along with (17) we get the following bound on relative error for zeroth order approximation:

$$\frac{\|\mathbf{u}_K - \prod_{k=0}^{K-1} \mathbf{A}_k \mathbf{u}_0\|}{\|\mathbf{u}_K\|} \leq \frac{\mathcal{O}\left(\left(2 + \frac{\varepsilon M}{2L}\right)^K (K\varepsilon) \varepsilon\right)}{\varepsilon \left(1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}\right)^K \sqrt{\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2} - \mathcal{O}\left(\|\mathbf{A}\|_2^K (K\varepsilon) \varepsilon\right)} \quad (193)$$

$$= \frac{\mathcal{O}\left(\frac{\left(2 + \frac{\varepsilon M}{2L}\right)^K}{\left(1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}\right)^K} (K\varepsilon)\right)}{\sqrt{\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2} - \mathcal{O}\left(\frac{\left(2 + \frac{\varepsilon M}{2L}\right)^K}{\left(1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L}\right)^K} (K\varepsilon)\right)} \quad (194)$$

$$\leq \frac{\mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)\right)}{\sqrt{\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2} - \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)\right)} \quad (195)$$

where we have substituted the upper bound on  $K_{exit}$  from (7) into  $K$ . Hence for  $\sqrt{\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2} > \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)\right)$  we have that:

$$\|\mathbf{u}_K\| \left(1 - \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)\right)\right) \leq \left\| \prod_{k=0}^{K-1} \mathbf{A}_k \mathbf{u}_0 \right\| \leq \|\mathbf{u}_K\| \left(1 + \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)\right)\right). \quad (196)$$

Now  $\mathbf{A}_k = \sum_{i \in \mathcal{N}_S} c_i^s(k) \mathbf{v}_i \mathbf{v}_i^T + \sum_{j \in \mathcal{N}_{US}} c_j^{us}(k) \mathbf{v}_j \mathbf{v}_j^T$  where  $\mathbf{v}_i$  and  $\mathbf{v}_j$  are the eigenvectors corresponding to the stable and unstable subspaces of  $\nabla^2 f(\mathbf{x}^*)$  and for  $\alpha = \frac{1}{L}$  we have the bounds  $1 + \frac{\beta}{L} - \frac{\varepsilon M}{2L} \leq c_j^{us}(k) \leq 2 + \frac{\varepsilon M}{2L}$  and  $-\frac{\varepsilon M}{2L} \leq c_i^s(k) \leq 1 - \frac{\beta}{L} + \frac{\varepsilon M}{2L}$ . Therefore we also get the bound:

$$\inf \left\| \prod_{k=0}^{K-1} \mathbf{A}_k \mathbf{u}_0 \right\| \leq \left\| \prod_{k=0}^{K-1} (\mathbf{I} - \alpha \nabla^2 f(\mathbf{x}^*)) \mathbf{u}_0 \right\| \leq \sup \left\| \prod_{k=0}^{K-1} \mathbf{A}_k \mathbf{u}_0 \right\|. \quad (197)$$

Combining this with (196) we get:

$$\|\mathbf{u}_K\| \left(1 - \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)\right)\right) \leq \left\| \prod_{k=0}^{K-1} (\mathbf{I} - \alpha \nabla^2 f(\mathbf{x}^*)) \mathbf{u}_0 \right\| \leq \|\mathbf{u}_K\| \left(1 + \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)\right)\right). \quad (198)$$

### Proof of Lemma 2

For values of  $\varepsilon$  sufficiently small and  $\sqrt{\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2} > \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)\right)$ , using (198) we have the following approximation:

$$\frac{\left\| \prod_{k=0}^{K-1} (\mathbf{I} - \alpha \nabla^2 f(\mathbf{x}^*)) \mathbf{u}_0 \right\|}{\left(1 + \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)\right)\right)} \leq \|\mathbf{u}_K\| \leq \frac{\left\| \prod_{k=0}^{K-1} (\mathbf{I} - \alpha \nabla^2 f(\mathbf{x}^*)) \mathbf{u}_0 \right\|}{\left(1 - \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)\right)\right)} \quad (199)$$

$$\Rightarrow \|\mathbf{u}_K\| \approx \left\| \prod_{k=0}^{K-1} (\mathbf{I} - \alpha \nabla^2 f(\mathbf{x}^*)) \mathbf{u}_0 \right\| = \|(\mathbf{I} - \alpha \nabla^2 f(\mathbf{x}^*))^K \mathbf{u}_0\| \quad (200)$$

where  $\mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)\right)$  term is neglected w.r.t. 1 for sufficiently small  $\varepsilon$  and  $K < K_{exit} \approx \mathcal{O}(\log(\varepsilon^{-1}))$ . Now, if  $\mathbf{u}_0$  has a projection value close to 0 on the unstable subspace of  $\nabla^2 f(\mathbf{x}^*)$ , then  $\|\mathbf{u}_K\|$  first approximately decreases exponentially such that  $\mathbf{x}_K$  reaches some  $\mathbf{x}_{critical}$  and from there onward it approximately increases exponentially until saddle region is escaped. For the case when  $\mathbf{x}_{critical} \rightarrow \mathbf{x}^*$ , we will have  $\|\mathbf{x}_{critical} - \mathbf{x}^*\| \rightarrow 0$ . The escape time for the  $\varepsilon$ -precision trajectories from this region  $\mathcal{B}_{\varepsilon'}(\mathbf{x}^*)$  where  $\varepsilon' = \|\mathbf{x}_{critical} - \mathbf{x}^*\|$  will be upper bounded by  $K < \mathcal{O}(\log(\varepsilon'^{-1}))$  from (7). This upper bound goes to infinity when  $\varepsilon' \rightarrow 0$  hence  $\varepsilon$ -precision trajectories fail to escape the saddle neighborhood when

$\mathbf{x}_{critical} = \mathbf{x}^*$ . It should also be noted that if for some  $K$ ,  $\mathbf{u}_K = \mathbf{0}$  or in other words  $\mathbf{x}_{critical} = \mathbf{x}^*$ , then for all  $J > K$  we have  $\mathbf{u}_J = \mathbf{0}$  since  $\nabla f(\mathbf{x}_J) = \mathbf{0}$  and the gradient trajectory can never escape the saddle region. ■

### Proof of Lemma 3

Let  $\{\mathbf{u}_K\}$  be any gradient trajectory with linear exit time that satisfies the condition  $\sqrt{\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2} > \mathcal{O}\left(\frac{1}{\sqrt{\epsilon}} \left(\log\left(\frac{1}{\epsilon}\right) \epsilon\right)\right)$ . Now if this trajectory curves around  $\mathbf{x}^*$  then the vectors  $\mathbf{u}_0$  and  $\mathbf{u}_K$  will form an obtuse angle for some finite values of  $K$ . Therefore in order to prove the first part, it is sufficient to show that:

$$\langle \mathbf{u}_K, \mathbf{u}_0 \rangle \geq 0$$

for any value of  $K$  such that  $\|\mathbf{u}_K\| < \epsilon$ . Now, for sufficiently small  $\epsilon$  where  $\epsilon$  is upper bounded by Theorem 1 from [195] we have  $\mathbf{u}_K = \prod_{k=0}^{K-1} \mathbf{A}_k \mathbf{u}_0 + \mathcal{O}\left(\frac{1}{\sqrt{\epsilon}} \left(\log\left(\frac{1}{\epsilon}\right) \epsilon^2\right)\right) \approx (\mathbf{I} - \alpha \nabla^2 f(\mathbf{x}^*))^K \mathbf{u}_0$  where we used the fact that  $\prod_{k=0}^{K-1} \mathbf{A}_k = (\mathbf{I} - \alpha \nabla^2 f(\mathbf{x}^*))^K + \mathcal{O}(K\epsilon)$  and dropped the term  $\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}} \left(\log\left(\frac{1}{\epsilon}\right) \epsilon^2\right)\right)$  for sufficiently small  $\epsilon$ . Using this approximate  $\mathbf{u}_K$  we get:

$$\langle \mathbf{u}_K, \mathbf{u}_0 \rangle \approx \mathbf{u}_0^T (\mathbf{I} - \alpha \nabla^2 f(\mathbf{x}^*))^K \mathbf{u}_0 \geq 0 \quad (201)$$

where the last inequality comes from the fact that  $(\mathbf{I} - \alpha \nabla^2 f(\mathbf{x}^*))^K$  will be a positive semi-definite matrix for  $\alpha \leq \frac{1}{L}$ . Therefore, vectors  $\mathbf{u}_0$  and  $\mathbf{u}_K$  will form an acute angle between them for all values of  $K$  such that  $\|\mathbf{u}_K\| < \epsilon$  and  $K \leq K_{exit} = \mathcal{O}(\log(\epsilon^{-1}))$ . Hence, the trajectory can never curve around  $\mathbf{x}^*$ .

The proof for second part follows the same method. Let us take any two points on the gradient trajectory denoted by vectors  $\mathbf{u}_{K_1}$  and  $\mathbf{u}_{K_2}$  w.r.t. stationary point  $\mathbf{x}^*$ . Then we have the following inner product:

$$\langle \mathbf{u}_{K_1}, \mathbf{u}_{K_2} \rangle \approx \langle \mathbf{u}_0, (\mathbf{I} - \alpha \nabla^2 f(\mathbf{x}^*))^{K_1+K_2} \mathbf{u}_0 \rangle \geq 0 \quad (202)$$

for  $K_1 + K_2 \leq \mathcal{O}(\log(\epsilon^{-1}))$ . Now with  $\langle \mathbf{u}_{K_1}, \mathbf{u}_{K_2} \rangle \gtrsim 0$  for any  $K_1, K_2$  where  $K_1 + K_2 \leq \mathcal{O}(\log(\epsilon^{-1}))$  such that  $\|\mathbf{u}_{K_1}\| < \epsilon$  and  $\|\mathbf{u}_{K_2}\| < \epsilon$ , the angle between the vectors  $\mathbf{u}_{K_1}$  and  $\mathbf{u}_{K_2}$  can never approximately exceed  $\frac{\pi}{2}$ . Hence the entire gradient descent trajectory approximately lies inside some orthant of the ball  $\mathcal{B}_\epsilon(\mathbf{x}^*)$ . ■

### Proof of Lemma 4

Let us denote the exit point on the ball  $\mathcal{B}_\epsilon(\mathbf{x}^*)$  by  $\mathbf{x}_{K+1}$  where  $\|\mathbf{x}_K - \mathbf{x}^*\| \leq \epsilon$  and  $\|\mathbf{x}_{K+1} - \mathbf{x}^*\| > \epsilon$ . Also,  $\|\mathbf{x}_{K+1} - \mathbf{x}^*\| \leq \|\mathbf{x}_K - \mathbf{x}^*\| + \frac{1}{L} \|\nabla f(\mathbf{x}_K)\| \leq 2\|\mathbf{x}_K - \mathbf{x}^*\|$  which

implies  $\|\mathbf{x}_K - \mathbf{x}^*\| \geq \frac{\|\mathbf{x}_{K+1} - \mathbf{x}^*\|}{2} \geq \frac{\epsilon}{2}$ . Now applying the Hessian Lipschitz condition around  $\mathbf{x}^*$  for  $\mathbf{x}_K$ , we get the following:

$$f(\mathbf{x}_K) \leq f(\mathbf{x}^*) + \langle \nabla f(\mathbf{x}^*), \mathbf{x}_K - \mathbf{x}^* \rangle + \frac{1}{2} \langle (\mathbf{x}_K - \mathbf{x}^*), \nabla^2 f(\mathbf{x}^*) (\mathbf{x}_K - \mathbf{x}^*) \rangle + \frac{M}{6} \|\mathbf{x}_K - \mathbf{x}^*\|^3 \quad (203)$$

$$\leq f(\mathbf{x}^*) + \frac{\langle \mathbf{x}_K - \mathbf{x}^*, \nabla f(\mathbf{x}_K) \rangle}{2} + \frac{1}{2} \left\langle (\mathbf{x}_K - \mathbf{x}^*), \left( \nabla^2 f(\mathbf{x}^*) - \nabla^2 f(\mathbf{x}^*) - \mathcal{O}(\epsilon) \right) (\mathbf{x}_K - \mathbf{x}^*) \right\rangle + \frac{M}{6} \|\mathbf{x}_K - \mathbf{x}^*\|^3 \quad (204)$$

$$\leq f(\mathbf{x}^*) + \frac{\langle \mathbf{x}_K - \mathbf{x}^*, \nabla f(\mathbf{x}_K) \rangle}{2} + \mathcal{O}(\epsilon^3) \quad (205)$$

where we have used  $\nabla f(\mathbf{x}_K) = \left( \nabla^2 f(\mathbf{x}^*) + \mathcal{O}(\epsilon) \right) (\mathbf{x}_K - \mathbf{x}^*)$  from Lemma 2 and substituted  $\|\mathbf{x}_K - \mathbf{x}^*\| \leq \epsilon$  in the last step.

Let us first analyze the term  $\frac{\langle \mathbf{x}_K - \mathbf{x}^*, \nabla f(\mathbf{x}_K) \rangle}{2}$ . Now,  $\|\mathbf{x}_K - \mathbf{x}^*\| < \|\mathbf{x}_{K+1} - \mathbf{x}^*\|$  since the gradient descent trajectory is exiting the ball  $\mathcal{B}_\epsilon(\mathbf{x}^*)$  at iteration  $K+1$  and therefore it has expansive dynamics at this iteration [13]. Squaring the condition  $\|\mathbf{x}_K - \mathbf{x}^*\| < \|\mathbf{x}_{K+1} - \mathbf{x}^*\|$  yields:

$$\|\mathbf{x}_K - \mathbf{x}^*\|^2 < \|\mathbf{x}_{K+1} - \mathbf{x}^*\|^2 \quad (206)$$

$$\|\mathbf{x}_K - \mathbf{x}^*\|^2 < \|\mathbf{x}_K - \mathbf{x}^*\|^2 + \alpha \|\nabla f(\mathbf{x}_K)\|^2 - 2\alpha \langle \mathbf{x}_K - \mathbf{x}^*, \nabla f(\mathbf{x}_K) \rangle \quad (207)$$

$$\langle \mathbf{x}_K - \mathbf{x}^*, \nabla f(\mathbf{x}_K) \rangle < \frac{\alpha}{2} \|\nabla f(\mathbf{x}_K)\|^2. \quad (208)$$

Next, by the gradient Lipschitz continuity for  $\mathbf{x}_K$  and  $\mathbf{x}_{K+1}$ , we have that:

$$f(\mathbf{x}_{K+1}) \leq f(\mathbf{x}_K) + \langle \nabla f(\mathbf{x}_K), \mathbf{x}_{K+1} - \mathbf{x}_K \rangle + \frac{L}{2} \|\mathbf{x}_{K+1} - \mathbf{x}_K\|^2 \quad (209)$$

$$f(\mathbf{x}_{K+1}) \leq f(\mathbf{x}_K) - \alpha \|\nabla f(\mathbf{x}_K)\|^2 + \frac{L}{2} \|\alpha \nabla f(\mathbf{x}_K)\|^2 \quad (210)$$

$$f(\mathbf{x}_{K+1}) + \frac{1}{2L} \|\nabla f(\mathbf{x}_K)\|^2 \leq f(\mathbf{x}_K) \quad (211)$$

where we substituted  $\alpha = \frac{1}{L}$ . Combining (211) with (205) followed by substitution of (208) yields:

$$f(\mathbf{x}_{K+1}) + \frac{1}{2L} \|\nabla f(\mathbf{x}_K)\|^2 \leq f(\mathbf{x}_K) \leq f(\mathbf{x}^*) + \frac{\langle \mathbf{x}_K - \mathbf{x}^*, \nabla f(\mathbf{x}_K) \rangle}{2} + \mathcal{O}(\epsilon^3) \quad (212)$$

$$\implies f(\mathbf{x}_{K+1}) + \frac{1}{2L} \|\nabla f(\mathbf{x}_K)\|^2 \leq f(\mathbf{x}^*) + \frac{\alpha}{4} \|\nabla f(\mathbf{x}_K)\|^2 + \mathcal{O}(\epsilon^3) \quad (213)$$

$$\implies f(\mathbf{x}_{K+1}) \leq f(\mathbf{x}^*) - \frac{1}{4L} \|\nabla f(\mathbf{x}_K)\|^2 + \mathcal{O}(\epsilon^3). \quad (214)$$

Next, using the bound  $\|\nabla f(\mathbf{x}_K)\| \geq \beta \|\mathbf{x}_K - \mathbf{x}^*\|$  from (158) in (214) and the fact that  $\|\mathbf{x}_K - \mathbf{x}^*\| \geq \frac{\epsilon}{2}$  we obtain:

$$f(\mathbf{x}_{K+1}) \leq f(\mathbf{x}^*) - \frac{\beta^2}{4L} \|\mathbf{x}_K - \mathbf{x}^*\|^2 + \mathcal{O}(\epsilon^3) \leq f(\mathbf{x}^*) - \frac{\beta^2}{16L} \epsilon^2 + \mathcal{O}(\epsilon^3) \quad (215)$$

$$\implies f(\mathbf{x}_{K+1}) < f(\mathbf{x}^*) \quad (216)$$

<sup>13</sup>Exit at iteration  $K+1$  implies  $\|\mathbf{x}_K - \mathbf{x}^*\| < \|\mathbf{x}_{K+1} - \mathbf{x}^*\|$ .

for sufficiently small  $\varepsilon$ .

*Proof of Lemma 5*

Let us take any two points  $\mathbf{x}_1, \mathbf{x}_2$  in the closed ball  $\bar{\mathcal{B}}_\varepsilon(\mathbf{x}^*)$ . Using gradient Lipschitz condition, we get the following inequalities:

$$f(\mathbf{x}_1) \leq f(\mathbf{x}^*) + \langle \nabla f(\mathbf{x}^*), \mathbf{x}_1 - \mathbf{x}^* \rangle + \frac{L}{2} \|\mathbf{x}_1 - \mathbf{x}^*\|^2 \quad (217)$$

$$\leq f(\mathbf{x}^*) + \frac{L}{2} \|\mathbf{x}_1 - \mathbf{x}^*\|^2 \quad (218)$$

and

$$f(\mathbf{x}^*) \leq f(\mathbf{x}_2) - \langle \nabla f(\mathbf{x}^*), \mathbf{x}_2 - \mathbf{x}^* \rangle + \frac{L}{2} \|\mathbf{x}_2 - \mathbf{x}^*\|^2 \quad (219)$$

$$\leq f(\mathbf{x}_2) + \frac{L}{2} \|\mathbf{x}_2 - \mathbf{x}^*\|^2 \quad (220)$$

Now adding (218) and (220) yields:

$$f(\mathbf{x}_1) - f(\mathbf{x}_2) \leq \frac{L}{2} \|\mathbf{x}_2 - \mathbf{x}^*\|^2 + \frac{L}{2} \|\mathbf{x}_1 - \mathbf{x}^*\|^2. \quad (221)$$

Next, using the fact that  $\|\mathbf{x}_2 - \mathbf{x}^*\| \leq \varepsilon$ ,  $\|\mathbf{x}_1 - \mathbf{x}^*\| \leq \varepsilon$  in (221), we get the following upper bound:

$$f(\mathbf{x}_1) - f(\mathbf{x}_2) \leq L\varepsilon^2. \quad (222)$$

Formally, this upper bound states that the function value gap between any two points in the closed ball  $\bar{\mathcal{B}}_\varepsilon(\mathbf{x}^*)$  surface cannot be more than  $L\varepsilon^2$ . Also notice that the result in (222) only depends on the gradient Lipschitz condition and therefore will hold true for any  $\varepsilon$ . Next, we assume that our gradient trajectory is currently exiting the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  at point  $\mathbf{x}_K$  s.t.  $\|\mathbf{x}_{K-1} - \mathbf{x}^*\| \leq \varepsilon$  and  $\|\mathbf{x}_K - \mathbf{x}^*\| > \varepsilon$ . Let us further assume that  $\hat{K}$  iterations after the current iteration, the gradient trajectory re-enters the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$ , i.e.,  $\|\mathbf{x}_{K+\hat{K}} - \mathbf{x}^*\| \leq \varepsilon$  and  $\|\mathbf{x}_{K+\hat{K}-1} - \mathbf{x}^*\| > \varepsilon$ . Using the update equation  $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)$  for  $0 \ll \alpha \leq \frac{1}{L}$  together with gradient Lipschitz condition, we get:

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \quad (223)$$

$$\implies f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{\alpha L}{2} \left( \frac{2}{L} - \alpha \right) \|\nabla f(\mathbf{x}_k)\|^2 \quad (224)$$

Taking the telescopic sum for these inequalities from  $k = K$  to  $k = K + \hat{K} - 1$  gives the following lower bound on  $f(\mathbf{x}_K) - f(\mathbf{x}_{K+\hat{K}})$ :

$$f(\mathbf{x}_{K+\hat{K}}) \leq f(\mathbf{x}_K) - \frac{\alpha L}{2} \left( \frac{2}{L} - \alpha \right) \sum_{k=K}^{K+\hat{K}-1} \|\nabla f(\mathbf{x}_k)\|^2 \quad (225)$$

$$\begin{aligned} \frac{\alpha L \beta^2}{2} \left( \frac{2}{L} - \alpha \right) \hat{K} \varepsilon^2 &< \frac{\alpha L}{2} \left( \frac{2}{L} - \alpha \right) \sum_{k=K}^{K+\hat{K}-1} \|\nabla f(\mathbf{x}_k)\|^2 \\ &\leq f(\mathbf{x}_K) - f(\mathbf{x}_{K+\hat{K}}) \leq f(\mathbf{x}_{K-1}) - f(\mathbf{x}_{K+\hat{K}}) \end{aligned} \quad (226)$$

where  $f(\mathbf{x}_K) \leq f(\mathbf{x}_{K-1})$  from monotonicity of  $\{f(\mathbf{x}_k)\}$  and we have substituted the lower bound

$$\|\nabla f(\mathbf{x}_k)\| \geq \beta \|\mathbf{x}_k - \mathbf{x}^*\| \geq \beta \varepsilon$$

from (158) since  $\|\mathbf{x}_k - \mathbf{x}^*\| > \varepsilon$  for all  $K \leq k \leq K + \hat{K} - 1$ . Combining (226) with (222) for  $\mathbf{x}_{K-1}, \mathbf{x}_{K+\hat{K}} \in \mathcal{B}_\varepsilon(\mathbf{x}^*)$  yields the following condition on  $\hat{K}$ :

$$\frac{\alpha L \beta^2}{2} \left( \frac{2}{L} - \alpha \right) \hat{K} \varepsilon^2 < L \varepsilon^2 \quad (227)$$

$$\hat{K} < \frac{2}{\alpha \beta^2 \left( \frac{2}{L} - \alpha \right)}. \quad (228)$$

Now, for sake of simplicity we substitute  $\alpha = \frac{1}{L}$ <sup>14</sup>. This yields the following bound on  $\hat{K}$ :

$$\hat{K} < \frac{2}{\kappa^2} \quad (229)$$

where  $\kappa = \frac{\beta}{L}$ . This inequality claims that if the gradient trajectory re-enters the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$ , it has to do so in fewer than  $\frac{2}{\kappa^2}$  iterations. From here onward we will develop a proof which contradicts this claim.

Let us first define some  $\xi > \varepsilon$  such that  $\xi = 2^{\frac{2}{\kappa^2}} \varepsilon (1+b)$  where  $\kappa = \frac{\beta}{L}$ ,  $b = \frac{\|\mathbf{x}_K - \mathbf{x}^*\|}{\varepsilon} - 1$  is a positive value and  $\xi$  is upper bounded from theorem 3. Note that  $\mathbf{x}_K$  as defined earlier is the exit point of the gradient trajectory, i.e.,  $\|\mathbf{x}_{K-1} - \mathbf{x}^*\| \leq \varepsilon$  and  $\|\mathbf{x}_K - \mathbf{x}^*\| > \varepsilon$ . Now for any  $\varepsilon \ll 2^{-\frac{2}{\kappa^2}}$  we will have  $\xi = \mathcal{O}(\varepsilon)$ . Therefore a gradient trajectory moving outwards from the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  is also bound to move out from the ball  $\mathcal{B}_\xi(\mathbf{x}^*)$  since we have already proved this in Theorem 2 for trajectories with expansive dynamics.

Under these conditions, let  $J$  represent the minimum number of iterations required to exit the ball  $\mathcal{B}_\xi(\mathbf{x}^*)$  for a trajectory which is just exiting  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  and is currently at the point  $\mathbf{x}_K$  s.t.  $\|\mathbf{x}_K - \mathbf{x}^*\| > \varepsilon$ . To this end, we rewrite the update equation of radial vector  $\mathbf{u}_k$  for any  $k \in \{K, K+1, \dots, K+J-1\}$ :

$$\mathbf{u}_{k+1} = \mathbf{u}_k + (\mathbf{x}_{k+1} - \mathbf{x}_k) = \mathbf{u}_k - \alpha \nabla f(\mathbf{x}_k) \quad (230)$$

where we have that  $\mathbf{u}_k = \mathbf{x}_k - \mathbf{x}^*$ . From the gradient Lipschitz condition we have the following bound for any  $\mathbf{u}_k$ :

$$\|\nabla f(\mathbf{x}_k)\| \leq L \|\mathbf{u}_k\| \quad (231)$$

where  $\mathbf{u}_k = \mathbf{x}_k - \mathbf{x}^*$ . Applying norm to (230) followed by triangle inequality and using the upper bound from (231) yields:

$$\|\mathbf{u}_{k+1}\| = \|\mathbf{u}_k + (\mathbf{x}_{k+1} - \mathbf{x}_k)\| \leq \|\mathbf{u}_k\| + \alpha \|\nabla f(\mathbf{x}_k)\| \leq 2 \|\mathbf{u}_k\| \quad (232)$$

for  $\alpha = \frac{1}{L}$ . Applying this bound recursively from  $k = K$  to  $k = K + J - 1$  and substituting  $\|\mathbf{u}_K\| = \varepsilon(1+b)$ , we have:

$$\|\mathbf{u}_{K+J}\| \leq 2^J \|\mathbf{u}_K\| = 2^J \varepsilon (1+b). \quad (233)$$

Since  $J$  is the minimum number of iterations required to exit the  $\xi$  radius ball for a trajectory which is just exiting the  $\varepsilon$  ball, we can set  $2^J \varepsilon (1+b) = \xi$ . This yields:

$$2^J \varepsilon (1+b) = \xi = 2^{\frac{2}{\kappa^2}} \varepsilon (1+b) \quad (234)$$

$$J = \frac{2}{\kappa^2}. \quad (235)$$

<sup>14</sup>It is to be noted that we can carry out a similar analysis for any other  $\alpha$  s.t.  $0 \ll \alpha \leq \frac{1}{L}$  and still obtain the same inference.

Now, the  $\hat{K}$  we defined as the time to re-enter the ball  $\mathcal{B}_\xi(\mathbf{x}^*)$  should be definitely greater than  $J$  since any trajectory will certainly take more than  $J$  iterations to traverse the shell present in between the concentric  $\xi$  and  $\varepsilon$  radii balls.

$$\hat{K} > J = \frac{2}{\kappa^2}. \quad (236)$$

However, this inequality contradicts the claim that  $\hat{K} < \frac{2}{\kappa^2}$  from (229) which completes our proof. ■

#### Proof of Lemma 6

Recall that from (221) and (222) in previous lemma, for any  $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{B}_\xi(\mathbf{x}^*)$  we have that:

$$f(\mathbf{x}_1) - f(\mathbf{x}_2) \leq L(\xi)^2. \quad (237)$$

Next, let  $\hat{K}$  be the minimum number of iterations in which the gradient trajectory re-enters the ball  $\mathcal{B}_\xi(\mathbf{x}^*)$ . Then following the same set of steps as in the previous lemma for obtaining (225), we get:

$$\begin{aligned} f(\mathbf{x}_{K+\hat{K}}) &\leq f(\mathbf{x}_K) - \frac{\alpha L}{2} \left( \frac{2}{L} - \alpha \right) \sum_{k=K}^{K+\hat{K}-1} \|\nabla f(\mathbf{x}_k)\|^2 \\ &\Rightarrow \frac{\alpha L^3}{4} \left( \frac{2}{L} - \alpha \right) \hat{K}(\xi)^2 < \frac{\alpha L}{2} \left( \frac{2}{L} - \alpha \right) \sum_{k=K}^{K+\hat{K}-1} \|\nabla f(\mathbf{x}_k)\|^2 \\ &\leq f(\mathbf{x}_K) - f(\mathbf{x}_{K+\hat{K}}) \\ &\leq f(\mathbf{x}_{K-1}) - f(\mathbf{x}_{K+\hat{K}}) \end{aligned} \quad (238) \quad (239)$$

where we substituted  $\|\nabla f(\mathbf{x}_k)\| \geq \gamma > \frac{1}{\sqrt{2}}L\xi$  and  $f(\mathbf{x}_K) \leq f(\mathbf{x}_{K-1})$  from monotonicity of  $\{f(\mathbf{x}_k)\}$ . Now if the trajectory re-enters the ball  $\mathcal{B}_\xi(\mathbf{x}^*)$  in  $\hat{K}$  iterations, then  $\mathbf{x}_{K-1}, \mathbf{x}_{K+\hat{K}} \in \mathcal{B}_\xi(\mathbf{x}^*)$  and hence  $\mathbf{x}_{K-1}, \mathbf{x}_{K+\hat{K}}$  satisfy (237). Therefore combining (239) with (237) yields the bound:

$$\frac{\alpha L^3}{4} \left( \frac{2}{L} - \alpha \right) \hat{K}(\xi)^2 < L(\xi)^2 \quad (240)$$

$$\Rightarrow \hat{K} < \frac{4}{\alpha L^2 \left( \frac{2}{L} - \alpha \right)}. \quad (241)$$

Now for  $\alpha = \frac{1}{L}$ , we have that  $\hat{K} < 4$ . Therefore the gradient trajectory has to re-enter the ball  $\mathcal{B}_\xi(\mathbf{x}^*)$  in three or less iterations. We now show that the gradient trajectory cannot return in three or less iterations.

Let the current iterate for the gradient trajectory be  $\mathbf{x}^-$  such that  $\|\mathbf{x}^- - \mathbf{x}^*\| < \xi$  and  $\|\mathbf{x} - \mathbf{x}^*\| \geq \xi$ , i.e., the iterate  $\mathbf{x}$  exits the ball  $\mathcal{B}_\xi(\mathbf{x}^*)$  where  $\xi$  is bounded from Theorem 3. Next, from Theorem 2 the iterate  $\mathbf{x}^+$  in the sequence  $\{\mathbf{x}^-, \mathbf{x}, \mathbf{x}^+\}$  will also have expansive dynamics, i.e.,  $\|\mathbf{x}^+ - \mathbf{x}^*\| > \|\mathbf{x} - \mathbf{x}^*\|$ . Let  $\mathbf{x}^{++}$  denote the next iterate in the sequence  $\{\mathbf{x}^-, \mathbf{x}, \mathbf{x}^+\}$ . Now, if the following condition:

$$\langle \mathbf{x}^{++} - \mathbf{x}^+, \mathbf{x}^+ - \mathbf{x}^* \rangle \geq 0 \quad (242)$$

is satisfied, then  $\mathbf{x}^{++} \notin \mathcal{B}_\xi(\mathbf{x}^*)$ . To check this, let the condition (242) be given and we have the contradiction  $\mathbf{x}^{++} \in \mathcal{B}_\xi(\mathbf{x}^*)$ ,

i.e.,  $\|\mathbf{x}^{++} - \mathbf{x}^*\| < \xi$ . Then we can write the following inequality:

$$\|\mathbf{x}^{++} - \mathbf{x}^*\|^2 < (\xi)^2 \quad (243)$$

$$\Rightarrow \|\mathbf{x}^{++} - \mathbf{x}^+ + \mathbf{x}^+ - \mathbf{x}^*\|^2 < (\xi)^2 \quad (244)$$

$$\Rightarrow \underbrace{\|\mathbf{x}^{++} - \mathbf{x}^+\|^2}_{>0} + \underbrace{\|\mathbf{x}^+ - \mathbf{x}^*\|^2}_{\geq (\xi)^2} + 2 \underbrace{\langle \mathbf{x}^{++} - \mathbf{x}^+, \mathbf{x}^+ - \mathbf{x}^* \rangle}_{\geq 0} < (\xi)^2 \quad (245)$$

which is not possible (left hand side is greater than right hand side). Hence,  $\mathbf{x}^{++} \notin \mathcal{B}_\xi(\mathbf{x}^*)$ .

Now, we are left to prove (242) condition, i.e.,  $\langle \mathbf{x}^{++} - \mathbf{x}^+, \mathbf{x}^+ - \mathbf{x}^* \rangle \geq 0$ . Manipulating the left hand side of this condition and using the substitutions  $\mathbf{x}^{++} - \mathbf{x}^+ = -\alpha \nabla f(\mathbf{x}^+)$ ,  $\mathbf{x}^+ - \mathbf{x} = -\alpha \nabla f(\mathbf{x})$  and  $\nabla f(\mathbf{x}^+) = \nabla f(\mathbf{x}) + \left( \int_{p=0}^1 \nabla^2 f(\mathbf{x} + p(\mathbf{x}^+ - \mathbf{x})) dp \right) (\mathbf{x}^+ - \mathbf{x})$ , we obtain:

$$\langle \mathbf{x}^{++} - \mathbf{x}^+, \mathbf{x}^+ - \mathbf{x}^* \rangle = \langle -\alpha \nabla f(\mathbf{x}^+), \mathbf{x}^+ - \mathbf{x}^* \rangle \quad (246)$$

$$= -\alpha \left\langle \nabla f(\mathbf{x}) + \left( \int_{p=0}^1 \nabla^2 f(\mathbf{x} + p(\mathbf{x}^+ - \mathbf{x})) dp \right) (\mathbf{x}^+ - \mathbf{x}), \mathbf{x}^+ - \mathbf{x}^* \right\rangle \quad (247)$$

$$= -\alpha \left\langle \left( \mathbf{I} - \alpha \int_{p=0}^1 \nabla^2 f(\mathbf{x} + p(\mathbf{x}^+ - \mathbf{x})) dp \right) \nabla f(\mathbf{x}), \mathbf{x}^+ - \mathbf{x}^* \right\rangle \quad (248)$$

$$= \left\langle \mathbf{x}^+ - \mathbf{x}^*, \left( \mathbf{I} - \alpha \int_{p=0}^1 \nabla^2 f(\mathbf{x} + p(\mathbf{x}^+ - \mathbf{x})) dp \right) (-\alpha \nabla f(\mathbf{x})) \right\rangle \quad (249)$$

$$= \left\langle \mathbf{x}^+ - \mathbf{x}^*, \left( \mathbf{I} - \alpha \int_{p=0}^1 \nabla^2 f(\mathbf{x} + p(\mathbf{x}^+ - \mathbf{x})) dp \right) (\mathbf{x}^+ - \mathbf{x}) \right\rangle \quad (250)$$

where  $\left( \mathbf{I} - \alpha \int_{p=0}^1 \nabla^2 f(\mathbf{x} + p(\mathbf{x}^+ - \mathbf{x})) dp \right)$  is a positive semi-definite matrix for  $\alpha \leq \frac{1}{L}$ . Next, recall that from (183) in the proof for theorem 3 for any tuple  $\{\mathbf{x}^-, \mathbf{x}, \mathbf{x}^+\}$  generated by the gradient descent method where  $\mathbf{x}^- \in \mathcal{B}_\xi(\mathbf{x}^*)$ , we have that:

$$\|\mathbf{x}^+ - \mathbf{x}^*\| \geq \left( \frac{\bar{\rho}(\mathbf{x}^-)}{1 + M\xi} \right) \|\mathbf{x} - \mathbf{x}^*\| > \|\mathbf{x} - \mathbf{x}^*\| \quad (251)$$

for  $\frac{\bar{\rho}(\mathbf{x}^-)}{1 + M\xi} > 1$ .

Using this fact that  $\|\mathbf{x}^+ - \mathbf{x}^*\| > \|\mathbf{x} - \mathbf{x}^*\|$  followed by the cosine identity of triangles we get:

$$\frac{\langle \mathbf{x}^+ - \mathbf{x}^*, \mathbf{x}^+ - \mathbf{x} \rangle}{\|\mathbf{x}^+ - \mathbf{x}^*\| \|\mathbf{x}^+ - \mathbf{x}\|} = \frac{\|\mathbf{x}^+ - \mathbf{x}\|^2 + \|\mathbf{x}^+ - \mathbf{x}^*\|^2 - \|\mathbf{x} - \mathbf{x}^*\|^2}{2 \|\mathbf{x}^+ - \mathbf{x}^*\| \|\mathbf{x}^+ - \mathbf{x}\|} > 0 \quad (252)$$

$$\Rightarrow \langle \mathbf{x}^+ - \mathbf{x}^*, \mathbf{x}^+ - \mathbf{x} \rangle > 0. \quad (253)$$

For any vectors  $\mathbf{a}$  and  $\mathbf{b}$  and any positive semi-definite matrix  $\mathbf{A}$ , if  $\langle \mathbf{a}, \mathbf{b} \rangle \geq 0$  then  $\langle \mathbf{a}, \mathbf{A}\mathbf{b} \rangle \geq 0$ . Using this property for  $\mathbf{A} = \left( \mathbf{I} - \alpha \int_{p=0}^1 \nabla^2 f(\mathbf{x} + p(\mathbf{x}^+ - \mathbf{x})) dp \right)$ ,  $\mathbf{b} = \mathbf{x}^+ - \mathbf{x}$  and  $\mathbf{a} = \mathbf{x}^+ - \mathbf{x}^*$ , we get that  $\left\langle \mathbf{x}^+ - \mathbf{x}^*, \left( \mathbf{I} - \alpha \int_{p=0}^1 \nabla^2 f(\mathbf{x} + p(\mathbf{x}^+ - \mathbf{x})) dp \right) (\mathbf{x}^+ - \mathbf{x}) \right\rangle > 0$ .

$\mathbf{x})dp)(\mathbf{x}^+ - \mathbf{x}) \geq 0$  since  $\langle \mathbf{x}^+ - \mathbf{x}, \mathbf{x}^+ - \mathbf{x}^* \rangle \geq 0$ . Hence from (250), we have that:

$$\begin{aligned} & \langle \mathbf{x}^{++} - \mathbf{x}^+, \mathbf{x}^+ - \mathbf{x}^* \rangle = \\ & \left\langle \mathbf{x}^+ - \mathbf{x}^*, \left( \mathbf{I} - \alpha \int_{p=0}^1 \nabla^2 f(\mathbf{x} + p(\mathbf{x}^+ - \mathbf{x})) dp \right) (\mathbf{x}^+ - \mathbf{x}) \right\rangle \geq 0 \end{aligned} \quad (254)$$

which completes the proof.  $\blacksquare$

## APPENDIX F

### Proof of Lemma 7

To establish the linear exit time of the proposed algorithm from any strict saddle neighborhood it is sufficient to prove the curvature condition (refer Step 15 from Algorithm 1). Now, for  $\|\nabla f(\mathbf{x})\| \leq \varepsilon$  and  $\Xi = 0$ , we have that:

$$\nabla f(\mathbf{x}) = \left( \nabla f(\mathbf{x}^*) + \int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) dp \right) (\mathbf{x} - \mathbf{x}^*). \quad (255)$$

With  $\varepsilon$  very small and upper bounded by Theorem 1 using Lemma 3.3 from [10] we can approximate the Hessian  $\nabla^2 f(\mathbf{x}^* + p(\mathbf{x} - \mathbf{x}^*)) = \nabla^2 f(\mathbf{x}^*) + \mathcal{O}(\varepsilon) \approx \nabla^2 f(\mathbf{x}^*)$  for any  $\mathbf{x} \in \mathcal{B}_\varepsilon(\mathbf{x}^*)$ . This is a valid approximation since we are no longer solving for rates of convergence and just need to approximately determine the unstable projection value. In particular our goal is to determine whether the estimated unstable projection value is of the order  $\Theta\left(\frac{1}{\log(\varepsilon^{-1})}\right)$  from Theorem 1 so that the linear exit time condition can be checked for the given trajectory. Now  $\Theta\left(\frac{1}{\log(\varepsilon^{-1})}\right)$  term will dominate all  $\mathcal{O}(\varepsilon)$  terms for any sufficiently small  $\varepsilon$  and so dropping these  $\mathcal{O}(\varepsilon)$  terms now will still yield the same result. Therefore, the equation (255) for  $\mathbf{x} = \mathbf{x}_k$  is approximated as:

$$\nabla f(\mathbf{x}_k) = \left( \nabla^2 f(\mathbf{x}^*) + \mathcal{O}(\varepsilon) \right) (\mathbf{x}_k - \mathbf{x}^*) \approx \nabla^2 f(\mathbf{x}^*) (\mathbf{x}_k - \mathbf{x}^*) \quad (256)$$

where  $\nabla f(\mathbf{x}^*)$  is zero vector. With  $\mathbf{y}_0 = \mathbf{x}_k$ ,  $\mathbf{y}_1 = \mathbf{x}_{k+1}$  and the approximation (256), we have the following terms:

$$\mathbf{y}_1 = \mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k) \quad (257)$$

$$= \mathbf{x}_k - \alpha \left( \nabla^2 f(\mathbf{x}^*) + \mathcal{O}(\varepsilon) \right) (\mathbf{x}_k - \mathbf{x}^*) \quad (258)$$

$$\approx \mathbf{x}_k - \alpha \nabla^2 f(\mathbf{x}^*) (\mathbf{x}_k - \mathbf{x}^*), \quad (259)$$

$$\nabla f(\mathbf{y}_1) = \nabla f(\mathbf{x}_{k+1}) = \left( \nabla^2 f(\mathbf{x}^*) + \mathcal{O}(\varepsilon) \right) (\mathbf{x}_{k+1} - \mathbf{x}^*) \quad (260)$$

$$= \left( \nabla^2 f(\mathbf{x}^*) + \mathcal{O}(\varepsilon) \right) \left( \mathbf{x}_k - \alpha \left( \nabla^2 f(\mathbf{x}^*) + \mathcal{O}(\varepsilon) \right) (\mathbf{x}_k - \mathbf{x}^*) - \mathbf{x}^* \right) \quad (261)$$

$$\approx \nabla^2 f(\mathbf{x}^*) \left( \mathbf{x}_k - \alpha \nabla^2 f(\mathbf{x}^*) (\mathbf{x}_k - \mathbf{x}^*) - \mathbf{x}^* \right). \quad (262)$$

Note that in the second last step we used the substitution from (258). Now, we define the terms  $V_1, V_2$  using  $\mathbf{y}_0, \mathbf{y}_1$ :

$$V_1 = \langle \mathbf{y}_1 - \mathbf{y}_0, \mathbf{y}_1 - \mathbf{y}_0 \rangle \approx (\mathbf{x}_k - \mathbf{x}^*)^T (\alpha \nabla^2 f(\mathbf{x}^*))^2 (\mathbf{x}_k - \mathbf{x}^*) \quad (263)$$

$$\begin{aligned} V_2 &= \alpha \langle \mathbf{y}_1 - \mathbf{y}_0, \nabla f(\mathbf{y}_1) - \nabla f(\mathbf{y}_0) \rangle \\ &\approx (\mathbf{x}_k - \mathbf{x}^*)^T (\alpha \nabla^2 f(\mathbf{x}^*))^3 (\mathbf{x}_k - \mathbf{x}^*) \end{aligned} \quad (264)$$

Next we use the following substitution:

$$\mathbf{x}_k - \mathbf{x}^* = \|\mathbf{x}_k - \mathbf{x}^*\| \left( \sum_{i \in \mathcal{N}_S} \theta_i^s \mathbf{v}_i(0) + \sum_{j \in \mathcal{N}_{US}} \theta_j^{us} \mathbf{v}_j(0) \right) \quad (265)$$

where we have that  $\|\mathbf{x}_k - \mathbf{x}^*\| \theta_i^s = \langle (\mathbf{x}_k - \mathbf{x}^*), \mathbf{v}_i(0) \rangle$ ,  $\|\mathbf{x}_k - \mathbf{x}^*\| \theta_j^{us} = \langle (\mathbf{x}_k - \mathbf{x}^*), \mathbf{v}_j(0) \rangle$  and  $\mathbf{v}_i(0), \mathbf{v}_j(0)$  are the eigenvectors of the scaled Hessian  $\alpha \nabla^2 f(\mathbf{x}^*)$ . On further simplifying  $V_1, V_2$  using (265) we get:

$$V_1 \approx \|\mathbf{x}_k - \mathbf{x}^*\|^2 \left( \sum_{i \in \mathcal{N}_S} (\lambda_i^s)^2 (\theta_i^s)^2 + \sum_{j \in \mathcal{N}_{US}} (\lambda_j^{us})^2 (\theta_j^{us})^2 \right) \quad (266)$$

$$V_2 \approx \|\mathbf{x}_k - \mathbf{x}^*\|^2 \left( \sum_{i \in \mathcal{N}_S} (\lambda_i^s)^3 (\theta_i^s)^2 + \sum_{j \in \mathcal{N}_{US}} (\lambda_j^{us})^3 (\theta_j^{us})^2 \right) \quad (267)$$

where  $\lambda_i^s$  and  $\lambda_j^{us}$  are the eigenvalues of stable subspace  $\mathcal{E}_S$  and unstable subspace  $\mathcal{E}_{US}$  of the scaled Hessian  $\alpha \nabla^2 f(\mathbf{x}^*)$  respectively. These eigenvalues are bounded by :

$$\frac{\beta}{L} \leq \lambda_i^s \leq 1 \quad (268)$$

$$-1 \leq \lambda_j^{us} \leq -\frac{\beta}{L}. \quad (269)$$

Evaluating  $V_1 - V_2$  and using the fact that  $\|\mathbf{x}_k - \mathbf{x}^*\| \leq \frac{1}{\beta} \|\nabla f(\mathbf{x}_k)\| \leq \frac{L\varepsilon}{\beta}$  from (158), we get the following expression:

$$\begin{aligned} V_1 - V_2 &\lesssim \frac{\varepsilon^2}{\kappa^2} \left( \sum_{i \in \mathcal{N}_S} ((\lambda_i^s)^2 - (\lambda_i^s)^3) (\theta_i^s)^2 + \right. \\ &\quad \left. \sum_{j \in \mathcal{N}_{US}} ((\lambda_j^{us})^2 - (\lambda_j^{us})^3) (\theta_j^{us})^2 \right) \end{aligned} \quad (270)$$

where  $\kappa = \frac{\beta}{L}$ . Now, the function  $h(y) = y^2 - y^3$  attains a maximum value of  $\frac{4}{27}$  in the interval  $y \in (0, 1]$  and a maximum value of 2 in the interval  $y \in [-1, 0)$ . Substituting  $y = \lambda_i^s$  in the interval  $y \in (0, 1]$  and  $y = \lambda_j^{us}$  in the interval  $y \in [-1, 0)$ , the upper bound for (270) becomes:

$$V_1 - V_2 \lesssim \frac{\varepsilon^2}{\kappa^2} \left( \sum_{i \in \mathcal{N}_S} \frac{4}{27} (\theta_i^s)^2 + \sum_{j \in \mathcal{N}_{US}} 2 (\theta_j^{us})^2 \right) \quad (271)$$

$$V_1 - V_2 \lesssim \frac{\varepsilon^2}{\kappa^2} \left( \frac{4}{27} - \frac{4}{27} \left( \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \right) + 2 \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \right) \quad (272)$$

$$V_1 - V_2 \lesssim \frac{\varepsilon^2}{\kappa^2} \left( \frac{4}{27} + \frac{50}{27} \left( \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \right) \right) \quad (273)$$

$$\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \gtrsim \frac{27(V_1 - V_2)\kappa^2 - 4}{50}. \quad (274)$$

The right-hand side in (274) can be considered as the lower bound estimate for  $\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2$ . Now, the sufficient condition for escaping the saddle neighborhood comes from the minimum unstable subspace projection value in (74). Let  $P_{min}(\varepsilon)$  be a function of  $\varepsilon$  equal to the lower bound from (74) where  $P_{min}(\varepsilon) = \Theta\left(\frac{1}{\log(\varepsilon^{-1})}\right)$ , then with the con-

dition  $\frac{27(V_1-V_2)\kappa^2}{\varepsilon^2} - 4 > P_{min}(\varepsilon)$  and (274), we can guarantee  $\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 \gtrsim P_{min}(\varepsilon)$  which implies that we have a sufficient unstable projection value to escape saddle region in almost linear time.

Recall that while developing the condition (274) we had dropped  $\mathcal{O}(\varepsilon)$  term in (256) to get approximate Hessian. However if we had retained that  $\mathcal{O}(\varepsilon)$  term, the condition (274) would be  $\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 > \frac{27(V_1-V_2)\kappa^2}{\varepsilon^2} - 4 + \mathcal{O}(\varepsilon)$  and then instead of checking  $\frac{27(V_1-V_2)\kappa^2}{\varepsilon^2} - 4 > P_{min}(\varepsilon)$  we will have to check the inequality  $\frac{27(V_1-V_2)\kappa^2}{\varepsilon^2} - 4 > P_{min}(\varepsilon) + \mathcal{O}(\varepsilon)$ . But since  $\mathcal{O}(\varepsilon)$  term is insignificant w.r.t.  $P_{min}(\varepsilon) = \Theta\left(\frac{1}{\log(\varepsilon^{-1})}\right)$ , we do not lose anything with dropping  $\mathcal{O}(\varepsilon)$  term in the first step.

Notice that the curvature condition from the step 15 in Algorithm 1 checks the inequality  $\frac{27(V_1-V_2)\kappa^2}{\varepsilon^2} - 4 < P_{min}(\varepsilon)$  which if true could imply  $\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 < P_{min}(\varepsilon)$ . Then the gradient trajectory may not necessarily have linear exit time from saddle neighborhood. Hence, we solve the eigenvector problem given by:

$$\mathbf{x}_{k+1} \in \arg \min_{\|\mathbf{x} - \mathbf{x}_k\| = \frac{\|\nabla f(\mathbf{x}_k)\|}{\beta}} \left( \frac{1}{2} (\mathbf{x} - \mathbf{x}_k)^T \mathbf{H} (\mathbf{x} - \mathbf{x}_k) \right) \quad (275)$$

which gives a solution with sufficient unstable projection. Notice that a possible solution to the unconstrained problem:

$$\mathbf{x}_{k+1} \in \arg \min_{\mathbf{x}} \left( \frac{1}{2} (\mathbf{x} - \mathbf{x}_k)^T \mathbf{H} (\mathbf{x} - \mathbf{x}_k) \right) \quad (276)$$

can be given by  $\mathbf{x}_{k+1} - \mathbf{x}_k = b \|\mathbf{x}_k - \mathbf{x}^*\| \mathbf{e}_j^{us}$  where  $\mathbf{e}_j^{us}$  is any eigenvector of the scaled Hessian  $\mathbf{H} = \alpha \nabla^2 f(\mathbf{x}_k) \approx \alpha \nabla^2 f(\mathbf{x}^*)$  corresponding to its least eigenvalue and  $b$  is any scalar. Although any vector in the subspace formed by the eigenvectors corresponding to the minimum eigenvalue can be used instead of  $\mathbf{e}_j^{us}$ , for sake of simplicity of the proof, we use the direction  $\mathbf{e}_j^{us}$ . Hence from the unconstrained eigenvector problem (276), we can write  $\mathbf{x}_{k+1} - \mathbf{x}^* = \mathbf{x}_k - \mathbf{x}^* + b \|\mathbf{x}_k - \mathbf{x}^*\| \mathbf{e}_j^{us}$ .

Using the substitution  $\mathbf{x}_k - \mathbf{x}^* = \|\mathbf{x}_k - \mathbf{x}^*\| \left( \sum_{i \in \mathcal{N}_S} \theta_i^s \mathbf{v}_i(0) + \sum_{j \in \mathcal{N}_{US}} \theta_j^{us} \mathbf{v}_j(0) \right)$  as before from (265) we get:

$$\begin{aligned} \mathbf{x}_{k+1} - \mathbf{x}^* &= \|\mathbf{x}_k - \mathbf{x}^*\| \left( \sum_{i \in \mathcal{N}_S} \theta_i^s \mathbf{v}_i(0) + \sum_{j \in \mathcal{N}_{US}} \theta_j^{us} \mathbf{v}_j(0) \right) \\ &\quad + b \|\mathbf{x}_k - \mathbf{x}^*\| \mathbf{e}_j^{us} \\ &= \|\mathbf{x}_k - \mathbf{x}^*\| \left( \sum_{i \in \mathcal{N}_S} \theta_i^s \mathbf{v}_i(0) + \sum_{j \in \mathcal{N}_{US}} \theta_j^{us} \mathbf{v}_j(0) \right) \end{aligned} \quad (277)$$

$$+ b \|\mathbf{x}_k - \mathbf{x}^*\| \left( \mathbf{v}_l(0) + \mathcal{O}(\varepsilon) \right) \quad (278)$$

$$\begin{aligned} &= \|\mathbf{x}_k - \mathbf{x}^*\| \sqrt{1+b^2} \left( \sum_{i \in \mathcal{N}_S} \frac{\theta_i^s}{\sqrt{1+b^2}} \mathbf{v}_i(0) \right. \\ &\quad \left. + \sum_{j \in \mathcal{N}_{US}} \frac{\theta_j^{us}}{\sqrt{1+b^2}} \mathbf{v}_j(0) + \frac{b}{\sqrt{1+b^2}} \mathbf{v}_l(0) \right) + \mathcal{O}(\varepsilon^2) \end{aligned} \quad (279)$$

$$\begin{aligned} &= \|\mathbf{x}_k - \mathbf{x}^*\| \sqrt{1+b^2} \left( \sum_{i \in \mathcal{N}_S} \tilde{\theta}_i^s \mathbf{v}_i(0) + \sum_{j \in \mathcal{N}_{US}} \tilde{\theta}_j^{us} \mathbf{v}_j(0) \right) \\ &\quad + \mathcal{O}(\varepsilon^2). \end{aligned} \quad (280)$$

where we have  $\sum_{i \in \mathcal{N}_S} (\tilde{\theta}_i^s)^2 + \sum_{j \in \mathcal{N}_{US}} (\tilde{\theta}_j^{us})^2 = 1$  for some positive  $\tilde{\theta}_i^s, \tilde{\theta}_j^{us}$ . Notice that we used the eigenvector perturbation bound  $\mathbf{e}_j^{us} = \mathbf{v}_l(0) + \mathcal{O}(\varepsilon)$  in the second step and  $\mathbf{v}_l(0)$  corresponds to the eigenvector for the smallest eigenvalue of  $\alpha \nabla^2 f(\mathbf{x}^*)$ . Notice that  $l \in \mathcal{N}_{US}$  where  $l$  is the index of  $\mathbf{v}_l(0)$  provided  $\mathbf{x}_k$  lies within some saddle neighborhood and not in a local minimum neighborhood. If  $\mathbf{x}_k$  were in a local minimum neighborhood, then the unstable subspace would have been the null space. Finally, in the second last step we normalized by dividing with  $\sqrt{1+b^2}$  because we require the condition:

$$\sum_{i \in \mathcal{N}_S} \left( \frac{\theta_i^s}{\sqrt{1+b^2}} \right)^2 + \underbrace{\sum_{j \in \mathcal{N}_{US}} \left( \frac{\theta_j^{us}}{\sqrt{1+b^2}} \right)^2 + \left( \frac{b}{\sqrt{1+b^2}} \right)^2}_{U_1} = 1 \quad (281)$$

where we have that  $\sum_{i \in \mathcal{N}_S} (\theta_i^s)^2 + \sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 = 1$ . From (279) and (280) using coefficient comparison, it can be checked that  $\frac{\theta_i^s}{\sqrt{1+b^2}} = \tilde{\theta}_i^s + \mathcal{O}(\varepsilon^2)$  for all  $i \in \mathcal{N}_S$ . Using this relation in (281) we get that  $U_1 = \sum_{j \in \mathcal{N}_{US}} (\tilde{\theta}_j^{us})^2 + \mathcal{O}(\varepsilon^2)$ . Next, dropping  $\mathcal{O}(\varepsilon^2)$  term<sup>15</sup> from the right-hand side of (280), we have:

$$\mathbf{x}_{k+1} - \mathbf{x}^* \approx \|\mathbf{x}_k - \mathbf{x}^*\| \sqrt{1+b^2} \left( \sum_{i \in \mathcal{N}_S} \tilde{\theta}_i^s \mathbf{v}_i(0) + \sum_{j \in \mathcal{N}_{US}} \tilde{\theta}_j^{us} \mathbf{v}_j(0) \right) \quad (282)$$

where  $\sum_{j \in \mathcal{N}_{US}} (\tilde{\theta}_j^{us})^2$  can be considered as the new unstable projection value of  $(\mathbf{x}_{k+1} - \mathbf{x}^*)$  and  $\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \approx \|\mathbf{x}_k - \mathbf{x}^*\| \sqrt{1+b^2}$ . Now, we require that the future gradient trajectory that starts from the point  $\mathbf{x}_{k+1}$  escapes the ball  $\mathcal{B}_{\tilde{\varepsilon}}(\mathbf{x}^*)$  in linear time where  $\tilde{\varepsilon} = \|\mathbf{x}_k - \mathbf{x}^*\| \sqrt{1+b^2}$ . Therefore we get that:

$$U_1 \approx \sum_{j \in \mathcal{N}_{US}} (\tilde{\theta}_j^{us})^2 \geq P_{min}(\tilde{\varepsilon}) \quad (283)$$

$$\Rightarrow \sum_{j \in \mathcal{N}_{US}} \left( \frac{\theta_j^{us}}{\sqrt{1+b^2}} \right)^2 + \left( \frac{b}{\sqrt{1+b^2}} \right)^2 \gtrsim P_{min}(\tilde{\varepsilon}) \quad (284)$$

$$= P_{min}(\|\mathbf{x}_k - \mathbf{x}^*\| \sqrt{1+b^2}) \quad (285)$$

<sup>15</sup>The  $\mathcal{O}(\varepsilon^2)$  term will be insignificant w.r.t.  $P_{min}(\tilde{\varepsilon})$  appearing in (283) which is of the order  $\Theta\left(\frac{1}{\log(\tilde{\varepsilon}^{-1})}\right)$  from Theorem 1 due to the fact that  $\tilde{\varepsilon} > \varepsilon$ , the function  $P_{min}(\varepsilon) = \Theta\left(\frac{1}{\log(\varepsilon^{-1})}\right)$  monotonically increases with  $\varepsilon$  and so  $P_{min}(\tilde{\varepsilon}) > P_{min}(\varepsilon) > \mathcal{O}(\varepsilon^2)$ .

$$> P_{\min} \left( \|\nabla f(\mathbf{x}_k)\| \frac{\sqrt{1+b^2}}{L} \right) \quad (286)$$

where in the last step we used  $P_{\min}(\|\mathbf{x}_k - \mathbf{x}^*\| \sqrt{1+b^2}) > P_{\min} \left( \|\nabla f(\mathbf{x}_k)\| \frac{\sqrt{1+b^2}}{L} \right)$  due to the fact that the function  $P_{\min}(\varepsilon) = \Theta \left( \frac{1}{\log(\varepsilon^{-1})} \right)$  monotonically increases with  $\varepsilon$  from (74) along with the property that  $\|\nabla f(\mathbf{x}_k)\| \leq L \|\mathbf{x}_k - \mathbf{x}^*\|$ . Now (286) will hold true whenever:

$$\left( \frac{b}{\sqrt{1+b^2}} \right)^2 > P_{\min} \left( \|\nabla f(\mathbf{x}_k)\| \frac{\sqrt{1+b^2}}{L} \right) \quad (287)$$

$$b > \frac{\sqrt{P_{\min} \left( \|\nabla f(\mathbf{x}_k)\| \frac{\sqrt{1+b^2}}{L} \right)}}{\sqrt{1 - P_{\min} \left( \|\nabla f(\mathbf{x}_k)\| \frac{\sqrt{1+b^2}}{L} \right)}}. \quad (288)$$

It can be checked that (288) will hold true for any positive  $b$  as long as it is bounded away from  $\varepsilon$ . Finally in the substitution  $\mathbf{x}_{k+1} - \mathbf{x}_k = b \|\mathbf{x}_k - \mathbf{x}^*\| \mathbf{e}_j^{us}$ , we can use the lower bound  $\|\nabla f(\mathbf{x}_k)\| \geq \beta \|\mathbf{x}_k - \mathbf{x}^*\|$  from (158) and the gradient Lipschitz bound  $\|\nabla f(\mathbf{x}_k)\| \leq L \|\mathbf{x}_k - \mathbf{x}^*\|$  to get the range  $\frac{\|\nabla f(\mathbf{x}_k)\|}{L \|\mathbf{x}_k - \mathbf{x}^*\|} \leq b \leq \frac{\|\nabla f(\mathbf{x}_k)\|}{\beta \|\mathbf{x}_k - \mathbf{x}^*\|}$ . Selecting the upper bound of  $b$  gives  $\mathbf{x}_{k+1} - \mathbf{x}_k = \frac{\|\nabla f(\mathbf{x}_k)\|}{\beta} \mathbf{e}_j^{us}$  provided  $\frac{\beta}{L} \gg 0$ . This particular choice of  $b$  is less conservative though it should be selected carefully and the selection criterion may vary from one problem to another. For the particular case of well-conditioned saddle neighborhood, a large  $b$  and hence a large step size can be afforded. Notice that  $\frac{\beta}{L} \leq b \leq \frac{L}{\beta}$  and any  $b$  in this range will satisfy (288) provided  $\frac{\beta}{L} \gg 0$ . Since  $\mathbf{x}_{k+1}$  is the desired solution, taking norm on both sides of  $\mathbf{x}_{k+1} - \mathbf{x}_k = \frac{\|\nabla f(\mathbf{x}_k)\|}{\beta} \mathbf{e}_j^{us}$  gives the constraint  $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| = \frac{\|\nabla f(\mathbf{x}_k)\|}{\beta}$  in the Step 17 of Algorithm 1.

Since evaluating the eigenvector  $\mathbf{e}_j^{us}$  will involve Hessian inversion operations, it will be solved in polynomial time though this step is invoked only once in the saddle neighborhood if required and hence does not add much computational complexity per iteration<sup>16</sup> (only  $\Theta(n^2 \log_{\lambda_1/\lambda_2}(\varepsilon^{-1}))$  complexity per saddle point).

Recall that the entire algorithmic analysis was carried out assuming there is just one eigenvector  $\mathbf{e}_j^{us}$  corresponding to the smallest eigenvalue of the Hessian  $\nabla^2 f(\mathbf{x}^*)$ . However, the same analysis can be done for the case of a subspace corresponding to the smallest eigenvalue. The bounds on  $b$  will still be the same however the steps involved are somewhat tedious and lengthy hence purposefully left out from the proof.

For the case of a local minimum we will have  $\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2 = 0$  since there is no unstable subspace. Substituting it in (274) yields:

$$\frac{4\varepsilon^2}{27\kappa^2} \gtrsim V_1 - V_2. \quad (289)$$

<sup>16</sup>The step 17 can be solved easily using power iteration type method which incurs a computation time of  $\Theta(n^2 \log_{\lambda_1/\lambda_2}(\varepsilon^{-1}))$  (here  $\lambda_1, \lambda_2$  are the top two dominant eigenvalues of  $\mathbf{I} - \alpha \nabla^2 f(\mathbf{x}_k)$  with  $\lambda_1 > \lambda_2$  and  $\varepsilon$  is the desired accuracy for power iteration).

Hence for  $\frac{4\varepsilon^2}{27\kappa^2} \lesssim V_1 - V_2$  we cannot have a local minimum neighborhood. Hence if (289) holds, then the region can be both a saddle neighborhood or a local minimum region. Therefore, the Step 15 in Algorithm 1 also checks if  $\frac{4\varepsilon^2}{27\kappa^2} < V_1 - V_2$  so as to rule out the possibility of local minimum. If however we have the inequality  $\frac{4\varepsilon^2}{27\kappa^2} > V_1 - V_2$  then a secondary condition  $\lambda_{\min}(\mathbf{H}) < 0$  ascertains it as a saddle neighborhood. This completes the proof. ■

#### Proof of Lemma 8

It can be very easily established that  $f(\mathbf{x}_{K+1}) < f(\mathbf{x}_K)$  where  $\mathbf{x}_{K+1}$  comes from the Step 17 in Algorithm 1.

Since  $\mathbf{x}_{K+1}$  is generated from Step 17 of Algorithm 1 we can use the particular update  $\mathbf{x}_{K+1} - \mathbf{x}_K = \frac{\|\nabla f(\mathbf{x}_K)\|}{\beta} \mathbf{e}_j^{us}$  (the more general update 17 is avoided for sake of simplicity) where  $\mathbf{e}_j^{us}$  is an eigenvector of  $\nabla^2 f(\mathbf{x}_K)$  belonging to its unstable subspace and  $\langle \mathbf{e}_j^{us}, \mathbf{x}_K - \mathbf{x}^* \rangle \lesssim \Theta \left( \frac{\varepsilon}{\sqrt{\log(\varepsilon^{-1})}} \right)$  (this approximate bound implies  $\mathbf{x}_K - \mathbf{x}^*$  does not have the required unstable subspace projection value from Theorem 1). As a consequence for  $\|\nabla f(\mathbf{x}_K)\| = \Theta(\varepsilon)$  we will have  $\langle \nabla f(\mathbf{x}_K), \mathbf{x}_{K+1} - \mathbf{x}_K \rangle \lesssim \Theta \left( \frac{\varepsilon^2}{\sqrt{\log(\varepsilon^{-1})}} \right)$  from the following steps where we use the substitutions  $\nabla f(\mathbf{x}_K) = (\nabla^2 f(\mathbf{x}^*) + \mathcal{O}(\varepsilon))(\mathbf{x}_K - \mathbf{x}^*)$  and  $\nabla^2 f(\mathbf{x}_K) = (\nabla^2 f(\mathbf{x}^*) + \mathcal{O}(\varepsilon))$  from Hessian Lipschitz condition.

$$\langle \nabla f(\mathbf{x}_K), \mathbf{x}_{K+1} - \mathbf{x}_K \rangle = \langle \nabla f(\mathbf{x}_K), \frac{\|\nabla f(\mathbf{x}_K)\|}{\beta} \mathbf{e}_j^{us} \rangle \quad (290)$$

$$= \frac{\|\nabla f(\mathbf{x}_K)\|}{\beta} \langle \mathbf{e}_j^{us}, (\nabla^2 f(\mathbf{x}^*) + \mathcal{O}(\varepsilon))(\mathbf{x}_K - \mathbf{x}^*) \rangle \quad (291)$$

$$= \frac{\|\nabla f(\mathbf{x}_K)\|}{\beta} \langle \mathbf{e}_j^{us}, (\nabla^2 f(\mathbf{x}_K) + \mathcal{O}(\varepsilon))(\mathbf{x}_K - \mathbf{x}^*) \rangle \quad (292)$$

$$= \frac{\|\nabla f(\mathbf{x}_K)\|}{\beta} \langle \lambda_j^{us} \mathbf{e}_j^{us}, (\mathbf{x}_K - \mathbf{x}^*) \rangle + \mathcal{O}(\varepsilon^3) \lesssim \Theta \left( \frac{\varepsilon^2}{\sqrt{\log(\varepsilon^{-1})}} \right) \quad (293)$$

where  $\nabla^2 f(\mathbf{x}_K) \mathbf{e}_j^{us} = \lambda_j^{us} \mathbf{e}_j^{us}$  and  $\Theta \left( \frac{\varepsilon^2}{\sqrt{\log(\varepsilon^{-1})}} \right) > \mathcal{O}(\varepsilon^3)$ .

Finally using Hessian Lipschitz condition for  $\mathbf{x}_{K+1}$  about  $\mathbf{x}_K$  along with (293) we get:

$$f(\mathbf{x}_{K+1}) \leq f(\mathbf{x}_K) + \langle \nabla f(\mathbf{x}_K), \mathbf{x}_{K+1} - \mathbf{x}_K \rangle + \underbrace{\frac{1}{2} \langle (\mathbf{x}_{K+1} - \mathbf{x}_K), \nabla^2 f(\mathbf{x}_K) (\mathbf{x}_{K+1} - \mathbf{x}_K) \rangle}_{<0} + \underbrace{\frac{M}{6} \|\mathbf{x}_{K+1} - \mathbf{x}_K\|^3}_{\Theta(\varepsilon^3)} \quad (294)$$

$$\leq f(\mathbf{x}_K) + \Theta \left( \frac{\varepsilon^2}{\sqrt{\log(\varepsilon^{-1})}} \right) + \frac{\|\nabla f(\mathbf{x}_K)\|^2}{2\beta^2} \underbrace{\langle \mathbf{e}_j^{us}, \nabla^2 f(\mathbf{x}_K) \mathbf{e}_j^{us} \rangle}_{<-\beta} + \Theta(\varepsilon^3) \quad (295)$$

$$\leq f(\mathbf{x}_K) + \Theta \left( \frac{\varepsilon^2}{\sqrt{\log(\varepsilon^{-1})}} \right) - \Theta(\|\nabla f(\mathbf{x})\|^2) + \Theta(\varepsilon^3) \quad (296)$$

$$\leq f(\mathbf{x}_K) + \Theta \left( \frac{\varepsilon^2}{\sqrt{\log(\varepsilon^{-1})}} \right) - \Theta(\varepsilon^2) + \Theta(\varepsilon^3) \\ = f(\mathbf{x}_K) + \Theta \left( \frac{\varepsilon^2}{\sqrt{\log(\varepsilon^{-1})}} \right) - \Theta(\varepsilon^2) \quad (297)$$

where we used the facts that  $\|\mathbf{x}_{K+1} - \mathbf{x}_K\| = \Theta(\varepsilon)$ ,  $\|\nabla f(\mathbf{x}_K)\| = \Theta(\varepsilon)$ ,  $\langle \mathbf{e}_j^{us}, \nabla^2 f(\mathbf{x}_K) \mathbf{e}_j^{us} \rangle = \lambda_j^{us} < -\beta$  and  $\frac{1}{2} \langle (\mathbf{x}_{K+1} - \mathbf{x}_K), \nabla^2 f(\mathbf{x}_K) (\mathbf{x}_{K+1} - \mathbf{x}_K) \rangle < 0$  from the Step 17 of Algorithm 1. Now for sufficiently small  $\varepsilon$ , the term  $\frac{\varepsilon^2}{\sqrt{\log(e^{-1})}} \rightarrow 0$  much faster than  $\varepsilon^2$  goes to 0. Hence for sufficiently small  $\varepsilon$  we will have  $f(\mathbf{x}_{K+1}) < f(\mathbf{x}_K)$ . For all other iterations when gradient descent update is used, the sequence  $\{f(\mathbf{x}_k)\}$  decreases monotonically. ■

## APPENDIX G ASYMPTOTIC CONVERGENCE

### Proof of Lemma 9

Let  $\{\mathbf{x}_k\}$  be the sequence generated by Algorithm 1. Then by Lemma 7 this sequence exits the  $\varepsilon$  neighborhood of any strict saddle point  $\mathbf{x}^*$  of a locally analytic Morse function in approximately linear time where  $\varepsilon$  is bounded from Theorem 1. Further,  $\varepsilon$  can be chosen in a way such that if the iterate  $\mathbf{x}_k$  exits the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  at some  $k = K$  then the trajectory of  $\{\mathbf{x}_k\}$  cannot return to this neighborhood  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  for any  $k > K$ . Such a choice of  $\varepsilon$  is guaranteed from Lemma 5. Hence the sequence  $\{\mathbf{x}_k\}$  cannot converge to the strict saddle point  $\mathbf{x}^*$  which completes the proof of the first part of the lemma.

For the second part notice that if any subsequence  $\{\mathbf{x}_{m_k}\}$  of the sequence  $\{\mathbf{x}_k\}$  converges to  $\mathbf{x}^*$  then  $\mathbf{x}^* \in \{\mathbf{x}_{m_k}\}$  i.o. or equivalently  $\mathbf{x}^* \in \{\mathbf{x}_k\}$  i.o. which implies  $\mathbf{x}^*$  is an accumulation point of the sequence  $\{\mathbf{x}_k\}$ . Since the mapping  $\text{id} - \alpha \nabla f$  is continuous where  $\text{id}$  is the identity (matrix) map, the set of accumulation points for the sequence generated by the gradient descent update  $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)$  will be connected (Lemma 21). Next, the set of accumulation points of the sequence  $\{\mathbf{x}_k\}$  generated by the gradient descent update  $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)$  will be the critical points of  $f(\cdot)$  since the critical points of  $f(\cdot)$  are the only fixed points of the mapping  $\text{id} - \alpha \nabla f$ . Since the second order step 17 in Algorithm 1 cannot generate fixed points of the sequence  $\{\mathbf{x}_k\}$ , it suffices to only consider the gradient descent step. Since  $f(\cdot)$  is Morse, it has isolated critical points and so the accumulation points of the sequence  $\{\mathbf{x}_k\}$  generated by the gradient descent update  $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)$  will be disconnected. Therefore the sequence generated by the update  $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)$  can have only a single isolated accumulation point since the only connected subsets of a set with isolated points are singletons. Next any sequence converges if and only if it has a unique limit/ accumulation point. Since  $\mathbf{x}^*$  is an accumulation point of the sequence  $\{\mathbf{x}_k\}$  and also a critical point of  $f(\cdot)$ , it will be the only isolated accumulation point of the sequence  $\{\mathbf{x}_k\}$  due to the fact that in the sequence  $\{\mathbf{x}_k\}$  second order step is invoked only once in  $\varepsilon$  neighborhood of  $\mathbf{x}^*$  and from there onward gradient descent is used for all  $k$ . Hence we get  $\mathbf{x}_k \rightarrow \mathbf{x}^*$ , a contradiction to first part of this lemma. Therefore no subsequence  $\{\mathbf{x}_{m_k}\}$  of the sequence  $\{\mathbf{x}_k\}$  can converge to the strict saddle point  $\mathbf{x}^*$  which completes the proof. ■

### Proof of Lemma 10

The sequence  $\{f(\mathbf{x}_k)\}$  decreases monotonically from Lemma 8. Since  $f$  is coercive i.e.  $\lim_{\|\mathbf{x}\| \rightarrow \infty} f(\mathbf{x}) = \infty$  and  $f$  is

continuous (and hence lower semi-continuous), we will have  $f(\mathbf{x}) \geq \inf_{\mathbf{x}} f(\mathbf{x}) > -\infty$  i.e. the infimum of the function values exists [55]. Then by the monotone convergence theorem,  $\lim_{k \rightarrow \infty} f(\mathbf{x}_k)$  exists and is finite. Since  $f$  is coercive and continuous, its sublevel sets given by  $\{\mathbf{x} \mid f(\mathbf{x}) \leq b\}$  for any  $b < \infty$  are compact. Since  $\lim_{k \rightarrow \infty} f(\mathbf{x}_k)$  exists and is finite, by the monotonicity of  $\{f(\mathbf{x}_k)\}$  it will belong to the compact sublevel set  $\{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$ , which completes the proof. ■

### Proof of Lemma 11

Let  $\mathbf{x}_0$  be the initialization of Algorithm 1, then by the previous lemma the sequence  $\{f(\mathbf{x}_k)\}$  converges over the compact sublevel set  $\{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$ . Combining this fact and the monotonicity of the sequence  $\{f(\mathbf{x}_k)\}$  we have that  $\mathbf{x}_k \in \{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$  for all  $k$ . Since a Morse function on a compact manifold has finitely many critical points [42], the compact sublevel set  $\{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$  can have at most finitely many saddle points. ■

### Proof of Theorem 5

In order to prove asymptotic convergence of the sequence  $\{\mathbf{x}_k\}$  generated by Algorithm 1 to a critical point we only need to show that the sequence  $\{\mathbf{x}_k\}$  satisfies all the conditions from Theorem 4. First, from Lemma 11 all points of the sequence  $\{\mathbf{x}_k\}$  are contained in a compact set  $D \subset X$  where  $D = \{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$  and  $X = \mathbb{R}^n$ . Next, the continuous function  $Z = f$  satisfies the strict monotonicity property where  $\{f(\mathbf{x}_k)\}$  is a strictly decreasing sequence provided  $\mathbf{x}_k \notin S$  and the solution set  $S \subset D$  is the set of critical points of  $f$  with  $f(\mathbf{x}_k) = f(\mathbf{x}_{k+1})$  for  $\mathbf{x}_k \in S$ .

Finally we are left to show that the mapping  $\mathbf{A}$  where  $\mathbf{x}_{k+1} = \mathbf{A}(\mathbf{x}_k)$  is closed outside  $S$ . It is easy to check that the mapping  $\mathbf{A}$  from Algorithm 1 is compact when  $\mathbf{A} := \text{id} - \alpha \nabla f$ . Notice that for the gradient descent update, the map  $\mathbf{A} := \text{id} - \alpha \nabla f$  is continuous due to  $f \in \mathcal{C}^2$ . Since  $\mathbf{x}_k \in D = \{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$  for all  $k$ , the map  $\mathbf{A} := \text{id} - \alpha \nabla f$  takes  $D$  to itself, i.e.  $\mathbf{A} : D \mapsto D$  where  $D$  is compact and Hausdorff [17]. Then by the closed map lemma (Lemma A.52 in [56]),  $\mathbf{A} := \text{id} - \alpha \nabla f$  is a closed map in  $D$  and hence closed in  $D \setminus S$ .

From the second-order step in Algorithm 1,  $\mathbf{x}_{k+1} \in \arg \min_{\|\mathbf{x} - \mathbf{x}_k\| = \frac{\|\nabla f(\mathbf{x}_k)\|}{\beta}} \left( \frac{1}{2} (\mathbf{x} - \mathbf{x}_k)^T \nabla^2 f(\mathbf{x}_k) (\mathbf{x} - \mathbf{x}_k) \right) = \mathbf{A}(\mathbf{x}_k)$  and it remains to show that this mapping is continuous. The second-order step can be simplified as  $\mathbf{x}_{k+1} \in \mathbf{x}_k - \frac{\|\nabla f(\mathbf{x}_k)\|}{\beta} \arg \min_{\|\mathbf{x}\| > 0} \frac{\mathbf{x}^T \nabla^2 f(\mathbf{x}_k) \mathbf{x}}{\|\mathbf{x}\|^2}$ . Since  $f$  is Hessian Lipschitz, the eigenvectors of  $\nabla^2 f(\mathbf{x})$  will vary continuously with  $\mathbf{x}$ ; hence  $\arg \min_{\|\mathbf{x}\| > 0} \frac{\mathbf{x}^T \nabla^2 f(\mathbf{x}_k) \mathbf{x}}{\|\mathbf{x}\|^2}$  is a continuous function and  $\|\nabla f(\cdot)\|$  is a continuous function by continuity of  $\nabla f(\cdot)$  and norm. Product of continuous functions is continuous therefore the map  $\mathbf{A}$  associated with the second order step is continuous. As before the map  $\mathbf{A}$  takes  $D$  to itself where  $D$  is compact and Hausdorff. Then by the closed map lemma, for the second order step,  $\mathbf{A}$  is closed in  $D \setminus S$ . Since  $\{\mathbf{x}_k\} \subset D$ , which is compact, there exists a convergent subsequence  $\{\mathbf{x}_{m_k}\}$  of  $\{\mathbf{x}_k\}$

<sup>17</sup>A Hausdorff space is a topological space with a separation property: any two distinct points can be separated by disjoint open sets.

and from Theorem 4 we have  $\lim_{k \rightarrow \infty} \mathbf{x}_{m_k} \in S \subset D$  where  $S$  is the set of critical points of  $f$ .

Finally from Lemma 9 since  $\{\mathbf{x}_{m_k}\}$  does not converge to any strict saddle point, we have  $\mathbf{x}_{m_k} \rightarrow \mathbf{x}^*$ , where  $\mathbf{x}^*$  is a local minimum. The fact that  $f(\cdot)$  has a local minimum holds because  $f(\cdot)$  is coercive, continuous and so has a global minimum and therefore has at least one local minimum. Next, the local minima of  $f(\cdot)$  belong to the set of critical points of  $f(\cdot)$  because  $f(\cdot)$  is  $\mathcal{C}^2$  smooth and hence  $\mathcal{C}^1$  smooth and we have that local minima of any  $\mathcal{C}^1$  smooth function are its critical points. Since  $\mathbf{x}^* \in \{\mathbf{x}_{m_k}\}$  i.o. hence  $\mathbf{x}^* \in \{\mathbf{x}_k\}$  i.o. which implies  $\mathbf{x}^*$  is an accumulation point of the sequence  $\{\mathbf{x}_k\}$ . Since the mapping  $\text{id} - \alpha \nabla f$  is continuous, the set of accumulation points for the sequence generated by the gradient descent update  $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)$  will be connected (Lemma 21). Next, the set of accumulation points of the sequence  $\{\mathbf{x}_k\}$  generated by the gradient descent update  $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)$  will be the critical points of  $f(\cdot)$  since the critical points of  $f(\cdot)$  are the only fixed points of the mapping  $\text{id} - \alpha \nabla f$ . Since the second order step 17 in Algorithm 1 cannot generate fixed points of the sequence  $\{\mathbf{x}_k\}$ , it suffices to only consider the gradient descent step. Moreover the sequence  $\{\mathbf{x}_k\}$  can encounter at most finitely many saddle point neighborhoods from Lemma 11 where each such  $\varepsilon$  saddle neighborhood is visited only once by Lemma 5. Since the second order step 17 in Algorithm 1 is only invoked once at worst in every  $\varepsilon$  saddle neighborhood, these second order steps can be at most finite in number implying that the sequence  $\{\mathbf{x}_k\}$  eventually follows the gradient descent update. Therefore working only with the set of accumulation points for the sequence generated by the gradient descent update  $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)$  is justified.

Since  $f(\cdot)$  is Morse, it has isolated critical points and so the accumulation points of the sequence  $\{\mathbf{x}_k\}$  generated by the gradient descent update  $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)$  will be disconnected. Therefore the sequence  $\{\mathbf{x}_k\}$  can have only a single isolated accumulation point since the only connected subsets of a set with isolated points are singletons. Next, any sequence converges if and only if it has a unique limit/accumulation point. Since  $\mathbf{x}^*$  is an accumulation point of the sequence  $\{\mathbf{x}_k\}$  and also a critical point of  $f(\cdot)$ , it will be the only isolated accumulation point of the sequence  $\{\mathbf{x}_k\}$  due to the fact that in the sequence  $\{\mathbf{x}_k\}$  second order step is invoked only once in  $\varepsilon$  neighborhood of  $\mathbf{x}^*$  and from there onward gradient descent is used for all  $k$ . Hence we get  $\mathbf{x}_k \rightarrow \mathbf{x}^*$  which completes the proof. ■

**Lemma 21.** *The set of accumulation points of any sequence  $\{\mathbf{x}_k\}$  in some compact metric space  $X$  generated from the relation  $\mathbf{x}_{k+1} = N(\mathbf{x}_k)$  is connected, provided the mapping  $N$  is continuous in  $X$ .*

*Proof.* Let  $A$  be the set of accumulation points of  $\{\mathbf{x}_k\}$  and suppose  $A$  is not connected. We know that the set of all subsequential limits of a sequence in a metric space is closed [57]. Hence  $A$  is disconnected and closed and so is separated by closed compact sets  $C$  and  $D$ . Hence there exists an  $\varepsilon > 0$  such that for all  $\mathbf{x} \in C$  and  $\mathbf{y} \in D$  we have  $d(\mathbf{x}, \mathbf{y}) > 2\varepsilon$  where  $d(\cdot, \cdot)$  is the metric.

Let  $W_C$  be the union of all the open  $\varepsilon$ -balls about elements of

$C$  and define  $W_D$  similarly. By the definition of accumulation point,  $\{\mathbf{x}_k\}$  is frequently in  $W_C$  and frequently in  $W_D$ . Suppose for the sake of contradiction that there exists  $\mathbf{x} \in W_C \cap W_D$ . Then there must be  $\mathbf{p} \in C$  and  $\mathbf{q} \in D$  such that  $d(\mathbf{x}, \mathbf{p}) < \varepsilon$  and  $d(\mathbf{x}, \mathbf{q}) < \varepsilon$ . By the triangle inequality,  $d(\mathbf{p}, \mathbf{q}) < 2\varepsilon$ , a contradiction. Thus we conclude that  $W_C \cap W_D = \emptyset$ .

Since  $X$  is compact and  $N$  is continuous,  $N$  is uniformly continuous by the Heine-Cantor theorem. Also for any  $\mathbf{x}, \mathbf{y}$  in  $X$  and any  $\varepsilon > 0$  we have that  $d(N(\mathbf{x}), N(\mathbf{y})) < \varepsilon$  whenever  $d(\mathbf{x}, \mathbf{y}) < \delta$ .

Thus there is a  $\delta$  such that  $0 < \delta < \varepsilon$  such that for all  $\mathbf{x}, \mathbf{y} \in X$ ,  $d(\mathbf{x}, \mathbf{y}) < \delta$  implies  $d(N(\mathbf{x}), N(\mathbf{y})) < \varepsilon$ .

Let  $U$  be the union of all the open  $\delta$ -balls about elements of  $A$ . Then  $\{\mathbf{x}_k\}$  is eventually in  $U$ : if not, then there is a subsequence in  $X \setminus U$ , which is closed, hence compact, so there is an accumulation point in  $X \setminus U$ , a contradiction.

Let  $K_0 > K$ , where  $K$  is sufficiently large, be such that  $k \geq K_0$  implies  $\mathbf{x}_k \in U$ . Let  $k > K_0$ . Since  $\delta < \varepsilon$ ,  $\mathbf{x}_k \in U \subseteq W_C \cup W_D$ . Assume without loss of generality that  $\mathbf{x}_k \in U \cap W_C$ . By the definition of  $U$ , there is an element  $\mathbf{a} \in A$  such that  $d(\mathbf{x}_k, \mathbf{a}) < \delta < \varepsilon$ . Since  $W_C$  and  $W_D$  are disjoint,  $\mathbf{x}_k \notin W_D$ , so  $\mathbf{a} \notin D$ , so  $\mathbf{a} \in C$ . By the choice of  $\delta$ , we get  $d(N(\mathbf{x}_k), N(\mathbf{a})) < \varepsilon$  thus  $d(\mathbf{x}_{k+1}, \mathbf{a}) < \varepsilon$ , so  $\mathbf{x}_{k+1} \in W_C$ . Here we used the fact that  $N(\mathbf{a}) = \mathbf{a}$  since  $\mathbf{a} \in A$ ,  $A$  is the accumulation set for the sequence  $\{\mathbf{x}_k\}$  and hence  $A$  is also the fixed point set for the mapping  $N$  where  $\mathbf{x}_{k+1} = N(\mathbf{x}_k)$ .

Thusm we have shown that  $\{\mathbf{x}_k\}$  is eventually in  $W_C$  and thus is not infinitely often in  $W_D$ , contradicting the fact that  $W_D$  is an open set containing elements of  $A$ . As this contradiction arose from the assumption that  $A$  is disconnected, we conclude that  $A$  is connected. ■

## APPENDIX H

### CONVERGENCE RATE TO A LOCAL MINIMUM (THEOREM 6 AND 7)

#### Proof of Theorem 6

For any  $\mathbf{x}, \mathbf{y}$  in  $\bar{\mathcal{B}}_{R_0}(\mathbf{x}_0^*)$  using (31) we have the following condition:

$$f(\mathbf{x}) - f(\mathbf{y}) \leq L \text{diam}(\mathcal{U}) \|\mathbf{x} - \mathbf{y}\| \leq 2L \text{diam}(\mathcal{U}) R_0. \quad (298)$$

Next, let the trajectory re-enter the ball  $\bar{\mathcal{B}}_{R_0}(\mathbf{x}_0^*)$  after  $J$  iterations and the current iteration index be  $K$  where we have that  $\mathbf{x}_K, \mathbf{x}_{K+J}$  belong to  $\bar{\mathcal{B}}_{R_0}(\mathbf{x}_0^*)$  whereas  $\mathbf{x}_{K+J-1} \notin \bar{\mathcal{B}}_{R_0}(\mathbf{x}_0^*)$ . Using gradient Lipschitz continuity on  $\mathbf{x}_k$  and  $\mathbf{x}_{k+1}$  we get:

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \quad (299)$$

$$\sum_{k=K}^{K+J-1} \left( \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}_{k+1} \rangle - \frac{L}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \right) \leq \sum_{k=K}^{K+J-1} \left( f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \right) \quad (300)$$

$$\sum_{k=K}^{K+J-1} \left( \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}_{k+1} \rangle - \frac{L}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \right) \leq f(\mathbf{x}_K) - f(\mathbf{x}_{K+J}) \leq 2L \text{diam}(\mathcal{U}) R_0 \quad (301)$$

where in the last step we used (298). Now from Algorithm 1 let  $\{k_l\}$  be the subsequence of  $\mathcal{J}$  where  $\mathcal{J} = \{K, \dots, K+J-1\}$  for which we have the update  $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)$  and  $\mathcal{J} \setminus \{k_l\}$  be the subsequence for which we have  $\mathbf{x}_{k+1} - \mathbf{x}_k = \frac{\|\nabla f(\mathbf{x}_k)\|}{\beta} \mathbf{e}_j^{us}$  (this update is a particular case of the Step 17 from Algorithm 1<sup>18</sup>). Further let  $\{k_{l_j}\}$  be the subsequence of  $\{k_l\}$  where  $\|\nabla f(\mathbf{x}_k)\| > \gamma$  and let  $r_k = \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}_{k+1} \rangle - \frac{L}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2$ . Now the left-hand side of (301) can be written as:

$$\sum_{k \in \mathcal{J}} r_k = \sum_{k \in \{k_l\}} r_k + \sum_{k \in \{k_l\} \setminus \{k_{l_j}\}} r_k + \sum_{k \in \mathcal{J} \setminus \{k_l\}} r_k \quad (302)$$

$$\begin{aligned} \sum_{k \in \mathcal{J}} r_k &= \sum_{k \in \{k_l\}} \left( \frac{1}{\alpha} \langle \mathbf{x}_k - \mathbf{x}_{k+1}, \mathbf{x}_k - \mathbf{x}_{k+1} \rangle - \frac{L}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \right) \\ &\quad + \sum_{k \in \{k_l\} \setminus \{k_{l_j}\}} \frac{1}{2L} \|\nabla f(\mathbf{x}_k)\|^2 \\ &\quad + \sum_{k \in \mathcal{J} \setminus \{k_l\}} \left( \langle \nabla f(\mathbf{x}_k), \frac{\|\nabla f(\mathbf{x}_k)\|}{\beta} \mathbf{e}_j^{us} \rangle - \frac{L}{2} \left\| \frac{\|\nabla f(\mathbf{x}_k)\|}{\beta} \mathbf{e}_j^{us} \right\|^2 \right) \end{aligned} \quad (303)$$

$$\begin{aligned} \sum_{k \in \mathcal{J}} r_k &= \sum_{k \in \{k_l\}} \frac{1}{2} \|\nabla f(\mathbf{x}_k)\| \|\mathbf{x}_{k+1} - \mathbf{x}_k\| \\ &\quad + \sum_{k \in \{k_l\} \setminus \{k_{l_j}\}} \frac{1}{2L} \|\nabla f(\mathbf{x}_k)\|^2 \\ &\quad + \sum_{k \in \mathcal{J} \setminus \{k_l\}} \left( \langle \nabla f(\mathbf{x}_k), \frac{\|\nabla f(\mathbf{x}_k)\|}{\beta} \mathbf{e}_j^{us} \rangle - \frac{L}{2\beta^2} \|\nabla f(\mathbf{x}_k)\|^2 \right) \end{aligned} \quad (304)$$

$$\begin{aligned} \sum_{k \in \mathcal{J}} r_k &> \frac{\gamma}{2} \sum_{k \in \{k_{l_j}\}} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| + \sum_{k \in \{k_l\} \setminus \{k_{l_j}\}} \frac{1}{2L} \|\nabla f(\mathbf{x}_k)\|^2 \\ &\quad - \sum_{k \in \mathcal{J} \setminus \{k_l\}} \left( \frac{1}{\beta} + \frac{L}{2\beta^2} \right) \|\nabla f(\mathbf{x}_k)\|^2. \end{aligned} \quad (305)$$

Substituting (305) into (301) yields:

$$\begin{aligned} \frac{\gamma}{2} \sum_{k \in \{k_{l_j}\}} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| &+ \sum_{k \in \{k_l\} \setminus \{k_{l_j}\}} \frac{1}{2L} \|\nabla f(\mathbf{x}_k)\|^2 \\ &- \sum_{k \in \mathcal{J} \setminus \{k_l\}} \left( \frac{1}{\beta} + \frac{L}{2\beta^2} \right) \|\nabla f(\mathbf{x}_k)\|^2 \leq 2L \text{diam}(\mathcal{U}) R_0 \end{aligned} \quad (306)$$

$$\begin{aligned} \frac{\gamma}{2} \sum_{k \in \{k_{l_j}\}} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| &- \sum_{k \in \mathcal{J} \setminus \{k_l\}} \left( \frac{1}{\beta} + \frac{L}{2\beta^2} \right) \|\nabla f(\mathbf{x}_k)\|^2 \leq \\ &2L \text{diam}(\mathcal{U}) R_0 \end{aligned} \quad (307)$$

$$\begin{aligned} \frac{\gamma}{2} \sum_{k \in \{k_{l_j}\}} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| &- \sum_{k \in \mathcal{J} \setminus \{k_l\}} \left( \frac{1}{\beta} + \frac{L}{2\beta^2} \right) L^2 \varepsilon^2 \leq \\ &2L \text{diam}(\mathcal{U}) R_0 \end{aligned} \quad (308)$$

<sup>18</sup>The more general update Step 17 from Algorithm 1 will also yield the same bound after taking norm but is not used here in the interest of simplifying analysis

where in the last step we used the fact that  $\|\nabla f(\mathbf{x}_k)\| \leq L\varepsilon$  for  $k \in \mathcal{J} \setminus \{k_l\}$ . Also note that for all  $k \in \mathcal{J} \setminus \{k_l\}$  we will have  $\mathbf{x}_k \in \bigcup_{\mathbf{x}_i^* \in \mathcal{S}_*} \mathcal{B}_\varepsilon(\mathbf{x}_i^*)$ . Similarly for all  $k \in \mathcal{J} \setminus \{k_{l_j}\}$  we will have  $\mathbf{x}_k, \mathbf{x}_{k+1}$  in the region  $\bigcup_{\mathbf{x}_i^* \in \mathcal{S}_*} \mathcal{B}_\varepsilon(\mathbf{x}_i^*)$  along with  $\mathcal{B}_\varepsilon(\mathbf{x}_i^*) \cap \mathcal{B}_\varepsilon(\mathbf{x}_s^*) = \emptyset$  for any  $\mathbf{x}_i^*, \mathbf{x}_s^*$  in  $\mathcal{S}_*$ . Now adding  $\frac{\gamma}{2} \sum_{k \in \mathcal{J} \setminus \{k_{l_j}\}} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|$  to both sides of (308) we get:

$$\begin{aligned} \frac{\gamma}{2} \sum_{k \in \mathcal{J} \setminus \{k_{l_j}\}} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| &+ \frac{\gamma}{2} \sum_{k \in \{k_{l_j}\}} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq 2L \text{diam}(\mathcal{U}) R_0 \\ &+ \sum_{k \in \mathcal{J} \setminus \{k_l\}} \left( \frac{1}{\beta} + \frac{L}{2\beta^2} \right) L^2 \varepsilon^2 + \frac{\gamma}{2} \sum_{k \in \mathcal{J} \setminus \{k_{l_j}\}} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| \quad (309) \\ \frac{\gamma}{2} \sum_{k \in \mathcal{J}} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| &\leq 2L \text{diam}(\mathcal{U}) R_0 + \sum_{k \in \mathcal{J} \setminus \{k_l\}} \left( \frac{1}{\beta} + \frac{L}{2\beta^2} \right) L^2 \varepsilon^2 \\ &+ \frac{\gamma}{2} \sum_{k \in \mathcal{J} \setminus \{k_{l_j}\}} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| \quad (310) \end{aligned}$$

$$\begin{aligned} \frac{\gamma}{2} \sum_{k \in \mathcal{J}} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| &\leq 2L \text{diam}(\mathcal{U}) R_0 + \sum_{k \in \mathcal{J} \setminus \{k_l\}} \left( \frac{1}{\beta} + \frac{L}{2\beta^2} \right) L^2 \varepsilon^2 \\ &+ \gamma \sum_{k \in \mathcal{J} \setminus \{k_{l_j}\}} \xi \quad (311) \end{aligned}$$

where in the last step we used the fact that  $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq 2\xi$  since  $\mathbf{x}_k, \mathbf{x}_{k+1}$  lie inside some ball  $\mathcal{B}_\varepsilon(\mathbf{x}_i^*)$  for  $k \in \mathcal{J} \setminus \{k_{l_j}\}$ . If the trajectory  $\{\mathbf{x}_k\}$  encounters  $N$  such  $\mathcal{B}_\varepsilon(\mathbf{x}_i^*)$  balls then (311) can be further simplified as:

$$\begin{aligned} \frac{\gamma}{2} \sum_{k \in \mathcal{J}} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| &\leq 2L \text{diam}(\mathcal{U}) R_0 \\ &+ N \left( \frac{1}{\beta} + \frac{L}{2\beta^2} \right) L^2 \varepsilon^2 + \gamma N (K_{\text{exit}} + K_{\text{shell}}) \xi \end{aligned} \quad (312)$$

where exit time from  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  ball is  $K_{\text{exit}}$  from Theorem 3.2 of [10], exit time from  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  ball is  $K_{\text{exit}} + K_{\text{shell}}$  after adding results from Theorem 3 and Theorem 3.2 of [10], and we have that  $\sum_{k \in \mathcal{J} \setminus \{k_l\}} \leq N$ ,  $\sum_{k \in \mathcal{J} \setminus \{k_{l_j}\}} \leq N(K_{\text{exit}} + K_{\text{shell}})$ .

Next we set  $K = 0$  in the subsequence  $\mathcal{J}$  where  $\mathcal{J} = \{K, \dots, K+J-1\}$  so that  $\mathbf{x}_0 \in \mathcal{B}_\varepsilon(\mathbf{x}_0^*)$ . In that case  $\sum_{k \in \mathcal{J}} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|$  is the total path length of the trajectory inside the ball  $\mathcal{B}_{R_\omega}(\mathbf{x}_0^*)$  where we have that  $R_\omega = \max_{k \in \mathcal{J}} \|\mathbf{x}_k - \mathbf{x}_0^*\|$  and  $\|\mathbf{x}_J - \mathbf{x}_0^*\| \leq R_0$ . Hence, for some  $K_\omega = \arg \max_{k \in \mathcal{J}} \|\mathbf{x}_k - \mathbf{x}_0^*\|$  we will have the condition:

$$\sum_{k \in \mathcal{J}} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| = \sum_{k=0}^{K_\omega-1} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| + \sum_{k=K_\omega}^J \|\mathbf{x}_{k+1} - \mathbf{x}_k\| \quad (313)$$

$$\geq \left\| \sum_{k=0}^{K_\omega-1} (\mathbf{x}_{k+1} - \mathbf{x}_k) \right\| + \left\| \sum_{k=K_\omega}^J (\mathbf{x}_{k+1} - \mathbf{x}_k) \right\| \quad (314)$$

$$\geq \|\mathbf{x}_{K_\omega} - \mathbf{x}_0\| + \|\mathbf{x}_J - \mathbf{x}_{K_\omega}\| \quad (315)$$

$$\geq \underbrace{\|\mathbf{x}_{K_\omega} - \mathbf{x}_0^*\|}_{=R_\omega} - \|\mathbf{x}_0 - \mathbf{x}_0^*\| + \|\mathbf{x}_{K_\omega} - \mathbf{x}_0^*\| - \|\mathbf{x}_J - \mathbf{x}_0^*\| \quad (316)$$

$$\geq 2R_\omega - R_0 - \xi. \quad (317)$$

Substituting (317) into (312) yields:

$$\begin{aligned} \frac{\gamma}{2}(2R_\omega - R_0 - \xi) &\leq \frac{\gamma}{2} \sum_{k \in \mathcal{J}} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq 2L \text{diam}(\mathcal{U}) R_0 \\ &\quad + N \left( \frac{1}{\beta} + \frac{L}{2\beta^2} \right) L^2 \varepsilon^2 + \gamma N (K_{\text{exit}} + K_{\text{shell}}) \xi. \end{aligned} \quad (318)$$

Next, recall that the distance between any two stationary points is greater than  $R$ . Hence, between two points  $\mathbf{x}, \mathbf{y}$  with  $\|\mathbf{x} - \mathbf{y}\| \leq D$ , there can be at most  $\frac{D}{R}$  stationary points along the straight line joining  $\mathbf{x}, \mathbf{y}$ . Now if the points  $\mathbf{x}, \mathbf{y}$  are connected by a path formed from the sequence of points  $\{\mathbf{v}_k\}_{k=1}^P$  then

there can be at most  $\frac{\sum_{k=1}^{P-1} \|\mathbf{v}_{k+1} - \mathbf{v}_k\|}{R}$  stationary points on the path connecting  $\mathbf{x}, \mathbf{y}$ . Using this result in (318) yields the following bound on  $N$ :

$$\begin{aligned} \frac{\gamma}{2} N &\leq \frac{\gamma \sum_{k \in \mathcal{J}} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|}{2R} \\ &\leq 2L \text{diam}(\mathcal{U}) \frac{R_0}{R} + N \left( \frac{1}{\beta} + \frac{L}{2\beta^2} \right) \frac{L^2 \varepsilon^2}{R} + \gamma N (K_{\text{exit}} + K_{\text{shell}}) \frac{\xi}{R} \end{aligned} \quad (319)$$

$$\begin{aligned} N \left( \frac{\gamma}{2} - \left( \frac{1}{\beta} + \frac{L}{2\beta^2} \right) \frac{L^2 \varepsilon^2}{R} - \gamma (K_{\text{exit}} + K_{\text{shell}}) \frac{\xi}{R} \right) &\leq 2L \text{diam}(\mathcal{U}) \frac{R_0}{R} \end{aligned} \quad (320)$$

$$N \leq \frac{2L \text{diam}(\mathcal{U}) \frac{R_0}{R}}{\left( \frac{\gamma}{2} - \left( \frac{1}{\beta} + \frac{L}{2\beta^2} \right) \frac{L^2 \varepsilon^2}{R} - \gamma (K_{\text{exit}} + K_{\text{shell}}) \frac{\xi}{R} \right)} \quad (321)$$

provided  $\left( \frac{\gamma}{2} - \left( \frac{1}{\beta} + \frac{L}{2\beta^2} \right) \frac{L^2 \varepsilon^2}{R} - \gamma (K_{\text{exit}} + K_{\text{shell}}) \frac{\xi}{R} \right) > 0$  which will hold true for  $\xi \ll R$ .

Finally, combining (318) and (321) yields the result:

$$\begin{aligned} R_\omega &\leq R_0 + 2L \text{diam}(\mathcal{U}) \frac{R_0}{\gamma} + N_0 K_{\text{exit}} \left( \frac{1}{\beta} + \frac{L}{2\beta^2} \right) \frac{L^2 \varepsilon^2}{\gamma} \\ &\quad + N_0 (K_{\text{exit}} + K_{\text{shell}}) \xi + \frac{1}{2} (R_0 + \xi) \end{aligned} \quad (322)$$

where  $N_0 = \frac{2L \text{diam}(\mathcal{U}) \frac{R_0}{R}}{\left( \frac{\gamma}{2} - \left( \frac{1}{\beta} + \frac{L}{2\beta^2} \right) \frac{L^2 \varepsilon^2}{R} - \gamma (K_{\text{exit}} + K_{\text{shell}}) \frac{\xi}{R} \right)}$  is the upper bound on the number of stationary point neighborhoods encountered by the trajectory of  $\{\mathbf{x}_k\}$ . ■

### Proof of Theorem 7

To obtain the total number of iterations in which the sequence  $\{\mathbf{x}_k\}$  converges to some  $\varepsilon$  neighborhood of a local minimum which is within a  $\xi$  neighborhood of  $\mathbf{x}_0$ , we first obtain the number of iterations the sequence  $\{\mathbf{x}_k\}$  spends in the region  $\mathcal{U} \setminus \bigcup_{j=1}^J \mathcal{B}_\xi(\mathbf{x}_j^*)$ , i.e., the region with  $\|\nabla f(\mathbf{x})\| > \gamma$ . Let  $K_1$  be the number of such iterations and  $T$  be the number of saddle neighborhoods encountered by the trajectory of  $\{\mathbf{x}_k\}$ .

In order to obtain  $K_1$  we make use of (308) by setting  $R_0 = \xi$  and  $K = 0$  in the subsequence  $\mathcal{J}$  where  $\mathcal{J} = \{K, \dots, K+J-1\}$  so that  $\mathbf{x}_0 \in \mathcal{B}_\xi(\mathbf{x}_0^*)$  to get:

$$\begin{aligned} \frac{\gamma}{2} \sum_{k \in \{k_{l_j}\}} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| - \sum_{k \in \mathcal{J} \setminus \{k_l\}} \left( \frac{1}{\beta} + \frac{L}{2\beta^2} \right) L^2 \varepsilon^2 &\leq \\ 2L \text{diam}(\mathcal{U}) \xi &\quad (323) \end{aligned}$$

$$\begin{aligned} \Rightarrow \frac{\gamma}{2} \sum_{k \in \{k_{l_j}\}} \|\alpha \nabla f(\mathbf{x}_k)\| - T \left( \frac{1}{\beta} + \frac{L}{2\beta^2} \right) L^2 \varepsilon^2 &\leq \\ 2L \text{diam}(\mathcal{U}) \xi &\quad (324) \end{aligned}$$

$$\Rightarrow K_1 \leq 4L \text{diam}(\mathcal{U}) \frac{\xi L}{\gamma^2} + 2T \left( \frac{1}{\beta} + \frac{L}{2\beta^2} \right) \frac{\varepsilon^2}{\gamma^2} \quad (325)$$

where we used the fact that  $\sum_{k \in \{k_{l_j}\}} \|\nabla f(\mathbf{x}_k)\| > \gamma K_1$  by definition of the subsequence  $\{k_{l_j}\}$  in (308) and  $\sum_{k \in \mathcal{J} \setminus \{k_l\}} = T < N_0 = \frac{2L \text{diam}(\mathcal{U}) \frac{R_0}{R}}{\left( \frac{\gamma}{2} - \left( \frac{1}{\beta} + \frac{L}{2\beta^2} \right) \frac{L^2 \varepsilon^2}{R} - \gamma (K_{\text{exit}} + K_{\text{shell}}) \frac{\xi}{R} \right)}$  by Theorem 6

for  $R_0 = \xi$  where  $T$  is the number of saddle neighborhoods encountered by the trajectory of  $\{\mathbf{x}_k\}$ . Since we have a bound on the number of saddle neighborhoods  $T$  and we also know the travel time within each saddle neighborhood we are only left to find the rate within the neighborhood of a local minimum.

### Local minimum neighborhood

When the trajectory  $\{\mathbf{x}_k\}$  is within a  $\xi$  neighborhood of local minimum  $\mathbf{x}_{\text{optimal}}^*$  for some  $k = K$ , we have linear rate of convergence to the neighborhood  $\mathcal{B}_\varepsilon(\mathbf{x}_{\text{optimal}}^*)$  from the following steps:

$$\begin{aligned} \mathbf{x}_{k+1} - \mathbf{x}_{\text{optimal}}^* &= \\ \left( \mathbf{I} - \alpha \int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}_{\text{optimal}}^* + p(\mathbf{x}_k - \mathbf{x}_{\text{optimal}}^*)) dp \right) (\mathbf{x}_k - \mathbf{x}_{\text{optimal}}^*) &\quad (326) \end{aligned}$$

$$\begin{aligned} \Rightarrow \|\mathbf{x}_{k+1} - \mathbf{x}_{\text{optimal}}^*\| &\leq \\ \left\| \mathbf{I} - \alpha \underbrace{\left( \int_{p=0}^{p=1} \nabla^2 f(\mathbf{x}_{\text{optimal}}^* + p(\mathbf{x}_k - \mathbf{x}_{\text{optimal}}^*)) dp \right)}_{=1 - \frac{\beta}{L}} \right\|_2 \|\mathbf{x}_k - \mathbf{x}_{\text{optimal}}^*\| &\quad (327) \end{aligned}$$

$$\Rightarrow \|\mathbf{x}_{K+K_{\text{convex}}} - \mathbf{x}_{\text{optimal}}^*\| \leq \left( 1 - \frac{\beta}{L} \right)^{K_{\text{convex}}} \|\mathbf{x}_K - \mathbf{x}_{\text{optimal}}^*\| \quad (328)$$

$$\Rightarrow K_{\text{convex}} \leq \frac{\log(\|\mathbf{x}_K - \mathbf{x}_{\text{optimal}}^*\|) - \log(\|\mathbf{x}_{K+K_{\text{convex}}} - \mathbf{x}_{\text{optimal}}^*\|)}{\log\left(1 - \frac{\beta}{L}\right)^{-1}}$$

$$\begin{aligned} &\leq \frac{\log\left(\frac{\xi}{\varepsilon}\right)}{\log\left(1 - \frac{\beta}{L}\right)^{-1}} \end{aligned} \quad (329)$$

where  $\mathbf{x}_K \in \mathcal{B}_\xi(\mathbf{x}_{\text{optimal}}^*)$  and  $\|\mathbf{x}_{K+K_{\text{convex}}} - \mathbf{x}_{\text{optimal}}^*\| = \varepsilon$ . Note that in the second step we used the facts that  $\alpha =$

$\frac{1}{L}$ ,  $\lambda_{\min}(f(\cdot)) \geq \int \lambda_{\min}(\cdot)$  and  $\lambda_{\min}(\nabla^2 f(\mathbf{x}_{\text{optimal}}^* + p(\mathbf{x}_k - \mathbf{x}_{\text{optimal}}^*))) = \beta$  for any  $\mathbf{x}_{\text{optimal}}^* + p(\mathbf{x}_k - \mathbf{x}_{\text{optimal}}^*)$  in the convex neighborhood  $\mathcal{B}_\xi(\mathbf{x}_{\text{optimal}}^*)$  from Assumption A4.

Finally putting everything together and using Theorem 3.2 from [10], Theorem 3 travel time from (325) and the convergence rate within a convex neighborhood from (329), the total time for the trajectory of  $\{\mathbf{x}_k\}$  to converge to an  $\varepsilon$  neighborhood of  $\mathbf{x}_{\text{optimal}}^*$  is bounded by:

$$K_{\max} \leq T(K_{\text{exit}} + K_{\text{shell}}) + K_1 + K_{\text{convex}} \quad (330)$$

$$\begin{aligned} &< T(K_{\text{exit}} + K_{\text{shell}}) + 4L \text{diam}(\mathcal{U}) \frac{\xi L}{\gamma^2} \\ &+ 2T\left(\frac{1}{\beta} + \frac{L}{2\beta^2}\right) \frac{\varepsilon^2}{\gamma^2} + \frac{\log\left(\frac{\xi}{\varepsilon}\right)}{\log\left(1 - \frac{\beta}{L}\right)^{-1}} \end{aligned} \quad (331)$$

where  $T < \frac{2L \text{diam}(\mathcal{U}) \frac{\xi}{R}}{\left(\frac{\gamma}{2} - \left(\frac{1}{\beta} + \frac{L}{2\beta^2}\right) \frac{L^2 \varepsilon^2}{R} - \gamma(K_{\text{exit}} + K_{\text{shell}}) \frac{\xi}{R}\right)}$  is the total number of saddle neighborhoods encountered.

We complete the proof of Theorem 7 by proving one last claim. Recall that  $K_{\text{exit}}$  was the exit time of the  $\varepsilon$ -precision trajectory from the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$  while we proved Theorem 7 for the exact gradient trajectory. Hence, we need to justify the use of the upper bound on  $K_{\text{exit}}$  from (7) in Theorem 7.

Let  $K_{\text{exit}}^o$  be the actual exit time of the gradient trajectory  $\{\mathbf{u}_K\}$  from the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$ , i.e.,  $K_{\text{exit}}^o = \inf_{K>0} \left\{ K \mid \|\mathbf{u}_K\| \geq \varepsilon \right\}$  where  $\mathbf{u}_K = \mathbf{x}_K - \mathbf{x}^*$  is the radial vector and  $\|\mathbf{u}_0\| = \varepsilon$ . Since  $K_{\text{exit}}$  is the exit time of the  $\varepsilon$ -precision trajectory  $\{\tilde{\mathbf{u}}_K\}$  from the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$ , i.e.,  $K_{\text{exit}} = \inf_{K>0} \left\{ K \mid \|\tilde{\mathbf{u}}_K\| \geq \varepsilon \right\}$ , by the definition of exit time we have that  $\|\tilde{\mathbf{u}}_{K_{\text{exit}}}\| \geq \varepsilon$ .

Now if the initial unstable subspace projection value  $\sum_{j \in \mathcal{N}_{US}} (\theta_j^{us})^2$  satisfies the condition of Theorem 1 then from the relative error bound (21) we have that:

$$\frac{\|\mathbf{u}_K - \tilde{\mathbf{u}}_K\|}{\|\mathbf{u}_K\|} \leq \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)^2\right) \quad (332)$$

$$\begin{aligned} \Rightarrow 1 - \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)^2\right) &\leq \frac{\|\tilde{\mathbf{u}}_K\|}{\|\mathbf{u}_K\|} \leq \\ &1 + \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)^2\right) \end{aligned} \quad (333)$$

$$\begin{aligned} \Rightarrow \frac{\|\tilde{\mathbf{u}}_K\|}{1 + \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)^2\right)} &\leq \|\mathbf{u}_K\| \leq \\ &\frac{\|\tilde{\mathbf{u}}_K\|}{1 - \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)^2\right)} \end{aligned} \quad (334)$$

$$\begin{aligned} \Rightarrow \frac{\varepsilon}{1 + \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)^2\right)} &\leq \|\mathbf{u}_{K_{\text{exit}}}\| \leq \\ &\frac{(1+d)\varepsilon}{1 - \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)^2\right)} \end{aligned} \quad (335)$$

where we substituted  $K = K_{\text{exit}}$  and used the bound  $(1+d)\varepsilon \geq \|\tilde{\mathbf{u}}_{K_{\text{exit}}}\| \geq \varepsilon$  for some  $d > 0$  in the last step. Next, from the definition of  $K_{\text{exit}}^o$  we have that  $\|\mathbf{u}_{K_{\text{exit}}^o}\| \geq \varepsilon$ . Hence, unless we have  $\|\mathbf{u}_{K_{\text{exit}}}\| \geq \varepsilon$  (which implies  $K_{\text{exit}}^o \leq K_{\text{exit}}$ ), the gradient trajectory  $\{\mathbf{u}_K\}$  will take not more than  $K_{\text{exit}}^o - K_{\text{exit}}$  iterations to travel the shell  $\mathcal{B}_\varepsilon(\mathbf{x}^*) \setminus \mathcal{B}_{\|\mathbf{u}_{K_{\text{exit}}}\|}(\mathbf{x}^*)$ . Next,  $K_{\text{exit}}^o - K_{\text{exit}}$  can be upper bounded by Theorem 3 provided the gradient trajectory has expansive dynamics at  $K_{\text{exit}}$  (from Theorem 2).

Now for sufficiently small  $\varepsilon$  and  $K_{\text{exit}} \geq 2$  (the minimal condition that ensures the gradient trajectory at-least enters the ball  $\mathcal{B}_\varepsilon(\mathbf{x}^*)$ ), there exists some  $K = K^v$  with  $K^v < K_{\text{exit}}$  such that:

$$\frac{\|\tilde{\mathbf{u}}_{K^v}\|}{1 - \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)^2\right)} \leq \frac{\|\tilde{\mathbf{u}}_{K_{\text{exit}}}\|}{1 + \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)^2\right)}. \quad (336)$$

Combining (336) with (334) for  $K = K_{\text{exit}}$  and  $K = K^v$  we get:

$$\begin{aligned} \|\mathbf{u}_{K^v}\| &\leq \frac{\|\tilde{\mathbf{u}}_{K^v}\|}{1 - \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)^2\right)} \leq \frac{\|\tilde{\mathbf{u}}_{K_{\text{exit}}}\|}{1 + \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)^2\right)} \\ &\leq \|\mathbf{u}_{K_{\text{exit}}}\| \end{aligned} \quad (337)$$

$$\Rightarrow \|\mathbf{u}_{K^v}\| \leq \|\mathbf{u}_{K_{\text{exit}}}\|. \quad (338)$$

which implies that the gradient trajectory has expansive dynamics at  $K = K_{\text{exit}}$  from Theorem 2. Hence, the gradient trajectory will also have expansive dynamics from  $K = K_{\text{exit}}$  to  $K = K_{\text{exit}}^o$ . Using Theorem 3 for  $\xi = \|\mathbf{u}_{K_{\text{exit}}^o - 1}\|$ ,  $\varepsilon = \|\mathbf{u}_{K_{\text{exit}}}\|$ ,  $\hat{K}_{\text{exit}} = K_{\text{exit}}^o - 1$  and  $K_e = K_{\text{exit}}$  we get:

$$K_{\text{exit}}^o - 1 - K_{\text{exit}} = \hat{K}_{\text{exit}} - K_e \leq \frac{\log(\|\mathbf{u}_{K_{\text{exit}}^o - 1}\|) - \log(\|\mathbf{u}_{K_{\text{exit}}}\|)}{\log\left(\frac{\inf\{\tilde{\rho}(\mathbf{x}_{k-2})\}}{1+M\xi}\right)} + 2 \quad (339)$$

$$\begin{aligned} &< \frac{\log(1 + \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)^2\right))}{\log\left(\frac{\inf\{\tilde{\rho}(\mathbf{x}_{k-2})\}}{1+M\xi}\right)} + 2 \\ &\lesssim 2 \end{aligned} \quad (340)$$

where we used the bound  $\|\mathbf{u}_{K_{\text{exit}}^o - 1}\| < \varepsilon$  from the definition of  $K_{\text{exit}}^o$ , the lower bound on  $\|\mathbf{u}_{K_{\text{exit}}}\|$  from (335) in the second last step and dropped the term  $\log(1 + \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} \left(\log\left(\frac{1}{\varepsilon}\right) \varepsilon\right)^2\right))$  for sufficiently small  $\varepsilon$ . Hence we have the condition  $K_{\text{exit}}^o \lesssim K_{\text{exit}} + 3$  where the constant 3 can be dropped w.r.t. order  $\mathcal{O}(\log(\varepsilon^{-1}))$  term after substituting the upper bound on  $K_{\text{exit}}$  from (7). This completes the proof. ■

### Proof of Theorem 9

Since  $f(\cdot)$  satisfies Assumption A1, it is locally gradient Lipschitz continuous in every compact set. Hence  $f(\cdot)$  is gradient Lipschitz continuous in some compact set  $\mathcal{K} \supset \{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$  where the compact set  $\mathcal{K}$  will be specified later. Suppose  $L$  is the gradient Lipschitz constant of  $f(\cdot)$  in this compact set or equivalently  $L$  is the local Lipschitz constant of the function  $\nabla f: \mathbb{R}^n \rightarrow \mathbb{R}^n$  when restricted to  $\mathcal{K}$ . Next, by the Kirszbraun Theorem (Theorem 8), there exists an extension  $G: \mathbb{R}^n \rightarrow \mathbb{R}^n$  of the function  $g \equiv \nabla f$  on the entire Euclidean space  $\mathbb{R}^n$  such that  $G \equiv \nabla f$  on the compact set  $\mathcal{K} \supset \{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$  and  $G$  is globally Lipschitz continuous with a Lipschitz constant of  $L$ . Next suppose  $F: \mathbb{R}^n \rightarrow \mathbb{R}$  is the primitive of  $G$  given by the line integral  $F(\mathbf{x}) = \int_{\gamma_0} \langle G(\mathbf{v}), d\mathbf{v} \rangle$  where  $\gamma_0$  is any smooth curve from  $\mathbf{0}$  to  $\mathbf{x}$ . Then taking directional derivative of  $F$  with respect to  $\mathbf{v}$  we get  $\frac{\partial F(\mathbf{x})}{\partial \mathbf{v}} = \langle G(\mathbf{x}), \mathbf{v} \rangle$  and so  $G(\mathbf{x}) = \nabla F(\mathbf{x})$  for all  $\mathbf{x} \in \mathbb{R}^n$ . Now  $\nabla F \equiv G \equiv \nabla f$  on the compact set  $\mathcal{K}$  so  $F \equiv f + c$  on this set for any constant  $c$ . Without loss of generality we can take  $c = 0$  so that  $F \equiv f$  on the set  $\mathcal{K} \supset \{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$ . Since  $G$  is  $L$  Lipschitz continuous and so is  $\nabla F$ , we can write:

$$F(\mathbf{x}) - F(\mathbf{y}) \geq \langle \nabla F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle - \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 \quad (341)$$

$$\implies F(\mathbf{x}) + \frac{L+\varepsilon}{2} \|\mathbf{x} - \mathbf{y}\|^2 \geq \langle \nabla F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\varepsilon}{2} \|\mathbf{x} - \mathbf{y}\|^2. \quad (342)$$

Next recall that since  $f$  is coercive and continuous it has a global minimum [55] which will belong to the compact sublevel set  $\{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\} \subset \mathcal{K}$ . Since  $f(\cdot) \in \mathcal{C}^2$ , this global minimum, say  $\mathbf{x}^*$ , will be a critical point of  $f$  and therefore a critical point of  $F$  because  $F \equiv f$  on the set  $\{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\} \subset \mathcal{K}$ . Then setting  $\mathbf{y} = \mathbf{x}^*$  in (342) we get the following for  $\varepsilon > 0$ :

$$F(\mathbf{x}) + \frac{L+\varepsilon}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 \geq \langle \nabla F(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle + \frac{\varepsilon}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 \\ \implies F(\mathbf{x}) + \frac{L+\varepsilon}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 \geq \frac{\varepsilon}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 \geq \frac{\varepsilon}{4} \|\mathbf{x}\|^2 - \frac{\varepsilon}{2} \|\mathbf{x}^*\|^2$$

where in the last step we used the inequality  $\|\mathbf{x} - \mathbf{x}^*\|^2 \geq \frac{1}{2} \|\mathbf{x}\|^2 - \|\mathbf{x}^*\|^2$ . Since  $\mathbf{x}^*$  belongs to a compact set,  $\|\mathbf{x}^*\|$  is bounded and so the function  $F(\mathbf{x}) + \frac{L+\varepsilon}{2} \|\mathbf{x} - \mathbf{x}^*\|^2$  is coercive by the following argument:

$$\limsup_{\|\mathbf{x}\| \rightarrow \infty} \left( F(\mathbf{x}) + \frac{L+\varepsilon}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 \right) \geq \\ \liminf_{\|\mathbf{x}\| \rightarrow \infty} \left( F(\mathbf{x}) + \frac{L+\varepsilon}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 \right) \geq \\ \liminf_{\|\mathbf{x}\| \rightarrow \infty} \left( \frac{\varepsilon}{4} \|\mathbf{x}\|^2 - \frac{\varepsilon}{2} \|\mathbf{x}^*\|^2 \right) \quad (343)$$

$$\implies \lim_{\|\mathbf{x}\| \rightarrow \infty} \left( F(\mathbf{x}) + \frac{L+\varepsilon}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 \right) = \infty. \quad (344)$$

Since  $F(\mathbf{x}) + \frac{L+\varepsilon}{2} \|\mathbf{x} - \mathbf{x}^*\|^2$  is coercive,  $\inf_{\mathbf{x} \in \mathbb{R}^n} \left( F(\mathbf{x}) + \frac{L+\varepsilon}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 \right) > -\infty$ .

Next, consider the function  $\tilde{F}: \mathbb{R}^n \rightarrow \mathbb{R}$  given by:

$$\tilde{F}(\mathbf{x}) = F(\mathbf{x}) + \frac{L+\varepsilon}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 \\ + (1 - \Phi_{\mathcal{K}}(\mathbf{x})) \left( f(\mathbf{x}_0) - \inf_{\mathbf{x} \in \mathbb{R}^n} \left( F(\mathbf{x}) + \frac{L+\varepsilon}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 \right) \right) \\ - \frac{L+\varepsilon}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 \Phi_{\mathcal{K}}(\mathbf{x}) \quad (345)$$

where  $\Phi_{\mathcal{K}}$  is a  $\mathcal{C}^\infty$  smooth bump function, where we choose the compact set  $\mathcal{K}$  to satisfy the condition  $\mathcal{K} \supset \{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\} + \mathcal{B}_\varepsilon(\mathbf{0})$  [19] and we have  $0 \leq \Phi_{\mathcal{K}} \leq 1$ ,  $\Phi_{\mathcal{K}} \equiv 1$  on the open set  $\{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\} + \mathcal{B}_\varepsilon(\mathbf{0})$  and  $\Phi_{\mathcal{K}} \equiv 0$  on  $\mathbb{R}^n \setminus \mathcal{K}$ . Such a smooth bump function exists by Proposition 2.25 in [56]. Clearly  $\tilde{F} \equiv f$  on the compact set  $\{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$  and  $\tilde{F}(\mathbf{x}) = F(\mathbf{x}) + \frac{L+\varepsilon}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 + f(\mathbf{x}_0) - \inf_{\mathbf{x} \in \mathbb{R}^n} \left( F(\mathbf{x}) + \frac{L+\varepsilon}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 \right)$  for  $\mathbf{x} \in \mathbb{R}^n \setminus \mathcal{K}$  and so  $\tilde{F}$  is coercive by (344). Moreover we have that  $\tilde{F}(\mathbf{x}) \geq f(\mathbf{x}_0)$  for  $\mathbf{x} \in \mathbb{R}^n \setminus \mathcal{K}$ . Since  $\mathbf{x}^* \in \{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\} \subset \mathcal{K}$  and  $F \equiv f$  on  $\mathcal{K}$  we will have  $\inf_{\mathbf{x} \in \mathbb{R}^n} \left( F(\mathbf{x}) + \frac{L+\varepsilon}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 \right) \leq F(\mathbf{x}^*) = f(\mathbf{x}^*) \leq f(\mathbf{x}_0)$  which implies  $f(\mathbf{x}_0) - \inf_{\mathbf{x} \in \mathbb{R}^n} \left( F(\mathbf{x}) + \frac{L+\varepsilon}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 \right) \geq 0$ . Therefore from (345) and the facts that  $F \equiv f$  on  $\mathcal{K}$ ,  $0 \leq \Phi_{\mathcal{K}} \leq 1$  we also have that  $\tilde{F}(\mathbf{x}) \geq f(\mathbf{x})$  for  $\mathbf{x} \in \mathcal{K} \setminus \{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$ . Now  $f(\mathbf{x}) > f(\mathbf{x}_0)$  on the complement of the set  $\{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$  so  $\tilde{F}(\mathbf{x}) > \tilde{F}(\mathbf{x}_0) = f(\mathbf{x}_0)$  on the set  $\mathbb{R}^n \setminus \{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$  where we used the fact that  $f$  is coercive. Therefore we have  $\{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\} = \{\mathbf{x} \mid \tilde{F}(\mathbf{x}) \leq \tilde{F}(\mathbf{x}_0)\}$ .

Since  $\tilde{F}$  is  $\mathcal{C}^2$  smooth, it will be gradient Lipschitz continuous on the compact set  $\mathcal{K} \supset \{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\} + \mathcal{B}_\varepsilon(\mathbf{0})$  with some gradient Lipschitz constant  $\tilde{L} > L > 0$ , where  $\tilde{L}$  is also the global gradient Lipschitz constant for  $\tilde{F}$ . Next, iterating CCRGD algorithm [1] on the function  $\tilde{F}$  with initialization  $\mathbf{x}_0$  and step size  $\alpha = \frac{1}{\tilde{L}}$  we will have that  $\tilde{F}(\mathbf{x}_{k+1}) \leq \tilde{F}(\mathbf{x}_k) \leq \tilde{F}(\mathbf{x}_0)$  for all  $k \geq 0$  by [20] Lemmas 8 and 10. Therefore  $\mathbf{x}_k \in \{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\} = \{\mathbf{x} \mid \tilde{F}(\mathbf{x}) \leq \tilde{F}(\mathbf{x}_0)\}$  for all  $k$ ,  $\mathbf{x}_k$  will converge to a local minimum in this compact set from Theorem 5 with a convergence rate given by Theorem 7. In particular, since the sequence  $\{\mathbf{x}_k\}$  stays within the compact set  $\{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$  and  $\tilde{F} \equiv f$  on this set, the CCRGD algorithm [1] for  $\alpha = \frac{1}{\tilde{L}}$  will generate the same iterate sequence it would have generated while iterating on the function  $f(\cdot)$ . This completes the proof. ■

<sup>19</sup>Here the operator  $+$  defines the Minkowski sum operation between sets. The openness of the set  $\{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\} + \mathcal{B}_\varepsilon(\mathbf{0})$  follows from the fact that for two sets  $A, B$  their Minkowski sum  $A+B$  is open even if only one of the sets (say,  $B$ ) is open [57].

<sup>20</sup>Since  $\tilde{F} \equiv f$  on  $\mathcal{K}$ ,  $\tilde{F}$  will be Hessian Lipschitz continuous in  $\Omega(\varepsilon)$  neighborhoods of strict saddle points in the set  $\{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\} \subset \{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\} + \mathcal{B}_\varepsilon(\mathbf{0}) \subset \mathcal{K}$  by Assumption A1. Now if for some  $k$ ,  $\mathbf{x}_k \in \mathcal{B}_\varepsilon(\mathbf{x}^*)$  where  $\mathbf{x}^* \in \{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$  is some strict saddle point of  $f$ , the second order step from algorithm [1] will decrease  $\tilde{F}$  by Lemma 8 because  $\mathcal{B}_\varepsilon(\mathbf{x}^*) \subset \mathcal{K}$ .

## REFERENCES

- [1] H. B. Curry, "The method of steepest descent for non-linear minimization problems," *Quarterly of Applied Mathematics*, vol. 2, no. 3, pp. 258–261, 1944.
- [2] M. R. Hestenes *et al.*, "Methods of conjugate gradients for solving linear systems," *Journal of research of the National Bureau of Standards*, vol. 49, no. 6, pp. 409–436, 1952.
- [3] H. Rosenbrock, "An automatic method for finding the greatest or least value of a function," *The Computer Journal*, vol. 3, no. 3, pp. 175–184, 1960.
- [4] N. Karmarkar, "A new polynomial-time algorithm for linear programming," in *Proceedings of the sixteenth annual ACM symposium on Theory of computing*. ACM, 1984, pp. 302–311.
- [5] S. Mehrotra, "On the implementation of a primal-dual interior point method," *SIAM Journal on optimization*, vol. 2, no. 4, pp. 575–601, 1992.
- [6] Y. Nesterov and A. Nemirovskii, *Interior-point polynomial algorithms in convex programming*. SIAM, 1994, vol. 13.
- [7] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.
- [8] T. D. Sanger, "Optimal unsupervised learning in a single-layer linear feedforward neural network," *Neural networks*, vol. 2, no. 6, pp. 459–473, 1989.
- [9] C. G. Broyden, "The convergence of a class of double-rank minimization algorithms 1. general considerations," *IMA Journal of Applied Mathematics*, vol. 6, no. 1, pp. 76–90, 1970.
- [10] R. Dixit, M. Gurbuzbalaban, and W. U. Bajwa, "Exit time analysis for approximations of gradient descent trajectories around saddle points," *arXiv preprint arXiv:2006.01106*, 2022.
- [11] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht, "First-order methods almost always avoid strict saddle points," *Mathematical programming*, vol. 176, no. 1, pp. 311–337, 2019.
- [12] A. Kelley, "The stable, center-stable, center, center-unstable, unstable manifolds," *Journal of Differential Equations*, 1966.
- [13] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.
- [14] Y. Nesterov and B. T. Polyak, "Cubic regularization of newton method and its global performance," *Mathematical Programming*, vol. 108, no. 1, pp. 177–205, 2006.
- [15] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the polyak-Łojasiewicz condition," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2016, pp. 795–811.
- [16] S. Łojasiewicz, "Sur le problème de la division," *Studia Mathematica*, vol. 18, pp. 87–136, 1959.
- [17] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized gauss–seidel methods," *Mathematical Programming*, vol. 137, no. 1–2, pp. 91–129, 2013.
- [18] J. Bolte, A. Daniilidis, and A. Lewis, "The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems," *SIAM Journal on Optimization*, vol. 17, no. 4, pp. 1205–1223, 2007.
- [19] Y. Kifer, "The exit problem for small random perturbations of dynamical systems with a hyperbolic fixed point," *Israel Journal of Mathematics*, vol. 40, no. 1, pp. 74–96, 1981.
- [20] J. Yang, W. Hu, and C. J. Li, "On the fast convergence of random perturbations of the gradient flow," *Asymptotic Analysis*, vol. 122, no. 3–4, pp. 371–393, 2021.
- [21] B. Shi, W. J. Su, and M. I. Jordan, "On learning rates and Schrödinger operators," *arXiv preprint*, 2020. [Online]. Available: <https://arxiv.org/abs/2004.06977>
- [22] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points—online stochastic gradient for tensor decomposition," in *Conference on learning theory*. PMLR, 2015, pp. 797–842.
- [23] S. S. Du, C. Jin, J. D. Lee, M. I. Jordan, A. Singh, and B. Póczos, "Gradient descent can take exponential time to escape saddle points," in *Advances in neural information processing systems*, 2017, pp. 1067–1077.
- [24] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," in *Proceedings of the 34th International Conference on Machine Learning—Volume 70*. JMLR. org, 2017, pp. 1724–1732.
- [25] Z. Zhou, P. Mertikopoulos, N. Bambos, S. Boyd, and P. W. Glynn, "Stochastic mirror descent in variationally coherent optimization problems," in *Advances in Neural Information Processing Systems*, 2017, pp. 7040–7049.
- [26] H. Daneshmand, J. Kohler, A. Lucchi, and T. Hofmann, "Escaping saddles with stochastic gradients," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1155–1164.
- [27] S. Reddi, M. Zaheer, S. Sra, B. Póczos, F. Bach, R. Salakhutdinov, and A. Smola, "A generic approach for escaping saddle points," in *International conference on artificial intelligence and statistics*. PMLR, 2018, pp. 1233–1242.
- [28] C. Jin, P. Netrapalli, and M. I. Jordan, "Accelerated gradient descent escapes saddle points faster than gradient descent," in *Conference On Learning Theory*. PMLR, 2018, pp. 1042–1085.
- [29] Y. Xu, J. Rong, and T. Yang, "First-order stochastic algorithms for escaping from saddle points in almost linear time," in *Advances in Neural Information Processing Systems*, 2018, pp. 5530–5540.
- [30] Z. Allen-Zhu, "Natasha 2: Faster non-convex optimization than sgd," in *Advances in Neural Information Processing Systems*, 2018, pp. 2675–2686.
- [31] Z. Allen-Zhu and Y. Li, "Neon2: Finding local minima via first-order oracles," in *Advances in Neural Information Processing Systems*, 2018, pp. 3716–3726.
- [32] C. Fang, Z. Lin, and T. Zhang, "Sharp analysis for nonconvex sgd escaping from saddle points," in *Conference on Learning Theory*. PMLR, 2019, pp. 1192–1234.
- [33] S. Paternain, A. Mokhtari, and A. Ribeiro, "A newton-based method for nonconvex optimization with fast evasion of saddle points," *SIAM Journal on Optimization*, vol. 29, no. 1, pp. 343–368, 2019.
- [34] A. Mokhtari, A. Ozdaglar, and A. Jadbabaie, "Escaping saddle points in constrained optimization," in *Advances in Neural Information Processing Systems*, 2018, pp. 3629–3639.
- [35] A. Anandkumar and R. Ge, "Efficient approaches for escaping higher order saddle points in non-convex optimization," in *Conference on learning theory*, 2016, pp. 81–102.
- [36] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, "Accelerated methods for nonconvex optimization," *SIAM Journal on Optimization*, vol. 28, no. 2, pp. 1751–1772, 2018.
- [37] M. Liu, Z. Li, X. Wang, J. Yi, and T. Yang, "Adaptive negative curvature descent with applications in non-convex optimization," *Advances in Neural Information Processing Systems*, vol. 31, pp. 4853–4862, 2018.
- [38] C. Zhang and T. Li, "Escape saddle points by a simple gradient-descent based algorithm," *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [39] D. G. Luenberger, Y. Ye *et al.*, *Linear and nonlinear programming*. Springer, 1984, vol. 2.
- [40] R. M. Corless, G. H. Gonnet, D. E. Hare, D. J. Jeffrey, and D. E. Knuth, "On the lambertw function," *Advances in Computational mathematics*, vol. 5, no. 1, pp. 329–359, 1996.
- [41] C. Ma, K. Wang, Y. Chi, and Y. Chen, "Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution," *Foundations of Computational Mathematics*, vol. 20, no. 3, 2020.
- [42] Y. Matsumoto, *An introduction to Morse theory*. American Mathematical Soc., 2002, vol. 208.
- [43] "Degenerate perturbation theory," <http://farside.ph.utexas.edu/teaching/qmech/Quantum/node105.html#e12.89>, accessed: 2019-08-19.
- [44] "Matrix perturbation theory," [https://ocw.mit.edu/courses/nuclear-engineering/22-51-quantum-theory-of-radiation-interactions-fall-2012/lecture-notes/MIT22\\_51F12\\_Ch11.pdf](https://ocw.mit.edu/courses/nuclear-engineering/22-51-quantum-theory-of-radiation-interactions-fall-2012/lecture-notes/MIT22_51F12_Ch11.pdf), accessed: 2019-08-19.
- [45] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent only converges to minimizers," in *Conference on learning theory*. PMLR, 2016, pp. 1246–1257.
- [46] J. T. Schwartz, *Nonlinear functional analysis*. CRC Press, 1969, vol. 4.
- [47] M. Kirsbraun, "Über die zusammenziehende und lipschitzsche transformationen," *Fundamenta Mathematicae*, vol. 22, no. 1, pp. 77–108, 1934.
- [48] R. Mises and H. Pollaczek-Geiringer, "Praktische verfahren der gleichungsauflösung," *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 9, no. 1, pp. 58–77, 1929.
- [49] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, "Generalized power method for sparse principal component analysis," *Journal of Machine Learning Research*, vol. 11, no. 2, 2010.

- [50] X.-T. Yuan and T. Zhang, "Truncated power method for sparse eigenvalue problems," *Journal of Machine Learning Research*, vol. 14, no. 4, 2013.
- [51] L. A. Rastrigin, "Systems of extremal control," *Nauka*, 1974.
- [52] H. Mühlenbein, M. Schomisch, and J. Born, "The parallel genetic algorithm as function optimizer," *Parallel computing*, vol. 17, no. 6-7, pp. 619–632, 1991.
- [53] F. Hoffmeister and T. Bäck, "Genetic algorithms and evolution strategies: Similarities and differences," in *International Conference on Parallel Problem Solving from Nature*. Springer, 1990, pp. 455–469.
- [54] J. R. Magnus and H. Neudecker, "Matrix differential calculus with applications to simple, hadamard, and kronecker products," *Journal of Mathematical Psychology*, vol. 29, no. 4, pp. 474–492, 1985.
- [55] D. Kinderlehrer and G. Stampacchia, *An introduction to variational inequalities and their applications*. SIAM, 2000.
- [56] J. M. Lee, "Smooth manifolds," in *Introduction to Smooth Manifolds*. Springer, 2013, pp. 1–31.
- [57] W. Rudin *et al.*, *Principles of mathematical analysis*. McGraw-hill New York, 1976, vol. 3.

**Rishabh Dixit** received the B. Tech. degree in electrical engineering from the Indian Institute of Technology, Kanpur, India, in 2015. He is currently working toward a Ph.D. degree in the Department of Electrical and Computer Engineering, Rutgers University–New Brunswick, NJ, USA. His research interests include algorithms development for convex and nonconvex optimization, dynamic optimization, and machine learning.

**Mert Gürbüzbalaban** is an associate professor at Rutgers University. Previously, he was an assistant professor at Rutgers University and a postdoctoral associate at the Laboratory for Information and Decision Systems (LIDS) at MIT. He is broadly interested in optimization and computational science driven by applications in network science and data science. He received his B.Sc. degrees in Electrical Engineering and Mathematics as a valedictorian from Boğaziçi University, Istanbul, Turkey, the "Diplôme d'ingénieur" degree from École Polytechnique, France, and the M.S. and Ph.D. degrees in Mathematics from the Courant Institute of Mathematical Sciences, New York University.

Dr. Gürbüzbalaban is a recipient of the Dean's Research Award, Dean's Young Research Award and Dean's Summer Fellowship Award at the Rutgers Business School and a co-recipient of the Honorable Mention for the Best Paper Award at the International Conference in Machine Learning (ICML 2019). He also received the Kurt Friedrichs Prize (given by the Courant Institute of New York University for an outstanding thesis) in 2013, Bronze Medal in the École Polytechnique Scientific Project Competition in 2006, the Nadir Orhan Bengisu Award (given by the electrical-electronics engineering department of Boğaziçi University to the best graduating undergraduate student) in 2005 and the Bülent Kerim Altay Award from the Electrical-Electronics Engineering Department of Middle East Technical University in 2001. He served as a special issue editor for the Probability in the Engineering and Informational Sciences journal, as a member of the Informs Nicholson Prize Committee in 2021 and as a track chair in Operations Research for the Institute of Industrial and Systems Engineering (IISE) Conference in 2019.

**Waheed U. Bajwa (S'98–M'09–SM'13)** received BE (with Honors) degree in electrical engineering from the National University of Sciences and Technology, Pakistan in 2001, and MS and PhD degrees in electrical engineering from the University of Wisconsin-Madison in 2005 and 2009, respectively. He has been with Rutgers University–New Brunswick since 2011, where he is currently a professor and graduate director in the Department of Electrical and Computer Engineering and an associate member of the graduate faculty of the Department of Statistics. He was also a postdoctoral research associate in the Program in Applied and Computational Mathematics at Princeton University from 2009 to 2010, a research scientist in the Department of Electrical and Computer Engineering at Duke University from 2010 to 2011, and a visiting fellow in the Center for Statistics and Machine Learning at Princeton University from 2021 to 2022. His research interests include statistical signal processing, high-dimensional statistics, machine learning, inverse problems, and networked systems.

Dr. Bajwa has received several research and teaching awards including the Army Research Office Young Investigator Award (2014), the National Science Foundation CAREER Award (2015), Rutgers Presidential Merit Award (2016), Rutgers Presidential Fellowship for Teaching Excellence (2017), Rutgers Engineering Governing Council ECE Professor of the Year Award (2016, 2017, 2019), Rutgers Warren I. Susman Award for Excellence in Teaching (2021), and Rutgers Presidential Outstanding Faculty Scholar Award (2022). He is a co-investigator on a work that received the Cancer Institute of New Jersey's Gallo Award for Scientific Excellence in 2017, a co-author on papers that received Best Student Paper Awards at IEEE IVMSP 2016 and IEEE CAMSAP 2017 workshops, and a Member of the Class of 2015 National Academy of Engineering Frontiers of Engineering Education Symposium.

Dr. Bajwa has also been involved in numerous professional activities. He has served as the Lead Guest Editor for a special issue of IEEE Signal Processing Magazine on "Distributed, Streaming Machine Learning" (2020), a Guest Editor for a special issue of Proceedings of the IEEE on "Optimization for Data-driven Learning and Control" (2020), an Associate Editor of IEEE Transactions on Signal and Information Processing over Networks (2018 – 2022), an Associate Editor of IEEE Signal Processing Letters (2014 – 2017), and a Guest Editor for a special issue of Elsevier Physical Communication Journal on "Compressive Sensing in Communications" (2012). He also served as a Technical Area Chair of Asilomar Conference on Signals, Systems, and Computers in 2021 and 2018, the US Liaison Chair of IEEE SPAWC 2019 Workshop, a Publicity & Publications Co-Chair of IEEE DSW 2019 Workshop, a Technical Co-Chair of IEEE SPAWC 2018 Workshop, the General Chair of 2017 DIMACS Workshop on Distributed Optimization, Information Processing, and Learning, the Publicity and Publications Chair of IEEE CAMSAP 2015 Workshop, and a General Co-Chair of IEEE GlobalSIP 2013 Symposium on New Sensing and Statistical Inference Methods and CP-SWeek 2013 Workshop on Signal Processing Advances in Sensor Networks. Additionally, he has served within the IEEE Signal Processing Society as an elected member of the Sensor Array and Multichannel (SAM) Technical Committee (2016 – 2021), an elected member of the Signal Processing for Communications and Networking (SPCOM) Technical Committee (2016 – 2021), an elected member of the Machine Learning for Signal Processing (MLSP) Technical Committee (2016 – 2018), an appointed member of the Data Science Initiative (2019 – 2021), and an elected member of the Big Data Special Interest Group (2019). He is currently serving as a Member of the Senior Editorial Board of the IEEE Signal Processing Magazine, a Senior Area Editor of IEEE Open Journal of Signal Processing and IEEE Signal Processing Letters, the Vice Chair of the SPCOM Technical Committee of the IEEE Signal Processing Society, and an elected member of the Signal Processing Theory and Methods (SPTM) Technical Committee of the IEEE Signal Processing Society.