Journal of Machine Learning Research 22 (2021) 1-63 Submitted 11/19; Revised 3/21; Published 3/21

Path Length Bounds for Gradient Descent and Flow

Chirag Gupta CHIRAGG@CMU.EDU
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15217, USA

Sivaraman Balakrishnan SIVA@STAT.CMU.EDU
Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA 15217, USA

Aaditya Ramdas ARAMDAS@CMU.EDU
Department of Statistics and Data Science, Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15217, USA

Editor: Ambuj Tewari

Abstract

We derive bounds on the path length ζ of gradient descent (GD) and gradient flow (GF) curves for various classes of smooth convex and nonconvex functions. Among other results, we prove that: (a) if the iterates are linearly convergent with factor $(1-c)$, then ζ is at most $\mathcal{O}(1/c)$; (b) under the Polyak-Kurdyka-Łojasiewicz (PKL) condition, ζ is at most $\mathcal{O}(\sqrt{\kappa})$, where κ is the condition number, and at least $\widetilde{\Omega}(\sqrt{d}\wedge\kappa^{1/4})$; (c) for quadratics, ζ is $\Theta(\min\{\sqrt{d},\sqrt{\log\kappa}\})$ and in some cases can be independent of κ; (d) assuming just convexity, ζ can be at most $2^{4d\log d}$; (e) for separable quasiconvex functions, ζ is $\Theta(\sqrt{d})$. Thus, we advance current understanding of the properties of GD and GF curves beyond rates of convergence. We expect our techniques to facilitate future studies for other algorithms.

Keywords: optimization, trajectory analysis, condition number, self-contracted curves, Polyak-Kurdyka-Łojasiewicz functions

1. Introduction

Recent work in machine learning has sought to understand the empirical success of gradient descent (GD) through trajectory analysis (Ge et al., 2015; Lee et al., 2019; Li et al., 2018; Oymak and Soltanolkotabi, 2019). Trajectory analysis techniques aim to discover desirable geometric properties satisfied by the GD curve that lead to good convergence behavior. One such property is an upper bound on the path length of the GD curve. Path length bounds have been used in recent convergence analyses for deep neural networks (Du et al., 2018, 2019; Allen-Zhu et al., 2019). These papers argue that fast convergence can be achieved even if desirable curvature properties are not satisfied globally, provided they hold within a local region around initialization. A path length bound is used to ensure that the iterates stay within this fast convergence region. In this work, we study the path length of GD and gradient flow (GF) curves as an independent object of interest. This problem has been considered from slightly different perspectives in machine learning (Oymak and Soltanolkotabi, 2019) and convex analysis (Bolte et al., 2010; Manselli and Pucci, 1991; Daniilidis et al., 2015). We draw from both these areas and further the study by establishing novel upper and lower path length bounds in a variety of settings.

Formally, consider the problem of minimizing an objective function $f : \mathbb{R}^d \to \mathbb{R}$ using an iterative update rule. An optimization curve refers to the sequence of iterates of the update rule and is denoted by a mapping $g : S \to \mathbb{R}^d$, where S is either $\mathbb{R}_0^+$ or $\mathbb{N}_0$. $S = \mathbb{R}_0^+$ corresponds to the continuous time setting, where the iterative update rule is specified using an ordinary differential equation. Here, we typically denote an element of S as t, to be thought of as 'time'. On the other hand, $S = \mathbb{N}_0$ corresponds to discrete update rules, where we denote an element of S as k, to be thought of as an iterate count. In both cases, we use $\mathbf{x}_s$ to denote $g(s)$. Iterative optimization techniques construct this mapping g using local update rules based on the gradient of f at $\mathbf{x}_s$, starting at some initial point $\mathbf{x}_0$. If f is differentiable, one such update rule takes the following form:

$$\text{Gradient flow (GF):} \qquad \dot{\mathbf{x}}_t = -\nabla f(\mathbf{x}_t).$$

A forward Euler discretization of the above ordinary differential equation with a fixed step-size η yields

$$\text{Gradient descent (GD):} \qquad \mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla f(\mathbf{x}_k).$$

In this paper, we bound the path lengths of the aforementioned update rules:

$$\text{(continuous)} \qquad \zeta(f, \mathbf{x}_0) := \int_0^\infty \|\dot{\mathbf{x}}_t\|_2 \, dt = \int_0^\infty \|\nabla f(\mathbf{x}_t)\|_2 \, dt.$$

$$\text{(discrete)} \qquad \zeta_\eta(f, \mathbf{x}_0) := \sum_{k=0}^{\infty} \|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 = \sum_{k=0}^{\infty} \|\eta \nabla f(\mathbf{x}_k)\|_2 \, .$$

Under the assumptions specified in the next section, the above integrals and sums are well defined—our results implicitly show that they converge to a finite value.

We use the terms *path* and *curve* to refer to the same mathematical object, but we make the following stylistic choices for their usage. The mathematical object itself is typically referred to as the 'GD curve' or 'GF curve', while when discussing its length we always use the term 'path length' instead of 'curve length'.

1.1 Notation and Assumptions

Throughout this paper, we assume that the objective function f is continuously differentiable on $\mathbb{R}^d$, and that the minimum of the function f is achieved by at least one point $\mathbf{x}^*$ with $\|\mathbf{x}^*\|_2 < \infty$. Denote the minimum value of the objective as

$$f^* := \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \ = \ f(\mathbf{x}^*),$$

and the set of all minimizers as $\mathbf{X}^* := \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) = f^*\}$. By first order optimality, $\nabla f(\mathbf{x}^*) = 0$ is a necessary (but not sufficient) condition for $\mathbf{x}^*$ to be an element of $\mathbf{X}^*$.

Theorem	Assumption	Upper bound	Lower bound
Theorem 6, 8, 17	LG, $(1, c)$-linear convergence	$\mathcal{O}(1/c)$	$\widetilde{\Omega}(\sqrt{1/c})$
Theorem 9*, 17	PKL, LG	$\mathcal{O}(\sqrt{\kappa})$	$\widetilde{\Omega}(\min\{\kappa^{1/4}, \sqrt{d}\})$
Theorem 10, 18	Quadratic objective	$\mathcal{O}(\min\{\sqrt{\log \kappa}, \sqrt{d}\})$	$\Omega(\min\{\sqrt{\log \kappa}, \sqrt{d}\})$
Theorem 15*, 18	Convexity, LG	$e^{\mathcal{O}(d \log d)}$	$\Omega(\sqrt{d})$
Theorem 16, 18	Quasiconvexity, LG, separability	$\mathcal{O}(\sqrt{d})$	$\Omega(\sqrt{d})$

Table 1: Summary of path length bounds for GF and GD. LG refers to Lipschitz Gradients (Definition 1) and PKL is the Polyak-Kurdyka-Łojasiewicz condition (Definition 3). Other terms are also defined in Section 1.1. *The GF version of Theorem 9 is due to Bolte et al. (2010). The GF version of Theorem 15 is due to Manselli and Pucci (1991). Further relationships to prior work are discussed in Section 1.2.

Under the standard curvature assumptions that we consider later in this paper $\mathbf{X}^*$ will be a convex set. Note that $\mathbf{X}^*$ is closed since f is continuous. Hence the projection of the initial point $\mathbf{x}_0$ on the optimal set $\mathbf{X}^*$ is uniquely defined. We denote this projection as $\Pi_{\mathbf{X}^*}(\mathbf{x}_0)$. We then denote the minimum distance between the initialization point $\mathbf{x}_0$ and the optimal set $\mathbf{X}^*$ as

$$\operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*) := \min_{\mathbf{x}^* \in \mathbf{X}^*} \|\mathbf{x}_0 - \mathbf{x}^*\|_2 = \|\mathbf{x}_0 \ - \ \Pi_{\mathbf{X}^*}(\mathbf{x}_0)\|_2 ,$$

which is finite since $\operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*) \le \operatorname{dist}(\mathbf{x}_0, \mathbf{x}^*) < \infty$. Many of our path length guarantees are a product of $\operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*)$ and a factor (> 1) that depends on curvature or smoothness properties of the function (defined shortly) or the dimension d. When we write path length bounds using $\mathcal{O}(\cdot)$ or $\Omega(\cdot)$ notation, we absorb the factor $\operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*)$ as a constant.

Most of our results are under the following standard smoothness assumption on f (for instance, see Nesterov (2013, Section 1.2.2)).

Definition 1 (Lipschitz Gradients (LG)) *For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and some $L > 0$,*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \le L \|\mathbf{x} - \mathbf{y}\|_2 .$$

The above condition implies $f(\mathbf{y}) - f(\mathbf{x}) \le \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|_2^2$.

We next introduce the curvature conditions under which path length bounds are studied in this paper.

Definition 2 (Convexity and quasiconvexity (QC)) *f is convex if $f(\mathbf{y}) \ge f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. f is quasiconvex if $f(\mathbf{y}) \le f(\mathbf{x}) \implies \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \le 0$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.*

Since f is differentiable, this condition for quasiconvexity is equivalent to saying that all sub-level sets of f are convex (Avriel et al., 2010, Chapter 3). It is also equivalent to the condition $f(t\mathbf{x} + (1 - t)\mathbf{y}) \le \max\{f(\mathbf{x}), f(\mathbf{y})\}$ for all $t \in [0, 1]$ (Fenchel, 1953). It is clear from the definition that convex functions are quasiconvex.

Definition 3 (Polyak-Kurdyka-Łojasiewicz condition (PKL)) *For all $\mathbf{x} \in \mathbb{R}^d$, and some $\mu > 0$,*

$$\|\nabla f(\mathbf{x})\|_2^2 \geq 2\mu \left(f(\mathbf{x}) - f^*\right).$$

The PKL condition implies quadratic growth: $f(\mathbf{x}) - f^* \geq \frac{\mu}{2} dist^2(\mathbf{x}, \mathbf{X}^*)$ (Karimi et al., 2016), and hence also linear growth for the gradient: $\|\nabla f(\mathbf{x})\| \geq \mu \cdot dist(\mathbf{x}, \mathbf{X}^*)$.

This inequality was introduced by Polyak (1963) to show linear convergence for gradient descent. Independently, Łojasiewicz (1963) showed that a generalized version of this inequality is true locally around a critical point for any real-analytic function. This result was further extended by Kurdyka (1998). The generalized inequality is now referred to as the Kurdyka-Łojasiewicz (KL) inequality—we refer the reader to Bolte et al. (2007, 2010) for more background.

The PKL inequality is weaker than strong convexity:

$$\text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d,\ f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|_2^2,$$

in the sense that μ-strongly convex functions are also μ-PKL. Linear convergence can be shown for GD if f satisfies μ-PKL (instead of the stronger μ-strong convexity) and has Lipschitz gradients (Polyak, 1963). Yet the PKL condition incorporates a significantly larger class of functions; in particular PKL functions can even be nonconvex. Nonconvex PKL function arise naturally in some machine learning optimization problems (for instance, see Lemma 26 (Soltanolkotabi et al., 2019), Section 3 (Fazel et al., 2018), Equation (3.4) (Balakrishnan et al., 2017) for nonconvex examples and Section 2.3 (Karimi et al., 2016) for convex but non-strongly convex examples).

Convergence of GD is often studied with respect to its dependence on the ratio of L and μ, called as the condition number (for instance, see Nesterov (2013, Section 2.1.3)).

Definition 4 (Condition number) *For a μ-PKL function f with L-Lipschitz gradients, we define the condition number κ as L/μ.*

Convex quadratic objective functions satisfy PKL and LG, and the κ here turns out to be the ratio of the largest and smallest non-zero singular values of the Hessian. We also consider path length bounds under a general linear convergence assumption, formalized next. The following definition is for any discrete-time iterative updates $\{\mathbf{x}_k\}_{k \in \mathbb{N}_0}$ or any continuous-time dynamics $\{\mathbf{x}_t\}_{t \geq 0}$ (not restricted to GD or GF).

Definition 5 ((A, c)-Linear convergence) *We say that a procedure is linearly convergent with constants $c \in (0, 1)$ and $A \geq 1$ if for every initial point $\mathbf{x}_0 \in \mathbb{R}^d$ and any $s \geq 0$,*

$$\text{dist}\left(\mathbf{x}_s, \mathbf{X}^*\right) \leq A(1 - c)^s \text{dist}\left(\mathbf{x}_0, \mathbf{X}^*\right). \tag{2}$$

In the discrete case, we have $s \in \mathbb{N}_0$ and in the continuous case we have $s \in \mathbb{R}_0$. In the discrete case c, should really be thought of as c_η since the linear convergence rate depends on the step-size η. However, we use c in both cases to simplify exposition.

Table 1 summarizes our results in these various settings. Figure 1 shows the containment relationships between various conditions introduced here (these are standard known

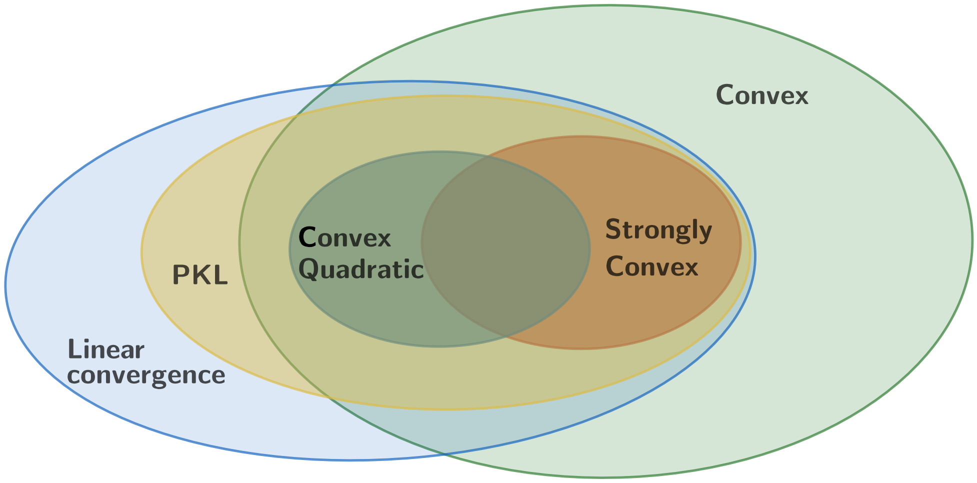


Figure 1: Venn diagram showing different curvature conditions (defined in Section 1.1) and their relationships. "Linear convergence" refers to functions for which GD or GF is linearly convergent. The quadratic losses we consider may have Hessians with zero singular values, in which case the loss is not strongly convex.

facts). Finally, we note that the smoothness and curvature assumptions we make are global. However, since path length bounds imply that the iterates always stay within a ball, these assumptions can always be restricted to that ball (as is done formally by Oymak and Soltanolkotabi (2019)).

1.2 Prior Work

The problem of bounding the length of GF curves of quasiconvex functions was first considered by Manselli and Pucci (1991). Their analysis is via a reduction to Lipschitz continuous curves (meaning that the curve map $g : \mathbb{R} \to \mathbb{R}^d$ is Lipschitz continuous) that exhibit the self-contracted property (see Definition 13). Under an a priori assumption that the GF curve has finite length, they prove that the length can be at most $2^{\mathcal{O}(d \log d)}$. Daniilidis et al. (2015, Corollary 2.4) show that the finiteness assumption can be dispensed. Further while Manselli and Pucci require the GF curve to be smooth, Daniilidis et al. provide a new analysis that includes non-smooth self-contracted curves (with a weaker 2^{10d^2} bound). This fact enables us to extend the analysis from GF curves to GD curves in the convex+LG case (Theorem 15), leading to the first bound in this setting. Path length bounds for self-contracted curves have been studied more generally in non-Euclidean settings (Stepanov and Teplitskaya, 2017; Daniilidis et al., 2018). It has also been shown that self-contracted curves that satisfy certain smoothness properties must be GF curves of convex functions (Durand-Cartagena and Lemenant, 2019). The papers in this line of work use geometric properties of self-contracted curves without referencing the specifics of the GF dynamics.

While these papers are concerned about fundamental questions about GF curves under the mild assumption of quasiconvexity, we know that quasiconvexity, or even convexity, is not enough to obtain the fast convergence rates that optimization algorithms exhibit empirically in many machine learning settings. Indeed, we may hope for path length bounds that are potentially independent of d when assumptions such as strong convexity or PKL hold. Bolte et al. (2010) showed a dimension independent path length bound for GF curves whenever f satisfies a Kurdyka-Łojasiewicz (KL) condition, a larger class that subsumes the PKL condition. Their result provides the best known upper bound of $\mathcal{O}(\sqrt{\kappa})$ for GF curves of PKL functions (reproduced in Theorem 9). However, for quadratic objective functions, we show an exponentially improved $\mathcal{O}(\sqrt{\log \kappa})$ bound (Theorem 10). The best known bound for strongly convex functions is the $\mathcal{O}(\sqrt{\kappa})$ bound for PKL functions, but our findings for quadratics suggests that this bound may be loose. We also show two novel lower bounds in these settings: $\widetilde{\Omega}(\kappa^{1/4})$ for PKL functions (Theorem 17) and $\Omega(\sqrt{\log \kappa})$ for quadratics (Theorem 18). Thus, although the KL condition holds very generally, an all purpose bound depending on the KL constant may not be tight. Oymak and Soltanolkotabi (2019) analyzed GD and SGD path lengths for nonlinear least square objectives, under spectral assumptions on the Jacobian of the non-linear mapping. Apart from least squares objectives, they also showed an $\mathcal{O}(\kappa)$ bound for PKL functions. We improve their result by showing an $\mathcal{O}(\sqrt{\kappa})$ bound for GD (Theorem 9).

1.3 Organization

The rest of the paper is organized as follows:

1. In Section 2.1, we prove a general purpose path length bound for GD and GF curves that is applicable under any set of curvature conditions for which linear convergence can be established. Using the technique of Section 2.1, in Appendix A we show path length bounds for Polyak's heavy ball method and projected gradient descent.

2. In Section 2.2 we show path length bounds for GD and GF curves of PKL functions. In Section 4.1, we present a worst case lower bound in the same setting.

3. In Section 2.3, we provide the tightest known path length bound for GD and GF curves of convex quadratic objective functions (potentially overparameterized). In Section 4.2, we provide a a matching lower bound construction.

4. In Section 3 we derive explicit dimension dependent path length bounds for quasiconvex functions (GF) and convex functions (GD and GF). This leads to the first known bounds for GD curves in this setting.

Table 1 summarizes the results in this paper. We remark that proofs in the GF case are often more straightforward, but lead to conclusions that continue to hold in the GD case with appropriate step-size restrictions.

2. Dimension Independent Path Length Bounds

In this section, we provide dimension independent bounds on the length GD and GF curves. All the bounds in this section hold for functions for which GD and GF exhibit linear convergence towards the optimal set. In particular we discuss PKL objectives, strongly convex objectives, and convex quadratic objectives. More generally if the function does not belong to one of the aforementioned function classes, but linear convergence is known to hold, we can prove path length bounds that depend on the constant of convergence. We first prove this meta-theorem and then discuss specific function classes (where the meta-theorem bound can be improved).

2.1 General Bounds Under Linear Convergence

Linear convergence (Definition 5) is known for GD (with appropriate step-size) if the function has Lipschitz gradients and is strongly convex with $A = 1$ and $c = 1/\kappa$. For PKL functions we have such a linear convergence result in function values instead of distance to the optimal set with $A = 1$ and $c = 1/\kappa$ (Karimi et al., 2016, Theorem 1). Using the LG condition and a quadratic growth result due to Karimi et al. (2016, Theorem 2), this convergence in function value can be converted to parametric linear convergence in the sense of Definition 5 with $A = \kappa$ and $c = 1/\kappa$. All of these convergence results are also known for GF with $c = 1 - e^{-\mu}$ in each case. GD is also known to be linearly convergent when performing maximum likelihood estimation in logistic regression with unseparable data (Freund et al., 2018, Theorem 3.3). In the theorem to follow, we guarantee a path length bound in each of these settings.

Theorem 6 *Suppose f has L-Lipschitz gradients. If the GF dynamics for f exhibits linear convergence with constants (A, c), then its path length is bounded as:*

$$\zeta \leq \left(AL/\log\left(1/(1-c)\right)\right)\, \mathrm{dist}\left(\mathbf{x}_0, \mathbf{X}^*\right), \tag{3}$$

and if the GD iterates with step-size η exhibit linear convergence with constants (A, c), then their path length is bounded as:

$$\zeta_\eta \leq (\eta AL/c)\, \mathrm{dist}\left(\mathbf{x}_0, \mathbf{X}^*\right). \tag{4}$$

Proof sketch The detailed proof of a more general result can be found in Theorem 20, Appendix A. As a consequence of linear convergence, as the distance to the optimal set decreases geometrically, so does the contribution of consecutive iterates to the path. As illustrated in Figure 2, we then have the bound

$$\zeta_\eta \leq \mathrm{dist}\left(\mathbf{x}_0, \mathbf{X}^*\right)\left(\eta AL + (1-c)\eta AL + (1-c)^2\eta AL + \ldots\right) = (\eta AL/c)\, \mathrm{dist}\left(\mathbf{x}_0, \mathbf{X}^*\right), \tag{5}$$

which proves claim (4). ■

Remark 7 *Variants of this theorem can be derived in more general settings:*

(a) Theorem 20 in Appendix A extends the result from linear convergence towards the globally optimal set $\mathbf{X}^*$ to linear convergence towards any convex set of points that satisfy first order optimality conditions.

(b) Theorem 23 in Appendix A extends the result to projected gradient descent, with a larger multiplicative constant.

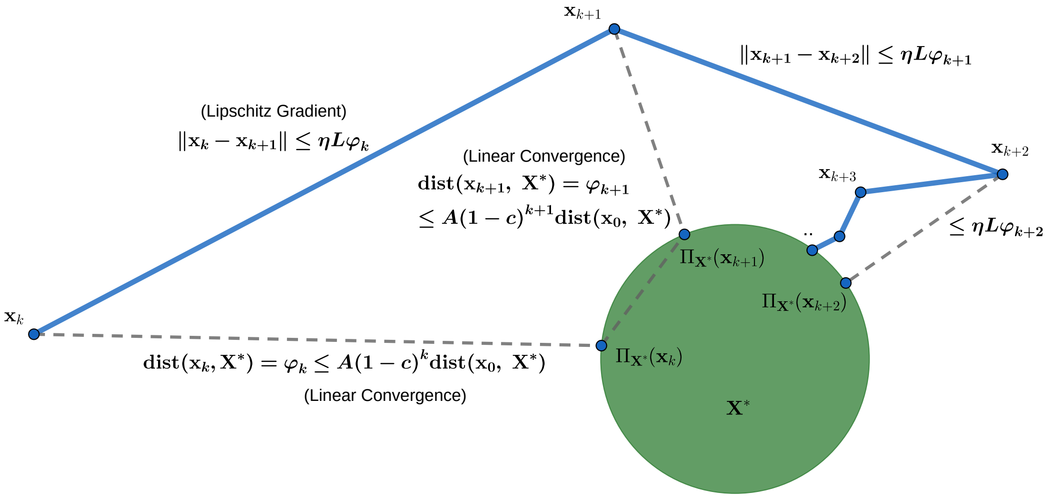

$\mathbf{x}_{k+1}$
$\|\mathbf{x}_{k+1} - \mathbf{x}_{k+2}\| \leq \eta L \varphi_{k+1}$
(Lipschitz Gradient)
$\|\mathbf{x}_k - \mathbf{x}_{k+1}\| \leq \eta L \varphi_k$
(Linear Convergence)
$\mathbf{dist}(\mathbf{x}_{k+1},\ \mathbf{X}^*) = \varphi_{k+1}$
$\leq A(1-c)^{k+1}\mathbf{dist}(\mathbf{x}_0,\ \mathbf{X}^*)$
$\mathbf{x}_{k+3}$
$\mathbf{x}_{k+2}$
..
$\leq \eta L \varphi_{k+2}$
$\Pi_{\mathbf{X}^*}(\mathbf{x}_{k+1})$
$\Pi_{\mathbf{X}^*}(\mathbf{x}_{k+2})$
$\mathbf{x}_k$
$\mathbf{dist}(\mathbf{x}_k, \mathbf{X}^*) = \varphi_k \leq A(1-c)^k\mathbf{dist}(\mathbf{x}_0,\ \mathbf{X}^*)$
(Linear Convergence)
$\Pi_{\mathbf{X}^*}(\mathbf{x}_k)$
$\mathbf{X}^*$

Figure 2: Path length bound under linear convergence. The GD curve is shown with solid blue lines. The inequalities in bold upper bound individual terms in the path length sum, as noted in Equation (5).

(c) Theorem 8 extends the result to any iterative update rule not limited to GD, if descent can be established with respect to all minimizers in $\mathbf{X}^*$*.*

For GD with a μ-strongly convex function that has L-Lipschitz gradients, it is a standard result that linear convergence holds for $\eta \leq 1/L$, with $A = 1$ and $c = \eta\mu$. This leads to the bound $\zeta_\eta \leq \kappa \|\mathbf{x}_0 - \mathbf{x}^*\|_2$. For GF in the same setting, linear convergence holds with $A = 1$ and $c = 1 - e^{-\mu}$. Since $\log(1/(1-c)) = \log(e^\mu) = \mu$, the bound also resolves to $\zeta \leq \kappa \|\mathbf{x}_0 - \mathbf{x}^*\|_2$. However, the $\mathcal{O}(\kappa)$ dependence can be improved. In Section 2.2, following the work of Bolte et al. (2010), we show that in fact the κ bound for μ-strongly convex functions can be improved to $\mathcal{O}(\sqrt{\kappa})$ with a weaker μ-PKL assumption. In Section 2.3 we study a specific strongly convex problem, namely convex quadratic objective functions. In this case we show that the above bound can be improved to $\mathcal{O}(\sqrt{\log \kappa})$.

Theorem 6 can be generalized to include any iterative update rule (not limited to GD), and any function class for which linear convergence holds, as long as descent holds with respect to all minimizers in $\mathbf{X}^*$. In the following, we denote the update for a point $\mathbf{x}$ as $\mathbf{x}^+$.

Theorem 8 *If the iterates for a discrete update rule satisfy the following conditions:*

(a) linear convergence with constants (A, c)*, and*

(b) descent towards all minimizers: for all $\mathbf{x}^* \in \mathbf{X}^*$*,* $\mathbf{x} \in \mathbb{R}^d$*,* $\|\mathbf{x}^+ - \mathbf{x}^*\|_2 \leq \|\mathbf{x} - \mathbf{x}^*\|_2$*,*

then their path length is bounded as:

$$\zeta_\eta \leq \left(\frac{2A}{c}\right) \operatorname{dist}\left(\mathbf{x}_0, \mathbf{X}^*\right). \tag{6}$$

The proof of a more general result can be found in Appendix A (see Theorem 21). Note that if $\mathbf{X}^*$ is singleton and $A = 1$, condition (b) is implied by condition (a). Using this observation, in Corollary 22 (Appendix A) we show a path length bound of $\sqrt{\kappa}\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2$ for Polyak's heavy ball method (Polyak, 1964) with a twice continuously differentiable and strongly convex function with Lipschitz gradients.

For GD, condition (b) of Theorem 8 is known if f is convex, has L-Lipschitz gradients and $\eta \leq 1/L$—in fact Lemma 14 argues that the stronger property of self-contractedness (Definition 13) holds. However, there exist scenarios where condition (a) is known but not condition (b), such as nonconvex PKL functions. In such cases, Theorem 6 can still be applied, as long as f has L-Lipschitz gradients.

2.2 Path Length Under the PKL Condition is $\mathcal{O}(\sqrt{\kappa})$

The bound in the PKL case can be improved by analyzing a particular potential/Lyapunov function:

$$\varepsilon_t = \sqrt{f(\mathbf{x}_t) - f^*}.$$

It can be shown by differentiating ε_t and applying the PKL condition that,

$$-\dot{\varepsilon}_t \geq \sqrt{\frac{\mu}{2}}\left\|\nabla f(\mathbf{x}_t)\right\|_2.$$

Integrating the above with respect to t and using LG leads to the following theorem path length bound for GF curves (the GD bound also follows a similar approach).

Theorem 9 *For any μ-PKL function f with L-Lipschitz gradients, the GF dynamics have a path length bounded as:*

$$\zeta \leq \sqrt{\kappa}\,\mathrm{dist}\left(\mathbf{x}_0, \mathbf{X}^*\right),$$

while the GD iterates with $\eta \leq 1/L$ have a path length bounded as: $\zeta_\eta \leq 2\sqrt{\kappa}\,\mathrm{dist}\left(\mathbf{x}_0, \mathbf{X}^*\right)$.

The proof can be found in Appendix B. The GF version of this bound is a special case of a general statement shown by Bolte et al. (2010, Theorem 27) for a larger class of functions, namely those that satisfy the Kurdyka-Łojasiewicz inequality. The GD version is new. Oymak and Soltanolkotabi (2019, Corollary 5.3) obtained an $\mathcal{O}(\kappa)$ bound while we show an $\mathcal{O}(\sqrt{\kappa})$ bound. Our result relies on Theorem 5.2 of their paper, where they show a path length bound in terms of $(f(\mathbf{x}_0) - f^*)$. However their final result in terms of $\mathrm{dist}\left(\mathbf{x}_0, \mathbf{X}^*\right)$ is weaker than ours.

2.3 Path Length for Convex Quadratic Objectives is $\mathcal{O}(\sqrt{\log \kappa})$

For convex quadratic objective functions we can explicitly write down the GD or GF iterates. This allows us to significantly improve the general linear convergence result. To make use of some standard notation, we write a general convex quadratic objective function as a linear regression problem specified by a matrix $\mathbf{A}$ of dimensions $n \times d$, and an output vector $\mathbf{y} \in \mathbb{R}^n$. The objective is

$$f(\mathbf{x}) = \frac{1}{2n}\left\|\mathbf{y} - \mathbf{A}\mathbf{x}\right\|_2^2. \tag{7}$$

The columns of $\mathbf{A}$ may be linearly dependent (which is necessarily true for the over-parameterized setting, when $d > n$). In this case the solution set $\mathbf{X}^*$ has more than one element. However, it is possible to show that both GF and GD converge to $\Pi_{\mathbf{X}^*}(\mathbf{x}_0)$, given by

$$\Pi_{\mathbf{X}^*}(\mathbf{x}_0) := (\mathbf{I}_{n\times n} - \mathbf{A}^T(\mathbf{A}^T)^\dagger)\mathbf{x}_0 + (\mathbf{A}^T\mathbf{A})^\dagger\mathbf{A}^T\mathbf{y},$$

where $B^\dagger$ denotes the Moore-Penrose inverse of a matrix B. If the columns of $\mathbf{A}$ are linearly independent, this reduces to the standard least squares solution $\Pi_{\mathbf{X}^*}(\mathbf{x}_0) = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}$.

Define $\Sigma := (\mathbf{A}^T\mathbf{A})/n$. The GF curve can be computed in closed form as:

$$\mathbf{x}_t = \Pi_{\mathbf{X}^*}(\mathbf{x}_0) - \Sigma^\dagger \exp(-t\Sigma)\Sigma(\Pi_{\mathbf{X}^*}(\mathbf{x}_0) - \mathbf{x}_0), \tag{8}$$

so that $\mathbf{x}_\infty = \Pi_{\mathbf{X}^*}(\mathbf{x}_0)$. By definition, Σ is positive semidefinite. If $\mathbf{A}$ has linearly dependent columns, Σ would have zero singular values. In general, suppose the number of non-zeros singular values of Σ is $d^+ \leq d$. We denote them as $\sigma_1 \geq \sigma_2 \cdots \geq \sigma_{d^+} > 0$. Then for any step-length $\eta \leq 1/\sigma_1$, the GD iterates converge to $\Pi_{\mathbf{X}^*}(\mathbf{x}_0)$ via the following updates for $k \geq 1$:

$$\mathbf{x}_k = \mathbf{x}_{k-1} - \frac{\eta\mathbf{A}^T(\mathbf{A}\mathbf{x}_{k-1} - \mathbf{y})}{n}. \tag{9}$$

For $i \in [d^+ - 1]$, define $\kappa_i := \sigma_i/\sigma_{i+1}$. The overall condition number is $\kappa := \sigma_1/\sigma_{d^+}$. The following theorem shows a path length bound for quadratic objectives in terms of each of the quantities: d^+, the κ_i's and κ.

Theorem 10 *For convex quadratic objective functions* (7)*, the GF dynamics* (8) *have a path length bounded as:*

$$\zeta \leq \min\left\{\sqrt{d^+}, 1 + \sum_{j=1}^{d^+-1} \kappa_j^{-1/(\kappa_j-1)}(1 - 1/\kappa_j), 1 + 2.5\sqrt{\log\kappa}\right\} \text{ dist}\,(\mathbf{x}_0, \mathbf{X}^*), \tag{10}$$

while the GD iterates (9) *with* $\eta \leq 1/\sigma_1$ *have a path length bounded as:*

$$\zeta_\eta \leq \text{ dist}\,(\mathbf{x}_0, \mathbf{X}^*) + \zeta. \tag{11}$$

To clarify, when $\kappa_j = 1$, $\kappa_j^{-1/(\kappa_j-1)}(1-1/\kappa_j)$ is defined as $\lim_{\kappa_j\to 1^+} \kappa_j^{-1/(\kappa_j-1)}(1-1/\kappa_j) = 0$. The proof of Theorem 10 can be found in Appendix C.

Remark 11 *We show in the proof that* $\sum_{j=1}^{d^+-1} \kappa_j^{-1/(\kappa_j-1)}(1-1/\kappa_j) \leq \frac{\log\kappa}{e}$. *This cannot be improved to* $o(\log\kappa)$ *as shown next. Suppose* $\kappa_j = 6$ *for every* $j \in [d^+ - 1]$. *It can be verified that* $6^{-1/(6-1)}(1 - 1/6) \geq 0.5$, *and thus* $\sum_{j=1}^{d^+-1} \kappa_j^{-1/(\kappa_j-1)}(1 - 1/\kappa_j) \geq 0.5(d^+ - 1) \geq \frac{\log\kappa}{2\log 6}$, *since* $\kappa = 6^{d^+-1}$. *This is true for any* d^+ *and consequently for all large* κ. *Thus the bound cannot be improved to* $o(\log\kappa)$ *in general, making the* $\mathcal{O}(\sqrt{\log\kappa})$ *bound the best asymptotic result with respect to* κ. *However for special cases, such as if there is only one large singular value, this bound may even be independent of* κ *and* d^+. *For instance, suppose* $\kappa_j = 1$ *for* $j \in [d^+ - 2]$, *and* $\kappa = \kappa_{d^+-1} \in [1, \infty)$. *Then, no matter the value of* d^+, κ, *we obtain the bound* $\zeta \leq 2 \text{ dist}\,(\mathbf{x}_0, \mathbf{X}^*)$.

There exists a family of quadratic functions that also satisfy a matching $\Omega(\sqrt{\log \kappa})$ lower bound on their path length, as shown in Theorem 18. However in Appendix G.2 we present experimental evidence to suggest that the constant 2.5 may not be sharp. This bound fundamentally improves the $\sqrt{\kappa}$ dependence we expect via the best known bound for strongly convex functions (Section 2.2). Although we show a poly(κ) lower bound for PKL functions (Section 4.1), we are not aware of poly(κ) lower bounds for strongly convex functions. In the absence of such lower bounds, the upper bound of this section suggests a potential improvement on the path length bound for strongly convex functions as well. To this end, we make preliminary progress on a special subclass of separable strongly convex functions. First, define a class of univariate functions

$$\mathcal{G}_{\mu,L} := \{g : g \text{ is } \mu\text{-strongly convex, has } L\text{-Lipschitz gradients, and } g'''(x) \geq 0 \text{ for all } x\}.$$

Equivalently, the second derivative g'' satisfies $\mu \leq g''(x) \leq L$ for all x and g'' is non-decreasing. Using the above as building blocks, define

$$\mathcal{F}_{\mathrm{sep},\mathcal{G}_{\mu,L}} := \{f : f(\mathbf{x}) = \sum_{i=1}^{d} g_{(i)}(\mathbf{x}_{(i)}) \text{ where } g_{(i)} \in \mathcal{G}_{\mu,L}\}.$$

To motivate the above definition, consider the quadratic function f_1 given by $f_1(\mathbf{x}) = \sum_{i=1}^{d} i\mathbf{x}_{(i)}^2$ and the nearly quadratic function f_2 given by $f_2(\mathbf{x}) = \sum_{i=1}^{d}(i\mathbf{x}_{(i)}^2 + 0.1\mathbf{x}_{(i)}^4)$. f_1 has $\mu = 2$ and $L = 2d$. Note that f_2 also has $\mu = 2$, and if we restrict it to a bounded domain $[-B, B]^d$, it has $L = (2 + 1.2B^2)d$. We expect that the path length of f_2 behaves like f_1; however the only applicable path length bound we know is the $\mathcal{O}(\sqrt{\kappa})$ bound of Section 2.2. Note however that $f_2 \in \mathcal{F}_{\mathrm{sep}}$ and so Theorem 12 below shows an $\mathcal{O}(\log \kappa)$ bound on its path length.

Theorem 12 *For any $f \in \mathcal{F}_{\mathrm{sep},\mathcal{G}_{\mu,L}}$, the GF path length is bounded as:*

$$\zeta \leq (2 + \log \kappa) \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*).$$

The proof of this theorem can be found in Appendix C.

3. Dimension Dependent Path Length Bounds Under Convexity

If our function class does not exhibit linear convergence, path length bounds can still be provided that depend on the dimension d. In this section, we analyze path lengths of GD and GF under convexity. In fact, the results of this section hold for GF under the weaker assumption of quasiconvexity (Definition 2). Under quasiconvexity or convexity, even finiteness of path length is a surprising result since there exist planar convex functions whose GF curves spiral around infinitely many times while going arbitrarily close to the minimum (Daniilidis et al., 2010). In other words, GF exhibits convergence but not the stronger notion of tangential convergence (Bolte and Pauwels, 2020, Section 5.5). Since there is no natural notion of a condition number here, we look for bounds that depend on the dimension d. The analysis of path lengths of GF and GD in the convex case goes via a reduction to the notion of self-contracted curves.

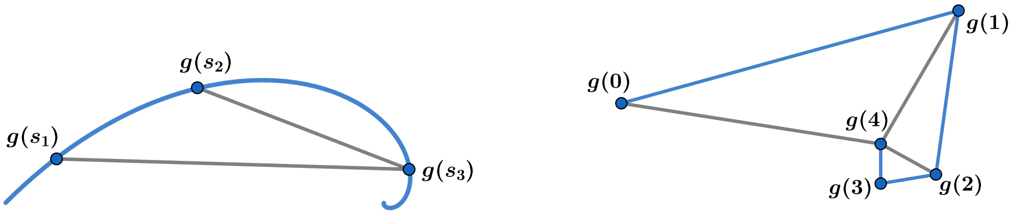


Figure 3: Self-contracted curves. In the continuous case (left), $S = \mathbb{R}_0^+$; in the discrete case (right), $S = \mathbb{N}_0$.

Definition 13 (Self-contracted curve (Daniilidis et al., 2010)) *A curve $g : S \to \mathbb{R}^d$ is self-contracted if for all $s_1, s_2, s_3 \in S$ such that $s_1 \leq s_2 \leq s_3$,*

$$\|g(s_3) - g(s_2)\|_2 \leq \|g(s_3) - g(s_1)\|_2 . \tag{12}$$

Setting s_3 such that $g(s_3) = \mathbf{x}^*$, we can see that self-contracted curves are descent curves, in the sense that consecutive iterates cannot go farther away from $\mathbf{x}^*$. However self-contracted curves require the descent condition (12) to hold more generally for any $s_1 \leq s_2 \leq s_3$. Figure 3 illustrates a self-contracted curve in two dimensions.

It is well known that the GF curve is a self-contracted curve (Manselli and Pucci, 1991; Daniilidis et al., 2015, 2010) for quasiconvex functions. To see this, first note that for any t, $\frac{df(\mathbf{x}_t)}{dt} = - \|\nabla f(\mathbf{x}_t)\|_2^2 \leq 0$, and thus for any $s \geq t$, $f(\mathbf{x}_s) \leq f(\mathbf{x}_t)$. Now, fix s and define the potential function $\varepsilon(t) = \|\mathbf{x}_s - \mathbf{x}_t\|_2^2$ for $t \leq s$. Then,

$$\dot{\varepsilon}(t) = 2 \langle \mathbf{x}_t - \mathbf{x}_s, \dot{\mathbf{x}}_t \rangle = 2 \langle \mathbf{x}_t - \mathbf{x}_s, -\nabla f(\mathbf{x}_t) \rangle = 2 \langle \nabla f(\mathbf{x}_t), \mathbf{x}_s - \mathbf{x}_t \rangle \leq \ 0,$$

where the inequality follows by quasiconvexity since $f(\mathbf{x}_s) \leq f(\mathbf{x}_t)$. Thus, $\|\mathbf{x}_s - \mathbf{x}_t\|_2^2$ is a decreasing function of t which is the same as self-contractedness for $s_3 = s$.

For GD, we prove self-contractedness under the additional assumptions of convexity and Lipschitz gradients:

Lemma 14 *For any convex function f with L-Lipschitz gradients, the GD curve with $\eta \leq 1/L$ is self-contracted.*

The proof of Lemma 14 is in Appendix D.1. We do not know if one can relax convexity to quasiconvexity. Also note the step-size restriction: for convex functions with L-Lipschitz gradients, GD with $\eta \in (0, 2/L]$ is a descent method; however for GD to be self-contracted, further restriction on step-size (such as $\eta \in (0, 1/L]$) is needed. This is discussed in Appendix D.1 after the proof of the lemma. Thus while self-contracted curves are descent curves (in terms of the iterates), there exist descent curves that are not self-contracted.

The GD curve in Lemma 14 refers to the iterates $g(0), g(1), \ldots,$ and not the affine extension of the iterates (obtained by connecting consecutive iterates by a line). It is unclear whether the affine extension itself is self-contracted. This precludes a direct application of path length bounds known for GF curves (Manselli and Pucci, 1991; Daniilidis et al., 2015) since these bounds require all points to be part of a self-contracted curve. Despite this limitation, we show that the self-contractedness guarantee provided by Lemma 14 is enough to show a path length bound for GD.

Theorem 15 *For any quasiconvex function f, the GF path length is bounded as:*

$$\zeta \leq 2^{2d\log d} \left\|\mathbf{x}_0 - \mathbf{x}_\infty\right\|_2 .$$

If f is convex with L-Lipschitz gradients, then the GD iterates with a step-size $\eta \leq 1/L$ admit a path length bound:

$$\zeta_\eta \leq 2^{10d^2} \left\|\mathbf{x}_0 - \mathbf{x}_\infty\right\|_2 ,$$

while the GD iterates with a step-size $\eta \leq 1/2L\sqrt{d}$ admit a path length bound:

$$\zeta_\eta \leq 2^{4d\log d} \left\|\mathbf{x}_0 - \mathbf{x}_\infty\right\|_2 .$$

The proof of Theorem 15 can be found in Appendix D. The GF bound is due to Manselli and Pucci (1991), while our contribution is the analysis for GD curves (nevertheless, we include the GF proof for completeness). To the best of our knowledge, ours is the only path length bound known for GD curves of convex LG functions (without further assumptions). However the step-size restriction of $\eta \leq 1/2L\sqrt{d}$ for the $2^{\mathcal{O}(d\log d)}$ result is much smaller than the usual step-sizes required for convergence to hold. It would be interesting to study if the $2^{\mathcal{O}(d\log d)}$ can be obtained with $\eta = \mathcal{O}(1/L)$.

Observe that this bound is with respect to $\left\|\mathbf{x}_0 - \mathbf{x}_\infty\right\|_2$ instead of $\operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*)$. For convex functions f with L-Lipschitz gradients, it is known that GD or GF both converge to a point in the optimal set, that is, $\mathbf{x}_\infty \in \mathbf{X}^*$. Thus if $\mathbf{X}^*$ is singleton, $\mathbf{x}_\infty = \mathbf{x}^*$. However in general, $\mathbf{x}_\infty$ may be distinct from $\Pi_{\mathbf{X}^*}(\mathbf{x}_0)$ and $\left\|\mathbf{x}_0 - \mathbf{x}_\infty\right\|_2$ may be larger than $\operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*)$.

The exponential bound of Theorem 15 can be significantly improved if the quasiconvex function is separable, that is it exhibits the decomposition

$$f(\mathbf{x}) = \sum_{i=1}^{d} g_{(i)}(\mathbf{x}_{(i)}), \tag{13}$$

for some functions $g_{(i)} : \mathbb{R} \to \mathbb{R}$. Note that if f is quasiconvex, each $g_{(i)}$ is quasiconvex.

Theorem 16 *Suppose f is quasiconvex and exhibits the decomposition* (13)*. Then the path length of GF is bounded as $\zeta \leq \sqrt{d}\ \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*)$. If f has L-Lipschitz gradients then the path length of GD with $\eta \leq 1/L$ also satisfies $\zeta_\eta \leq \sqrt{d}\ \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*)$.*

The proof of this theorem can be found in Appendix D. The decomposition (13) ensures that GD/GF always follows a descend direction in each of the components. The theorem generalizes easily to a larger class of functions that exhibit component-wise descent for any orthogonal basis (and not necessarily the canonical basis). It would be interesting to study if this can be shown for some standard class of functions larger than separable quasiconvex functions. Theorem 18 shows a matching $\Omega(\sqrt{d})$ lower bound for separable quasiconvex functions.

4. Lower Bounds

In this section we provide lower bounds on the path length for quadratic functions, PKL functions, and separable quasiconvex functions. In each case, given problem parameters (d, κ), we construct a worst-case lower bound—that is we exhibit a function f that satisfies the problem parameters and specify an initial point $\mathbf{x}_0$ for which the path length is lower bounded by some function of d and κ, times the length of the shortest path $\operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*)$.

4.1 An $\widetilde{\Omega}(\sqrt{d} \wedge \kappa^{1/4})$ Lower Bound for PKL Functions

In Section 2.3 we obtained an $\mathcal{O}(\sqrt{\log \kappa})$ dependence for the path length of quadratics objectives. Thus, a natural question is whether the $\mathcal{O}(\sqrt{\kappa})$ bound for the path length of PKL objectives can be improved to $\mathcal{O}(\text{polylog}(\kappa))$. In this section, we show that such a dependence is precluded for PKL functions without further assumptions. Previously, Oymak and Soltanolkotabi (2019, Theorem 5.4) have presented a lower bound in terms of $f(\mathbf{x}_0) - f^*$. However, this bound when translated in terms of $\operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*)$ leads to a trivial result. Theorem 17 is the first non-trivial lower bound for functions that satisfy PKL. The constructed function also satisfies linear convergence, leading to a lower bound for linearly convergent functions as well. Let $\mathcal{F}_\kappa$ be the class of real-valued functions on $\mathbb{R}^d$ such that every $f \in \mathcal{F}_\kappa$ satisfies:

- f is continuously differentiable.
- There exist constants $\mu, L > 0$ such that $\kappa \geq L/\mu$ and a) f has L-Lipschitz gradients, b) f satisfies the PKL inequality with constant μ.

Theorem 17 *For every $d \geq 6$ and $\kappa \geq 216$, there exists a function $f \in \mathcal{F}_\kappa$ and an initial point $\mathbf{x}_0$ such that the GF dynamics on f with the initial point $\mathbf{x}_0$ satisfies*

$$\zeta \geq \min\left\{\frac{\sqrt{d}}{6\log d}, \frac{\kappa^{1/4}}{6\log \kappa}\right\} \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*).$$

Similarly, there exists a function $f \in \mathcal{F}_\kappa$, an initial point $\mathbf{x}_0$, and some step-size $\eta \in [1/2L, 1/L]$ such that the GD iterates on f with the initial point $\mathbf{x}_0$ satisfy

$$\zeta_\eta \geq \min\left\{\frac{\sqrt{d}}{16\log d}, \frac{\kappa^{1/4}}{16\log \kappa}\right\} \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*).$$

The same construction guarantees the following: for any $c \in (0, 5.8 \cdot 10^{-3})$, there exists a function $f \in \mathcal{F}_\kappa$ and a point $\mathbf{x}_0$, such that f satisfies $(1, c)$-linear convergence (for GF and GD) and the path length with the initial point $\mathbf{x}_0$ can be bounded as

$$\zeta \geq \left(\frac{\sqrt{1/c}}{12\log^{1.5}(1/c)}\right) \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*); \quad \zeta_\eta \geq \left(\frac{\sqrt{1/c}}{64\log^{1.5}(1/c)}\right) \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*),$$

for GF and GD respectively.

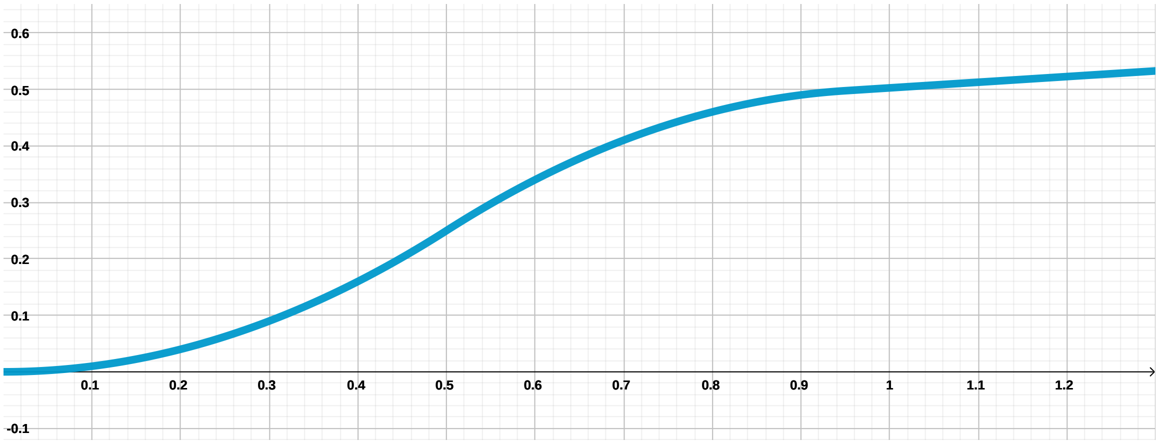


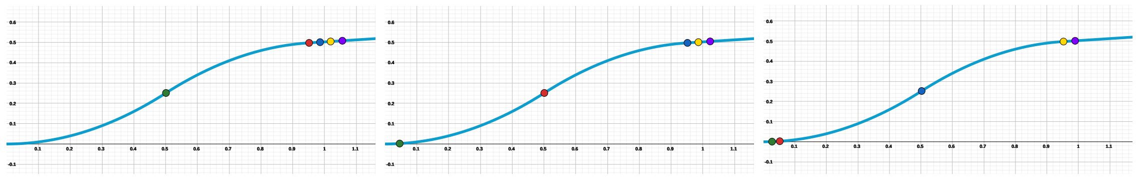
Figure 4: Component function g for PKL lower bound.

Proof sketch The function f that we construct decomposes as $f(\mathbf{x}) = \sum_{i=1}^{d} g(\mathbf{x}_{(i)})$, where the function g (Figure 4) is L-Lipschitz and μ-PKL (and thus so is f) with $\kappa \geq L/\mu$. g is designed so that it is equal to x^2 in the interval $[0, 0.5]$, so that it is strongly convex in that region. In $[0.5, 1]$, g is not strongly convex (or convex) and in some sense tapers off. However, g continues to maintain the PKL curvature condition with some constant μ globally. Next we stagger the components of the initial point $\mathbf{x}_0$ so that at every consecutive time interval, a single component starts has value 0.5 at the beginning of the time interval and decreases to almost 0.0 at the end of the time interval (Figure 5). In this way, at every time interval, a single additional component is captured. Loosely speaking, GD while optimizing approximately follows the edges of a cube instead of the diagonal. This ensures that the path length is a factor $\approx 0.5\sqrt{d}$ larger than the shortest path. Then, we compute κ and relate it to d to obtain the final bound. Similarly, we show that the function satisfies $(1, c)$-linear convergence and we relate the linear convergence constant c to the dimension d to obtain the linear convergence result. See Appendix E for details. ■

These lower bounds do not match the $O(\sqrt{\kappa})$ upper bound for PKL functions and the $O(1/c)$ upper bound for linearly convergent functions. We do not know which of these bounds are tight. In Appendix G.1 we simulate the lower bound constructed in the proof of Theorem 18 with GD. This simulation allows us to verify that the dependence of the path length of the lower bound construction is indeed $\Omega(\kappa^{1/4}/\log\kappa)$ (and not something larger like $\Omega(\sqrt{\kappa})$). We observed that the dependence is $\approx 3\kappa^{1/4}/\log\kappa$. Thus the dependence on κ is correct, but the constants are loose. Observe that the function constructed above is not strongly convex. Thus proving a poly(κ) lower bound for strongly convex functions remains an open problem.

4.2 An $\Omega(\sqrt{d} \wedge \sqrt{\log\kappa})$ Lower Bound for Quadratics

In this section, we show that the upper bound for quadratic objectives proved in Section 2.3 is tight by constructing an instance of GD and GF where the path length is $\Omega(\sqrt{\log\kappa})$, if the dimension can be set arbitrarily (this is the 'large-scale' optimization assumption made in



(a) $k = 0$

(b) $k = 1$

(c) $k = 2$

Figure 5: Illustration of lower bound path length construction of Theorem 17 for GD. Every colored circle denotes the value of $\mathbf{x}_k$ in a different component. In iterate $k = 1$, the green component goes from 0.5 to almost 0.0. Then in iterate $k = 2$ the red components decrease in the same manner and so on.

many lower bounds; for instance, see Guzmán and Nemirovski (2015) and Nesterov (2013, Theorem 2.1.7). Let $\mathcal{Q}_\kappa$ be the class of quadratic functions on $\mathbb{R}^d$ such that the Hessian has non-negative singular values and the ratio of the largest and smallest non-zero singular values is at most κ.

Theorem 18 *For every $\kappa \geq 5$, there exists a quadratic function $f \in \mathcal{Q}_\kappa$ and an initial point $\mathbf{x}_0$ such that the GF dynamics on f with the initial point $\mathbf{x}_0$ satisfies*

$$\zeta \geq \min\left\{0.7\sqrt{d}, 0.45\sqrt{\log\kappa}\right\}\ \operatorname{dist}\left(\mathbf{x}_0, \mathbf{X}^*\right). \tag{14}$$

Similarly, there exists a function $f \in \mathcal{Q}_\kappa$ and an initial point $\mathbf{x}_0$, such that for step-size $\eta = {}^1\!/_{2L}$ the GD iterates on f with the initial point $\mathbf{x}_0$ satisfy

$$\zeta_\eta \geq \min\left\{0.5\sqrt{d}, 0.3\sqrt{\log\kappa}\right\}\ \operatorname{dist}\left(\mathbf{x}_0, \mathbf{X}^*\right). \tag{15}$$

The quadratic functions constructed are separable (see (13)) and convex. Thus the $\Omega(\sqrt{d})$ bound holds for separable quasiconvex functions and convex functions as well.

Proof sketch The quadratic function we consider has geometrically increasing spectra—the eigenvalues are $1, \omega, \omega^2, \ldots \omega^{d-1}$. Similar to the proof of Theorem 17 for the PKL case, in each time interval (or iterate), a single component is captured. This leads to a lower bound of $\Omega(\sqrt{d})$. We then relate $\sqrt{d} = \Theta(\sqrt{\log\kappa})$ to write the final bound. Details can be found in Appendix F. ■

In Appendix G.2 we simulate the quadratic lower bound constructed in the proof of Theorem 18 and compare the lower bound and upper bound (Theorem 10) with the empirically observed path length.

5. Discussion

Bounds on the path length of GD (and related algorithms like SGD) have implicitly been studied in some recent papers seeking to understand optimization for deep neural networks.

In this paper, we provide unified results for GD and GF under common smoothness and curvature assumptions, obtaining the tightest known bounds for quadratics and convex objectives. We also presented a meta-theorem that gives us a path length bound for any linearly convergent iterative algorithm. To complement these results, for PKL objectives, we also show a lower bound, which to our knowledge is the first such lower bound for path lengths. For quadratics, we give a matching lower bound thus completely characterizing path lengths in this setting. For separable quasiconvex objectives, we give matching (up to constants) upper and lower bounds on the path length.

Our meta-theorem suggests that path lengths are intricately tied with convergence properties. While we focused on GD and GF, it would be interesting to consider the path length bounds exhibited by other optimization algorithms that have good convergence properties, such as SGD, projected gradient descent, accelerated methods, heavy ball methods, second order methods, proximal methods, and so on. In Appendix A, we show some preliminary results to this end. Attouch and Peypouquet (2016, Fact 4) showed a path length bound style result for Nesterov's acceleration method.

A broader open direction concerns understanding better the statistical implications of our path length bounds. As an example, it is known that stable optimization algorithms can exhibit better generalization guarantees (Bousquet and Elisseeff, 2002; Chen et al., 2018; Hardt et al., 2016) and it is natural to expect similar qualitative behaviour from optimization algorithms that have short path lengths. Some recent works (Balakrishnan et al., 2017; Mei et al., 2018), have provided statistical guarantees for optimization-based estimators via uniform (statistical) convergence over the possible algorithm iterates, and we believe these might also be strengthened by a deeper understanding of the path followed by these algorithms.

While the focus of this paper is on path length bounds for GD and GF, we also note other problems for which path length bounds have been considered. Bounds on the ℓ_1 path length of lasso and forward stagewise regression have been studied by Hastie et al. (2007). Adaptive regret bounds for bandits have been shown that depend on the total path length of the losses at each step (Wei and Luo, 2018; Bubeck et al., 2019). Argue et al. (2019) showed a result for nested convex body chasing using a bound on the path length of self-contracted curves (Manselli and Pucci, 1991), which was originally developed for proving the convergence of GF for any quasiconvex function.

Acknowledgments

SB is grateful to Marco Molinaro for inspiring discussions. We thank Sébastien Bubeck for pointing us to some references on self-contracted curves, and Anupam Gupta for informing us about the nested convex body chasing problem that sparked our interest in path length bounds. We thank the anonymous JMLR reviewers and action editor Ambuj Tewari for their comments that helped improve an earlier version of this paper. SB and CG were supported in part by the NSF grant DMS-1713003. All figures in the paper were created using GeoGebra (Hohenwarter et al., 2013).

Appendix

Table of Contents

A Proofs and Additional Results from Section 2.1 **18**
- A.1 Generalized Version of Theorem 6 19
- A.2 Generalized Version of Theorem 8 20
- A.3 Path Length Bound for Polyak's Heavy Ball Method 22
- A.4 Path Length Bound for Projected Gradient Descent 22

B Proof of Theorem 9 **24**

C Proofs of Results in Section 2.3 **26**
- C.1 Proof of Theorem 10 26
- C.2 Proof of Theorem 12 31

D Proofs of Results in Section 3 **33**
- D.1 Proof of Lemma 14 33
- D.2 Proof of Theorem 15 35
- D.3 Proof of Theorem 16 45

E Proof of Theorem 17 **45**
- E.1 Construction of $\boldsymbol{g}$ 47
- E.2 Identification of $\mathbf{x}_0$ 48
- E.3 GF Analysis 49
- E.4 GD Analysis 52
- E.5 Lower Bound Under Linear Convergence 53

F Proof of Theorem 18 **54**
- F.1 GF Analysis 55
- F.2 GD Analysis 56

G Lower Bound Simulations **57**
- G.1 PKL 57
- G.2 Quadratics 57

Appendix A. Proofs and Additional Results from Section 2.1

In this section, we state and prove more general versions of Theorem 6 and Theorem 8. These provide a path length bound if linear convergence can be established with respect to a local minimum rather than the global minimum. We also show path length bounds for Polyak's heavy ball method and projected gradient descent.

A.1 Generalized Version of Theorem 6

Theorem 6 can be generalized for linear convergence to any set $\widehat{\mathbf{X}}$ as long as it is stationary, defined next.

Definition 19 (Stationary convex set) *A stationary convex set is a convex set $\widehat{\mathbf{X}}$ such that for every $\widehat{\mathbf{x}} \in \widehat{\mathbf{X}}$, $\nabla f(\widehat{\mathbf{x}}) = 0$.*

Given any convex set $\widehat{\mathbf{X}}$, we can generalize Definition 5 to allow linear convergence to this set instead of the globally optimal set $\mathbf{X}^*$. Consider the following modification of condition (2) replacing $\mathbf{X}^*$ with $\widehat{\mathbf{X}}$:

$$\operatorname{dist}\left(\mathbf{x}_s, \widehat{\mathbf{X}}\right) \leq A(1-c)^s \operatorname{dist}\left(\mathbf{x}, \widehat{\mathbf{X}}\right).$$

Here s may belong to $\mathbb{N}_0$ (discrete-time) or $\mathbb{R}_0$ (continuous-time).

Theorem 20 *Suppose f has L-Lipschitz gradients. If the GF dynamics for f exhibits linear convergence towards a stationary convex set $\widehat{\mathbf{X}}$ with constants (A, c), then its path length is bounded as:*

$$\zeta \leq \left(AL/\log\left(1/(1-c)\right)\right) \operatorname{dist}\left(\mathbf{x}_0, \widehat{\mathbf{X}}\right), \tag{16}$$

and if the GD iterates with step-size η exhibit linear convergence towards $\widehat{\mathbf{X}}$ with constants (A, c), then their path length is bounded as:

$$\zeta_\eta \leq (\eta A L/c) \operatorname{dist}\left(\mathbf{x}_0, \widehat{\mathbf{X}}\right). \tag{17}$$

Proof We first prove the GF bound (16). Using LG and the fact that for every $\widehat{\mathbf{x}} \in \widehat{\mathbf{X}}$, $\nabla f(\widehat{\mathbf{x}}) = 0$, we bound the path length increment at every instance t:

$$\|\dot{\mathbf{x}}_t\|_2 \, dt = \|\nabla f(\mathbf{x}_t)\|_2 \, dt = \left\|\nabla f(\mathbf{x}_t) - \nabla f(\Pi_{\widehat{\mathbf{X}}}(\mathbf{x}_t))\right\|_2 dt \leq L \operatorname{dist}\left(\mathbf{x}_t, \widehat{\mathbf{X}}\right) dt.$$

Thus,

$$\begin{aligned}
\int_0^\infty \|\dot{x}_t\|_2 \; dt &\leq \int_0^\infty AL(1-c)^t \operatorname{dist}\left(\mathbf{x}_0, \widehat{\mathbf{X}}\right) \; dt \\
&= \int_0^\infty AL e^{t\log(1-c)} \operatorname{dist}\left(\mathbf{x}_0, \widehat{\mathbf{X}}\right) \; dt \\
&= \left(\frac{AL}{\log\left(1/(1-c)\right)}\right) \operatorname{dist}\left(\mathbf{x}_0, \widehat{\mathbf{X}}\right).
\end{aligned}$$

This concludes the proof of claim (16).

For claim (17), using LG we have the following regularity condition on the distance travelled at every step:

$$\begin{aligned}
\left\|\mathbf{x} - \mathbf{x}^+\right\|_2 &= \|\mathbf{x} - (\mathbf{x} - \eta \nabla f(\mathbf{x}))\|_2 \\
&= \|\eta \nabla f(\mathbf{x})\|_2 \\
&= \eta \left\|\nabla f(\mathbf{x}) - \nabla f(\Pi_{\widehat{\mathbf{X}}}(\mathbf{x}))\right\|_2
\end{aligned}$$

$$\overset{\text{LG}}{\leq} \eta L \operatorname{dist}\left(\mathbf{x}, \widehat{\mathbf{X}}\right).$$

The equality on the second last line follows by the stationarity assumption on $\widehat{\mathbf{X}}$. Thus we have the following bound on the overall path length:

$$\begin{aligned}
\zeta_\eta &= \sum_{k=0}^{\infty} \|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 \\
&\leq \sum_{k=0}^{\infty} \eta L \operatorname{dist}\left(\mathbf{x}, \widehat{\mathbf{X}}\right) \\
&\leq \sum_{k=0}^{\infty} \eta AL \ (1-c)^k \operatorname{dist}\left(\mathbf{x}_0, \widehat{\mathbf{X}}\right) \\
&= \left(\frac{\eta AL}{c}\right) \operatorname{dist}\left(\mathbf{x}_0, \widehat{\mathbf{X}}\right),
\end{aligned}$$

completing the proof.

■

Next, we show a similar generalization of Theorem 8 for linear convergence to a local minimum $\widehat{\mathbf{X}}$, but without requiring a stationarity assumption.

A.2 Generalized Version of Theorem 8

The following result applies for any iterative algorithm (not just GD) and any function class (not necessarily convex) for which linear convergence to a set $\widehat{\mathbf{X}}$ holds (i.e., Definition 5 with $\mathbf{X}^*$ replaced with $\widehat{\mathbf{X}}$). We additionally assume that descent holds with respect to all points in $\widehat{\mathbf{X}}$. We do not require $\widehat{\mathbf{X}}$ to consist of stationary points or be a convex set, as long as the conditions specified above hold.

Theorem 21 *If the iterates for a discrete update rule satisfy the following conditions:*

(a) linear convergence with constants (A, c) to a set $\widehat{\mathbf{X}}$, and

(b) descent towards all points in $\widehat{\mathbf{X}}$: for all $\widehat{\mathbf{x}} \in \widehat{\mathbf{X}}$, $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}^+ - \widehat{\mathbf{x}}\|_2 \leq \|\mathbf{x} - \widehat{\mathbf{x}}\|_2$,

then their path length is bounded as:

$$\zeta_\eta \leq \left(\frac{2A}{c}\right) \operatorname{dist}\left(\mathbf{x}_0, \widehat{\mathbf{X}}\right). \tag{18}$$

Proof Using triangle inequality for any two consecutive iterates $\mathbf{x} \to \mathbf{x}^+$,

$$\begin{aligned}
\left\|\mathbf{x} - \mathbf{x}^+\right\|_2 &\leq \left\|\mathbf{x} - \Pi_{\widehat{\mathbf{X}}}(\mathbf{x})\right\|_2 + \left\|\mathbf{x}^+ - \Pi_{\widehat{\mathbf{X}}}(\mathbf{x})\right\|_2 \\
&\overset{\text{(i)}}{\leq} \left\|\mathbf{x} - \Pi_{\widehat{\mathbf{X}}}(\mathbf{x})\right\|_2 + \left\|\mathbf{x} - \Pi_{\widehat{\mathbf{X}}}(\mathbf{x})\right\|_2 \\
&= 2\left\|\mathbf{x} - \Pi_{\widehat{\mathbf{X}}}(\mathbf{x})\right\|_2.
\end{aligned}$$

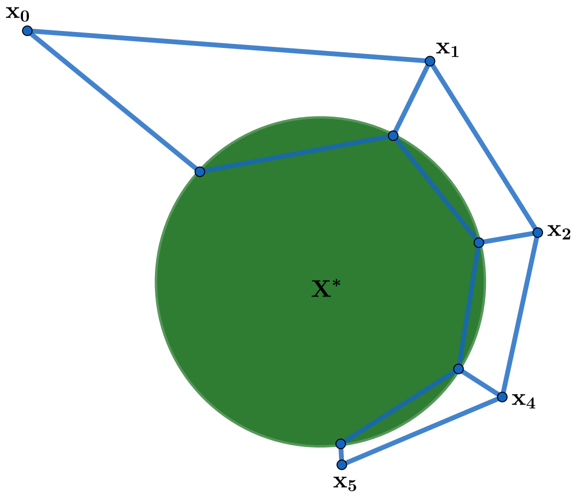


Figure 6: The iterates $\mathbf{x}_k$ converge linearly to the optimal set $\mathbf{X}^*$ but the path length is high. This situation is forbidden by condition (b) in Theorem 21.

Above, inequality (i) holds by condition (b) since $\Pi_{\widehat{\mathbf{X}}}(\mathbf{x}) \in \widehat{\mathbf{X}}$. We use this inequality for each iterate $\mathbf{x}_k$ in order to bound each term in the path length summation:

$$\begin{aligned}
\zeta_\eta &= \sum_{k=0}^{\infty} \|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 \\
&= \sum_{k=0}^{\infty} \left\|\mathbf{x}_k - (\mathbf{x}_k)^+\right\|_2 \\
&\leq 2\sum_{k=0}^{\infty} \left\|\mathbf{x}_k - \Pi_{\widehat{\mathbf{X}}}(\mathbf{x}_k)\right\|_2 \\
&\leq 2\sum_{k=0}^{\infty} A(1-c)^k \left\|\mathbf{x}_0 - \Pi_{\widehat{\mathbf{X}}}(\mathbf{x}_0)\right\|_2 && \text{(by condition (a))} \\
&= \left(\frac{2A}{c}\right) \operatorname{dist}\left(\mathbf{x}_0, \widehat{\mathbf{X}}\right),
\end{aligned}$$

as was to be shown. ■

If the convergence set $\widehat{\mathbf{X}}$ or $\mathbf{X}^*$ is singleton and $A = 1$, condition (b) of Theorem 21 is implied by condition (a). However, condition (a) is not sufficient if the convergence set is not singleton, without additional assumptions on f or the specific algorithm used. Figure 6 illustrates that there may exist algorithms that satisfy condition (a), but not condition (b), and do not exhibit short path lengths.

As a corollary to Theorem 21, we prove a path length bounds for Polyak's heavy ball method (Polyak, 1964).

A.3 Path Length Bound for Polyak's Heavy Ball Method

Given suitable α, β the Polyak's Heavy Ball (HB) method takes the following update. Below, let $\mathbf{x}$ be the current iterate (initialized at some $\mathbf{x}_0$), $\mathbf{x}^-$ be the previous iterate (initialized as $\mathbf{x}_0$), and $\mathbf{x}^+$ be the update or the next iterate.

$$\text{Polyak's heavy ball (HB):} \qquad \mathbf{x}^+ = \mathbf{x} - \alpha \nabla f(\mathbf{x}) + \beta(\mathbf{x} - \mathbf{x}^-). \tag{19}$$

Polyak (1964, Theorem 9) showed that the HB method has linear convergence for a twice continuously differentiable and μ-strongly convex function f with L-Lipschitz gradients, for the following choice of α, β:

$$\alpha = \frac{4}{\left(\sqrt{L} + \sqrt{\mu}\right)^2}, \qquad \beta = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^2. \tag{20}$$

The linear convergence parameters are given by $A = 1$ and

$$c = 1 - \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} = \frac{2}{\sqrt{\kappa} + 1}.$$

Note that a strongly convex function has a unique minimum and thus $\mathbf{X}^*$ is singleton. In this case, condition (a) of Theorem 21 with $A = 1$ implies condition (b). Thus, we have the following corollary to Theorem 21.

Corollary 22 *For any twice continuously differentiable and μ-strongly convex function f with L-Lipschitz gradients, the heavy ball method with α, β chosen according to* (20)*, has a path length bounded as:*

$$\zeta_\eta \leq \sqrt{\kappa} \left\| \mathbf{x}_0 - \mathbf{x}^* \right\|_2.$$

Proof Given the linear convergence parameters of HB: $A = 1$ and $c = 2/(\sqrt{\kappa} + 1)$, we compute

$$\frac{A + A^2(1-c)}{c} = \frac{2-c}{c} = \sqrt{\kappa}.$$

Applying Theorem 21 we obtain the claimed result. ■

Ghadimi et al. (2015) showed that the HB method also admits a linear convergence bound assuming just continuous differentiability (not twice continuous differentiability), but for conservative values of α, β and at a slower rate: $c = \mathcal{O}(1/\kappa)$. This would lead to a path length bound of $\mathcal{O}(\kappa \left\| \mathbf{x}_0 - \mathbf{x}^* \right\|_2)$.

A.4 Path Length Bound for Projected Gradient Descent

Consider a convex constraint set $\Omega \subseteq \mathbb{R}^d$ and let $\Pi_\Omega(\mathbf{x})$ denote the unique projection of a point $\mathbf{x} \in \mathbb{R}^d$ on Ω. Projected gradient descent (PGD) is an iterative optimization technique that ensures that the iterates stay within the constraint set. PGD corresponds to taking a GD step and projecting it onto Ω as follows:

$$\text{Projected gradient descent (PGD):} \qquad \mathbf{x}_{k+1} = \Pi_\Omega(\mathbf{x}_k - \eta \nabla f(\mathbf{x}_k)). \tag{21}$$

Clearly, GD is a special case of PGD with $\Omega = \mathbb{R}^d$. In this section, we provide a path length bound for PGD if f has L-Lipschitz gradients, and the iterates satisfy linear convergence. This result is closely related to the GD bound of Theorem 6. Technically, the significant difference is that in this case the stationarity assumption $\nabla f(\mathbf{x}^*) = 0$ may not be satisfied by the constrained optimal set $\mathbf{X}^*$. More work is needed to prove the result without the stationarity condition, and it leads to a larger constant (factor $(1+\sqrt{2})$ below while Theorem 6 essentially obtains the factor 1 after translation to the same setting). If stationarity does indeed hold (which is true if $\mathbf{X}^*$ is in the interior of Ω), then it can be shown that Theorem 6 applies directly and we get constant 1.

Theorem 23 *Suppose f has L-Lipschitz gradients. If the PGD iterates with step-size η exhibit linear convergence towards the optimal set $\mathbf{X}^*$ with constants (A, c), then their path length is bounded as:*

$$\zeta_\eta \leq \left(\frac{\eta L + 1}{2} + \sqrt{\eta L + \left(\frac{\eta L + 1}{2} \right)^2} \right) (A/c) \ \mathrm{dist}\left(\mathbf{x}_0, \mathbf{X}^*\right).$$

If $\eta \leq 1/L$, the above simplifies to $\zeta_\eta \leq (1+\sqrt{2})(A/c)\ \mathrm{dist}\left(\mathbf{x}_0, \mathbf{X}^*\right)$.

Proof For any $\mathbf{x} \in \Omega$, denote $\widetilde{\mathbf{x}} = \mathbf{x} - \eta \nabla f(\mathbf{x})$, and $\mathbf{x}^+ = \Pi_\Omega(\widetilde{\mathbf{x}})$ which corresponds to the PGD update for $\mathbf{x}$. Also let $\mathbf{x}^* = \Pi_{\mathbf{X}^*}(\mathbf{x})$ be the closest optimum point to $\mathbf{x}$. The next two inequalities use the following standard fact about projections to a convex set (for instance, see Bubeck (2015, Lemma 3.1)):

$$\text{for all } \mathbf{x} \in \mathbb{R}^d \text{ and } \mathbf{z} \in \Omega, \text{ the projection } \mathbf{y} = \Pi_\Omega(\mathbf{x}) \text{ satisfies } \langle \mathbf{y} - \mathbf{x}, \mathbf{z} - \mathbf{y} \rangle \geq 0.$$

Since $\mathbf{x}^+$ is the projection of $\widetilde{\mathbf{x}}$, and $\mathbf{x}^* \in \Omega$:

$$\left\langle \mathbf{x}^+ - \widetilde{\mathbf{x}}, \mathbf{x}^* - \mathbf{x}^+ \right\rangle \geq 0. \tag{22}$$

Further, by the optimality of $\mathbf{x}^*$, $\Pi_\Omega(\mathbf{x}^* - \eta \nabla f(\mathbf{x}^*)) = \mathbf{x}^*$, thus since $\mathbf{x}^+ \in \Omega$,

$$\left\langle \mathbf{x}^* - (\mathbf{x}^* - \eta \nabla f(\mathbf{x}^*)), \mathbf{x}^+ - \mathbf{x}^* \right\rangle \geq 0. \tag{23}$$

Adding (22) and (23),

$$\left\langle \mathbf{x}^+ - \widetilde{\mathbf{x}} - \eta \nabla f(\mathbf{x}^*), \mathbf{x}^* - \mathbf{x}^+ \right\rangle \geq 0.$$

Substituting $\widetilde{\mathbf{x}} = \mathbf{x} - \eta \nabla f(\mathbf{x})$ and rearranging we get,

$$\eta \left\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*), \mathbf{x}^* - \mathbf{x}^+ \right\rangle + \left\langle \mathbf{x}^+ - \mathbf{x}, \mathbf{x}^* - \mathbf{x} \right\rangle \geq \left\| \mathbf{x}^+ - \mathbf{x} \right\|_2^2.$$

Using Cauchy-Schwarz and triangle inequality, this implies

$$\eta \left\| \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*) \right\|_2 \left(\left\| \mathbf{x}^* - \mathbf{x} \right\|_2 + \left\| \mathbf{x} - \mathbf{x}^+ \right\|_2 \right) + \left\| \mathbf{x}^+ - \mathbf{x} \right\|_2 \left\| \mathbf{x}^* - \mathbf{x} \right\|_2 \geq \left\| \mathbf{x}^+ - \mathbf{x} \right\|_2^2.$$

Denote $a = \|\mathbf{x} - \mathbf{x}^+\|_2$ and $b = \|\mathbf{x} - \mathbf{x}^*\|_2$. Using LG, $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*)\|_2 \leq L\,\|\mathbf{x} - \mathbf{x}^*\|_2 = Lb$; thus the above can be re-written as

$$\eta Lb(a+b) + ab \geq a^2.$$

Completing squares,

$$\left(a - \left(\frac{\eta L + 1}{2}\right) b\right)^2 \leq \eta L b^2 + \left(\frac{\eta L + 1}{2}\right)^2 b^2,$$

which gives

$$a \leq \left(\frac{\eta L + 1}{2} + \sqrt{\eta L + \left(\frac{\eta L + 1}{2}\right)^2}\right) b. \tag{24}$$

Denote the factor multiplied to b as u so that $a \leq ub$. We now compute the path length bound by instantiating the above result for $\mathbf{x}$ given by the iterates of PGD so that for $\mathbf{x} = \mathbf{x}_k$, $a = \|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2$ and $b = \text{dist}\,(\mathbf{x}_k, \mathbf{X}^*)$:

$$\begin{aligned}
\zeta_\eta &= \sum_{k=0}^{\infty} \|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 \\
&\leq \sum_{k=0}^{\infty} u \;\text{dist}\,(\mathbf{x}_k, \mathbf{X}^*) && \text{(using (24))} \\
&\leq \sum_{k=0}^{\infty} uA\;(1-c)^k \text{dist}\,(\mathbf{x}_0, \mathbf{X}^*) && \text{(using linear convergence)} \\
&= (uA/c)\;\text{dist}\,(\mathbf{x}_0, \mathbf{X}^*)\,.
\end{aligned}$$

This completes the proof. ■

Notice that the only property of $\mathbf{X}^*$ used in the proof is stationarity with respect to the PGD updates, required in equation (23). Thus, like Theorem 20, Theorem 23 can be extended to include convergence to any locally optimal set $\widehat{\mathbf{X}}$ that is stationary with respect to the PGD updates.

Appendix B. Proof of Theorem 9

We first prove the statement for GF. Consider some t such that $\mathbf{x}_t \notin \mathbf{X}^*$, so that $f(\mathbf{x}_t) \neq f(\mathbf{x}^*)$. Note that once $\mathbf{x}_t \in \mathbf{X}^*$, $\nabla f(\mathbf{x}_\tau) = 0$ for all $\tau \geq t$, and hence the path length is 0 henceforth. Consider the following Lyapunov function:

$$\varepsilon_t = \sqrt{f(\mathbf{x}_t) - f^*}.$$

Suppose $\mathbf{x}_t \notin \mathbf{X}^*$. Taking the derivative of ε_t with respect to time, and using chain rule,

$$\dot{\varepsilon}_t = \frac{\frac{df(\mathbf{x}_t)}{dt}}{2\sqrt{f(\mathbf{x}_t) - f^*}}$$

$$= \frac{\left\langle \frac{df(\mathbf{x}_t)}{d\mathbf{x}_t}, \frac{d\mathbf{x}_t}{dt} \right\rangle}{2\sqrt{f(\mathbf{x}_t) - f^*}}$$
$$= -\frac{\|\nabla f(\mathbf{x}_t)\|_2^2}{2\sqrt{f(\mathbf{x}_t) - f^*}}$$
$$\overset{\text{PKL}}{\leq} -\sqrt{\frac{\mu}{2}}\, \|\nabla f(\mathbf{x}_t)\|_2 .$$

Although the above proof assumes $\mathbf{x}_t \notin \mathbf{X}^*$, the conclusion is true even for $\mathbf{x}_t \in \mathbf{X}^*$, since both sides are equal to 0. Using the fundamental theorem of calculus,

$$\int_0^\infty \|\nabla f(\mathbf{x}_t)\|_2 \; dt \leq -\sqrt{\frac{2}{\mu}}\, [\varepsilon_t]_0^\infty = -\sqrt{\frac{2}{\mu}} \left[\sqrt{f(\mathbf{x}_t) - f^*}\right]_0^\infty = \sqrt{\frac{2(f(\mathbf{x}_0) - f^*)}{\mu}}.$$

We can then use LG to bound the right side in terms of the distance of the point $\mathbf{x}_0$ from the optimal set:

$$f(\mathbf{x}_0) - f^* \leq \frac{L}{2}\, \|\mathbf{x}_0 - \Pi_{\mathbf{X}^*}(\mathbf{x}_0)\|_2^2 = \frac{L}{2} \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*)^2 .$$

Substituting this back in, we obtain our result:

$$\zeta = \int_0^\infty \|\nabla f(\mathbf{x}_t)\|_2 \; dt \leq \sqrt{\frac{L}{\mu}}\, \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*),$$

as was to be shown.

The GD proof technique is directly inspired by the proof of Oymak and Soltanolkotabi (2019): Equation (5.3) of Theorem 5.2 in the paper. (However, their path length bound in terms of $\operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*)$ is weaker than ours. Equation (5.5) in their paper corresponds to the bound $\frac{2L}{\mu} \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*)$ instead of $2\sqrt{\frac{L}{\mu}} \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*)$ as we show.) Like the GF case, consider some k such that $\mathbf{x}_k \notin \mathbf{X}^*$. Then by LG,

$$\begin{aligned}
\sqrt{f(\mathbf{x}_{k+1}) - f^*} &= \sqrt{f(\mathbf{x}_k - \eta \nabla f(\mathbf{x}_k)) - f^*} \\
&\leq \sqrt{f(\mathbf{x}_k) - \langle \nabla f(\mathbf{x}_k), \eta \nabla f(\mathbf{x}_k) \rangle + \frac{L}{2}\, \|\eta \nabla f(\mathbf{x}_k)\|_2^2 - f^*} \\
&= \sqrt{f(\mathbf{x}_k) - f^* - (\eta - \eta^2 L/2)\, \|\nabla f(\mathbf{x}_k)\|_2^2} \\
&\leq \sqrt{f(\mathbf{x}_k) - f^*} - \frac{(\eta - \eta^2 L/2)\, \|\nabla f(\mathbf{x}_k)\|_2^2}{2\sqrt{f(\mathbf{x}_k) - f^*}}.
\end{aligned}$$

The final inequality holds since $\sqrt{a - b} \leq \sqrt{a} - \frac{b}{2\sqrt{a}}$. To simplify the second term, note that $\eta \leq 1/L$, implies $\eta - \eta^2 L/2 \geq \eta/2$. Using this and the PKL inequality, we get

$$\frac{(\eta - \eta^2 L/2)\, \|\nabla f(\mathbf{x}_k)\|_2^2}{2\sqrt{f(\mathbf{x}_k) - f^*}} \geq (\eta/2) \cdot \frac{\|\nabla f(\mathbf{x}_k)\|_2}{2\sqrt{f(\mathbf{x}_k) - f^*}} \cdot \|\nabla f(\mathbf{x}_k)\|_2 \geq \eta \sqrt{\mu/8}\;\; \|\nabla f(\mathbf{x}_k)\|_2 .$$

Rearranging, we get

$$\sqrt{f(\mathbf{x}_{k+1}) - f^*} \leq \sqrt{f(\mathbf{x}_k) - f^*} - \eta\sqrt{\mu/8}\ \ \|\nabla f(\mathbf{x}_k)\|_2 \,.$$

The above is trivially also true if $\mathbf{x}_k \in \mathbf{X}^*$, since both sides are 0. Note that $\mathbf{x}_{k+1} - \mathbf{x}_k = -\eta\nabla f(\mathbf{x}_k)$; thus for all $k \geq 0$,

$$\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 \leq \sqrt{8/\mu}\ \ \left(\sqrt{f(\mathbf{x}_k) - f^*} - \sqrt{f(\mathbf{x}_{k+1}) - f^*}\right).$$

Telescoping this from $k = 0, \ldots \infty$,

$$\zeta_\eta = \sum_{k=0}^{\infty} \|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 \leq \sqrt{8/\mu}\left(\sqrt{f(\mathbf{x}_0) - f^*} - \sqrt{f(\mathbf{x}_\infty) - f^*}\right) \leq \sqrt{8/\mu}\left(\sqrt{f(\mathbf{x}_0) - f^*}\right).$$

As an immediate consequence of LG, we have $\sqrt{f(\mathbf{x}_0) - f^*} \leq \sqrt{\frac{L}{2}}\ \mathrm{dist}\left(\mathbf{x}_0, \mathbf{X}^*\right)$, and plugging this into the above bound yields

$$\zeta_\eta \leq 2\sqrt{\frac{L}{\mu}}\ \mathrm{dist}\left(\mathbf{x}_0, \mathbf{X}^*\right),$$

as claimed. ■

Appendix C. Proofs of Results in Section 2.3

The proofs of Theorem 10 and Theorem 12 are organized as subsections.

C.1 Proof of Theorem 10

We first write the proof in the GF case. Let α_i be the component of $(\mathbf{x}_0 \ - \ \Pi_{\mathbf{X}^*}(\mathbf{x}_0))$ in the direction of the eigenvector of Σ that corresponds to the eigenvalue σ_i. Observe that

$$\begin{aligned}\zeta &= \int_0^\infty \|\dot{\mathbf{x}}_t\|_2 \ dt \\ &= \int_0^\infty \|\exp(-t\Sigma)\Sigma(\mathbf{x}_0 \ - \ \Pi_{\mathbf{X}^*}(\mathbf{x}_0))\|_2 \ dt \\ &= \int_0^\infty \sqrt{\sum_{i=1}^{d^+} \exp\left(-2t\sigma_i\right)\sigma_i^2\alpha_i^2} \ dt.\end{aligned}$$

The $\sqrt{d^+}$ bound is straightforward. Since $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for nonnegative a and b, we have

$$\begin{aligned}\int_0^\infty \sqrt{\sum_{i=1}^{d^+} \exp\left(-2t\sigma_i\right)\sigma_i^2\alpha_i^2} \ dt &\leq \int_0^\infty \sum_{i=1}^{d^+} \sqrt{\exp\left(-2t\sigma_i\right)\sigma_i^2\alpha_i^2} \ dt \\ &= \int_0^\infty \sum_{i=1}^{d^+} \exp\left(-t\sigma_i\right)\sigma_i \left|\alpha_i\right| \ dt\end{aligned}$$

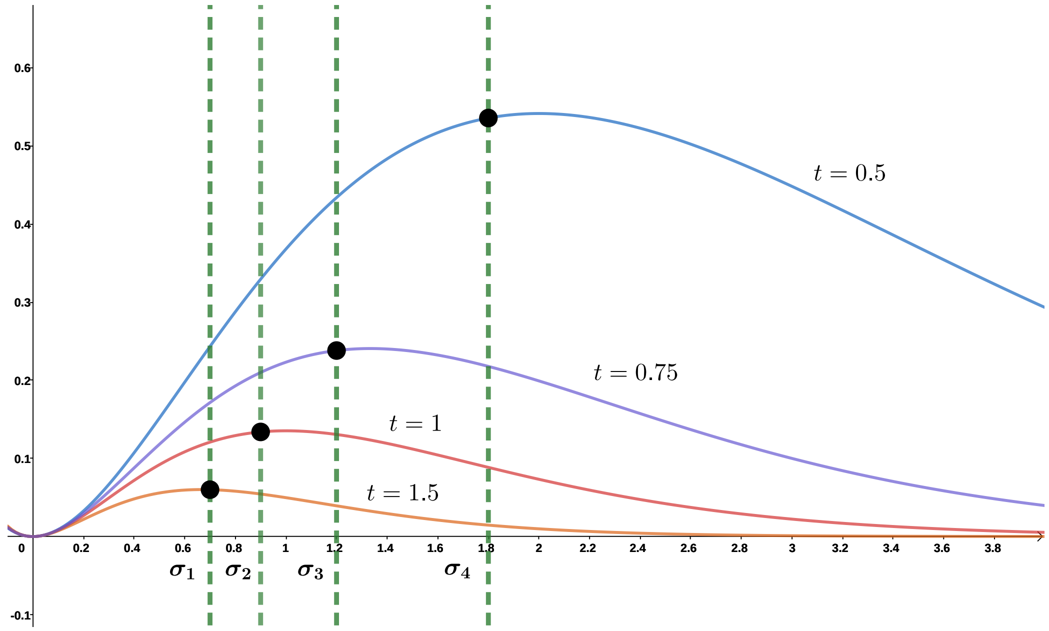


Figure 7: $g_t(x)$ for some values of t. Suggestive values of $\sigma_i = 0.7, 0.9, 1.2, 1.8$ are also shown. Observe that for every t, there is a different σ_i that maximizes g_t (indicated with large black points).

$$
\begin{aligned}
&= \sum_{i=1}^{d^+} \int_0^\infty \exp\left(-t\sigma_i\right) \sigma_i \left|\alpha_i\right| \, dt \\
&= \sum_{i=1}^{d^+} \left|\alpha_i\right| \\
&\leq \sqrt{d^+} \|\alpha\|_2 \;=\; \sqrt{d^+}\ \mathrm{dist}\left(\mathbf{x}_0, \mathbf{X}^*\right).
\end{aligned}
$$

We now prove the bound in (10) that depends on the κ_i's. For every $t \in \mathbb{R}^+$, consider a function $g_t : \mathbb{R}^+ \to \mathbb{R}$, $g_t(x) = \exp(-2tx)x^2$. For every value of t, a term in the path length integral is a linear combination of evaluations of g_t at the points $\sigma_1, \sigma_2, \ldots, \sigma_{d^+}$. We will bound each $g_t(\sigma_i)$ in the linear combination with $\max_j g_t(\sigma_j)$. Notice that for different values of t, $\arg\max_j g_t(\sigma_j)$ is different. The maximum of g_t occurs at $x_m = 1/t$, and g_t is an increasing function in x before x_m and decreasing in x after x_m. To bound the path length we use this observation to identify $\arg\max_j g_t(\sigma_j)$ for each t (see Figure 7 for reference).

$$
\zeta \;=\; \int_0^\infty \sqrt{\sum_{i=1}^{d^+} \exp\left(-2t\sigma_i\right) \sigma_i^2 \alpha_i^2}\ dt = \underbrace{\int_0^{1/\sigma_1} \sqrt{\sum_{i=1}^{d^+} \exp\left(-2t\sigma_i\right) \sigma_i^2 \alpha_i^2}\ dt}_{=:E_1}
$$

$$+ \sum_{j=1}^{d^+-1} \underbrace{\int_{1/\sigma_j}^{1/\sigma_{j+1}} \sqrt{\sum_{i=1}^{d^+} \exp\left(-2t\sigma_i\right) \sigma_i^2 \alpha_i^2 \; dt}}_{=:T_j} \;+\; \underbrace{\int_{1/\sigma_{d^+}}^{\infty} \sqrt{\sum_{i=1}^{d^+} \exp\left(-2t\sigma_i\right) \sigma_i^2 \alpha_i^2 \; dt}}_{=:E_2}.$$

For the first integral, when $t \leq 1/\sigma_1$, we have that $\sigma_i \leq \sigma_1 < 1/t$, and at this stage g_t is an increasing function of x, so every term is upper bounded by $\exp(-2t\sigma_1)\sigma_1^2$. This leads to a bound for the first term above:

$$E_1 \leq \int_0^{1/\sigma_1} \exp(-t\sigma_1)\sigma_1 \|\alpha\|_2 dt = (1 - 1/e)\|\mathbf{x}_0 \;-\; \Pi_{\mathbf{X}^*}(\mathbf{x}_0)\|_2. \tag{25}$$

Similarly, for the last integral, when $t \geq 1/\sigma_{d^+}$, we have all $\sigma_i \geq \sigma_{d^+} > 1/t$, and here g_t is a decreasing function of x, so every term is upper bounded by $\exp(-2t\sigma_{d^+})\sigma_{d^+}^2$, so the last term is upper bounded as:

$$E_2 \leq \int_{1/\sigma_{d^+}}^{\infty} \exp(-t\sigma_{d^+})\sigma_{d^+} \|\alpha\|_2 dt = (1/e)\|\mathbf{x}_0 \;-\; \Pi_{\mathbf{X}^*}(\mathbf{x}_0)\|_2. \tag{26}$$

Last, for the middle integral, consider a particular term T_j. If $\sigma_j = \sigma_{j+1}$, $T_j = 0 = \kappa_j^{-1/(\kappa_j - 1)}(1 - 1/\kappa_j)$. Else, define $t_j := \log(\kappa_j)/(\sigma_j - \sigma_{j+1})$ and observe that that $t_j \in (1/\sigma_j, 1/\sigma_{j+1})$. We can split T_j into two parts:

$$T_j = \int_{1/\sigma_j}^{t_j} \sqrt{\sum_{i=1}^{d} \exp\left(-2t\sigma_i\right) \sigma_i^2 \alpha_i^2 \; dt} + \int_{t_j}^{1/\sigma_{j+1}} \sqrt{\sum_{i=1}^{d} \exp\left(-2t\sigma_i\right) \sigma_i^2 \alpha_i^2 \; dt}.$$

Whenever $1/\sigma_j < t < 1/\sigma_{j+1}$, $\sigma_{j+1} < 1/t < \sigma_j$. Thus, for every t, the value of $\max\{g_t(\sigma_j), g_t(\sigma_{j+1})\}$ dominates every $g_t(\sigma_i)$. Which one of these two is larger depends on which side of t_j we consider. In the first term, $g_t(\sigma_j)$ dominates, and in the second $g_t(\sigma_{j+1})$ dominates. This yields an upper bound of:

$$\begin{aligned} T_j \;&\leq \int_{1/\sigma_j}^{t_j} \exp\left(-t\sigma_j\right) \sigma_j \|\alpha\|_2 \; dt + \int_{t_j}^{1/\sigma_{j+1}} \exp\left(-t\sigma_{j+1}\right) \sigma_{j+1} \|\alpha\|_2 \; dt \\ &= (\exp(-t_j\sigma_{j+1}) - \exp(-t_{j+1}\sigma_j))\|\alpha\|_2 \\ &= \kappa_j^{-1/(\kappa_j-1)}(1 - 1/\kappa_j) \left\|\mathbf{x}_0 \;-\; \Pi_{\mathbf{X}^*}(\mathbf{x}_0)\right\|_2. \end{aligned} \tag{27}$$

Summing up the bounds in Equations (25), (26), (27), we get

$$\zeta \;\leq\; \left(1 + \sum_{j=d-1}^{1} \kappa_j^{-1/(\kappa_j-1)}(1 - 1/\kappa_j)\right) \|\mathbf{x}_0 \;-\; \Pi_{\mathbf{X}^*}(\mathbf{x}_0)\|_2.$$

Next, we simplify the above expression in terms of κ. Note the following fact for all $x \geq 1$ (this can be seen graphically as shown in Figure 8):

$$x^{-1/(x-1)}(1 - 1/x) \leq \frac{\log x}{e}.$$

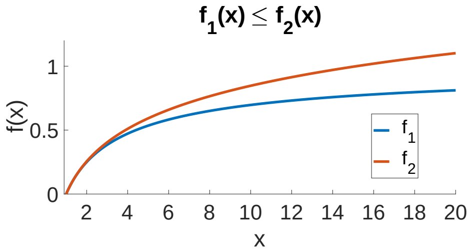


Figure 8: Graphical proof for $x^{-1/(x-1)}(1-1/x) \le \frac{\log x}{e}$. $f_1(x) = x^{-1/(x-1)}(1-1/x)$ and $f_2(x) = \frac{\log x}{e}$. Clearly $f_1(x) \le 1$ so for $x \in [20, \infty)$ the inequality is trivial.

Thus

$$\sum_{j=1}^{d^+-1} \kappa_j^{-1/(\kappa_j-1)}(1-1/\kappa_j) \le \sum_{j=1}^{d^+-1} \frac{\log \kappa_j}{e} = \frac{\log \kappa}{e},$$

which gives an $\mathcal{O}(\log \kappa)$ bound. For the $\mathcal{O}(\sqrt{\log \kappa})$ bound, we first split ζ as before:

$$\begin{aligned}\zeta &= E_1 + \left(\int_{1/\sigma_1}^{1/\sigma_{d^+}} \sqrt{\sum_{i=1}^{d^+} \exp\left(-2t\sigma_i\right) \sigma_i^2 \alpha_i^2 \; dt} \right) + E_2 \\ &\le 1 + \int_{1/\sigma_1}^{1/\sigma_{d^+}} \sqrt{\sum_{i=1}^{d^+} \exp\left(-2t\sigma_i\right) \sigma_i^2 \alpha_i^2} \; dt.\end{aligned}$$

Now we bound the second summation. Fix a constant $b > 1$, to be specified later. Let $r = \lceil \log_b \kappa \rceil = \lceil \log_b(\sigma_1/\sigma_{d^+}) \rceil$ and consider the r intervals given by $I_k = [b^{k-1}\sigma_{d^+}, b^k \sigma_{d^+})$ for $k \in [r-1]$ and $I_r = [b^{r-1}\sigma_{d^+}, b^r \sigma_{d^+}]$. If $\sigma_i \in I_k$, that is if $b^k \sigma_{d^+} \ge \sigma_i \ge b^{k-1}\sigma_{d^+}$ then we have for any t

$$\exp\left(-2t\sigma_i\right) \sigma_i^2 \le \exp\left(-2t(b^{k-1}\sigma_{d^+})\right) b^{2k} \sigma_{d^+}^2.$$

Define $\theta_k := \sqrt{\sum_{i:\sigma_i \in I_k} \alpha_i^2}$ and note that $\sqrt{\sum_{k=1}^r \theta_k^2} = \|\mathbf{x}_0 \; - \; \Pi_{\mathbf{X}^*}(\mathbf{x}_0)\|_2$. Then

$$\begin{aligned}\int_{1/\sigma_1}^{1/\sigma_{d^+}} \sqrt{\sum_{i=1}^{d^+} \exp\left(-2t\sigma_i\right) \sigma_i^2 \alpha_i^2 \; dt} &= \int_{1/\sigma_1}^{1/\sigma_{d^+}} \sqrt{\sum_{k=1}^{r} \sum_{i:\sigma_i \in I_k} \exp\left(-2t\sigma_i\right) \sigma_i^2 \alpha_i^2 \; dt} \\ &\le \int_{1/\sigma_1}^{1/\sigma_{d^+}} \sqrt{\sum_{k=1}^{r} \sum_{i:\sigma_i \in I_k} \exp\left(-2t(b^{k-1}\sigma_{d^+})\right) b^{2k} \sigma_{d^+}^2 \alpha_i^2 \; dt} \\ &\le \int_{1/\sigma_1}^{1/\sigma_{d^+}} \left(\sum_{k=1}^{r} \sqrt{\sum_{i:\sigma_i \in I_k} \exp\left(-2t(b^{k-1}\sigma_{d^+})\right) b^{2k} \sigma_{d^+}^2 \alpha_i^2} \right) \; dt \\ &= \int_{1/\sigma_1}^{1/\sigma_{d^+}} \left(\sum_{k=1}^{r} \theta_k \exp\left(-t(b^{k-1}\sigma_{d^+})\right) b^k \sigma_{d^+} \right) \; dt\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{r} \theta_k \int_{1/\sigma_1}^{1/\sigma_{d^+}} \exp\left(-t(b^{k-1}\sigma_{d^+})\right) b^k \sigma_{d^+}\ dt \\
&= b \sum_{k=1}^{r} \theta_k \left(\exp\left(-(b^{k-1}\sigma_{d^+})/\sigma_1\right) - \exp\left(-(b^{k-1}\sigma_{d^+})/\sigma_{d^+})\right)\right) \\
&\le b \sum_{k=1}^{r} \theta_k \\
&\le b\sqrt{r} \sqrt{\sum_{k=1}^{r} \theta_k^2} \\
&= b\sqrt{\log_b \kappa}\, \|\mathbf{x}_0 \ - \ \Pi_{\mathbf{X}^*}(\mathbf{x}_0)\|_2 \\
&= b\sqrt{\log_b e}\ \sqrt{\log \kappa}\, \|\mathbf{x}_0 \ - \ \Pi_{\mathbf{X}^*}(\mathbf{x}_0)\|_2 \,.
\end{aligned}
$$

Setting $b = 2$, which is close to the minima of $b\sqrt{\log_b e}$, and bounding $2\sqrt{\log_2 e} \le 2.5$ gives the result. This concludes the proof in the GF case.

For GD, since $\mathbf{A}^T y = (\mathbf{A}^T\mathbf{A})\Pi_{\mathbf{X}^*}(\mathbf{x}_0)$, Equation (9) leads to the following recurrence for $k \ge 1$:

$$
\mathbf{x}_k - \Pi_{\mathbf{X}^*}(\mathbf{x}_0) = \mathbf{x}_{k-1} - \Pi_{\mathbf{X}^*}(\mathbf{x}_0) - \eta\left(\frac{\mathbf{A}^T\mathbf{A}}{n}\right)(\mathbf{x}_{k-1} - \Pi_{\mathbf{X}^*}(\mathbf{x}_0)) = (I - \eta\Sigma)\,(\mathbf{x}_{k-1} - \Pi_{\mathbf{X}^*}(\mathbf{x}_0)).
$$

Thus for $k \in \mathbb{N}_0$,

$$
(\mathbf{x}_k - \Pi_{\mathbf{X}^*}(\mathbf{x}_0)) = (I - \eta\Sigma)^k\,(\mathbf{x}_0 - \Pi_{\mathbf{X}^*}(\mathbf{x}_0))
$$

Then we can compute the path length as follows:

$$
\begin{aligned}
\text{For } k \ge 1,\ \|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2 &= \|(\mathbf{x}_k - \Pi_{\mathbf{X}^*}(\mathbf{x}_0)) - (\mathbf{x}_{k-1} - \Pi_{\mathbf{X}^*}(\mathbf{x}_0))\|_2 \\
&= \left\|(I - \eta\Sigma)^{k-1}\,(\eta\Sigma)\,(\mathbf{x}_0 - \Pi_{\mathbf{X}^*}(\mathbf{x}_0))\right\|_2 .
\end{aligned}
$$

Taking a sum over all iterates from $k = 0$ to ∞, we obtain the bound

$$
\zeta_\eta = \sum_{k=0}^{\infty} \left\|(I - \eta\Sigma)^k\,(\eta\Sigma)\,(\mathbf{x}_0 - \Pi_{\mathbf{X}^*}(\mathbf{x}_0))\right\|_2 .
$$

Note that η is such that the singular values of $\eta\Sigma$ are $1 \ge \sigma_1' \ge \sigma_2' \cdots \ge \sigma_p' > 0$. For every $j \in [d^+ - 1]$, we have $\kappa_j = \sigma_j'/\sigma_{j+1}'$. Also observe that α_i is the component of $(\mathbf{x}_0 - \Pi_{\mathbf{X}^*}(\mathbf{x}_0))$ in the direction of the eigenvector of $\eta\Sigma$ that corresponds to the eigenvalue σ_i'. Thus,

$$
\begin{aligned}
\zeta_\eta &= \sum_{k=0}^{\infty} \sqrt{\sum_{i=1}^{d^+} (1 - \sigma_i')^{2k} \sigma_i'^2 \alpha_i^2} \\
&\le \sum_{k=0}^{\infty} \sqrt{\sum_{i=1}^{d^+} \exp\left(-2k\sigma_i'\right) \sigma_i'^2 \alpha_i^2}
\end{aligned}
$$

$$
\begin{aligned}
&\le \ \|\mathbf{x}_0 \ - \ \Pi_{\mathbf{X}^*}(\mathbf{x}_0)\|_2 + \sum_{k=1}^{\infty} \sqrt{\sum_{i=1}^{d^+} \exp\left(-2k\sigma_i'\right) \sigma_i'^2 \alpha_i^2} \\
&\le \ \|\mathbf{x}_0 \ - \ \Pi_{\mathbf{X}^*}(\mathbf{x}_0)\|_2 + \int_0^{\infty} \sqrt{\sum_{i=1}^{d^+} \exp\left(-2t\sigma_i'\right) \sigma_i'^2 \alpha_i^2 \ dt} \\
&= \ \|\mathbf{x}_0 \ - \ \Pi_{\mathbf{X}^*}(\mathbf{x}_0)\|_2 + \int_0^{\infty} \sqrt{\sum_{i=1}^{d^+} \exp\left(-2t\sigma_i\right) \sigma_i^2 \alpha_i^2 \ dt} \qquad \text{(reparameterizing } t \to \eta t\text{)} \\
&= \ \|\mathbf{x}_0 \ - \ \Pi_{\mathbf{X}^*}(\mathbf{x}_0)\|_2 + \zeta,
\end{aligned}
$$

as was to be shown. ■

C.2 Proof of Theorem 12

In each component i, let $\mathbf{x}^*_{(i)}$ denote the unique minimum. We will consider GF on f with some initial point $\mathbf{x}_0$. For every index i consider the following potential function for any time t such that $(\mathbf{x}_t)_{(i)} \neq \mathbf{x}^*_{(i)}$:

$$\phi(t) = \frac{g'_{(i)}((\mathbf{x}_t)_{(i)})}{(\mathbf{x}_t)_{(i)} - \mathbf{x}^*_{(i)}}.$$

First note that by convexity, $\phi(t) \geq 0$. We will show that $\phi(t)$ is decreasing in t.

$$
\begin{aligned}
\phi'(t) &= \frac{d\phi(t)}{d\mathbf{x}_{(i)}} \cdot \frac{d\mathbf{x}_{(i)}}{dt} \\
&= \left(\frac{g''_{(i)}((\mathbf{x}_t)_{(i)})}{(\mathbf{x}_t)_{(i)} - \mathbf{x}^*_{(i)}} - \frac{g'_{(i)}((\mathbf{x}_t)_{(i)})}{((\mathbf{x}_t)_{(i)} - \mathbf{x}^*_{(i)})^2} \right) \left(-g'_{(i)}(\mathbf{x}_{(i)}) \right).
\end{aligned}
$$

Suppose $((\mathbf{x}_t)_{(i)} - \mathbf{x}^*_{(i)}) \geq 0$. Then by convexity, $g'_{(i)}(\mathbf{x}_{(i)}) \geq 0$. Now observe that

$$
\begin{aligned}
g'_{(i)}((\mathbf{x}_t)_{(i)}) &= \int_{\mathbf{x}^*_{(i)}}^{(\mathbf{x}_t)_{(i)}} g''_{(i)}(x) dx \\
&\leq \int_{\mathbf{x}^*_{(i)}}^{(\mathbf{x}_t)_{(i)}} g''_{(i)}((\mathbf{x}_t)_{(i)}) dx \qquad \text{(since } g'' \text{ is assumed to be non-decreasing)} \\
&= g''_{(i)}((\mathbf{x}_t)_{(i)})((\mathbf{x}_t)_{(i)} - \mathbf{x}^*_{(i)}).
\end{aligned}
$$

Thus $\phi'(t) \leq 0$. Alternately, suppose $((\mathbf{x}_t)_{(i)} - \mathbf{x}^*_{(i)}) \leq 0$. Then by convexity, $g'_{(i)}(\mathbf{x}_{(i)}) \leq 0$. Now observe that

$$
\begin{aligned}
g'_{(i)}((\mathbf{x}_t)_{(i)}) &= \int_{\mathbf{x}^*_{(i)}}^{(\mathbf{x}_t)_{(i)}} g''_{(i)}(x) dx \\
&\geq \int_{\mathbf{x}^*_{(i)}}^{(\mathbf{x}_t)_{(i)}} g''_{(i)}((\mathbf{x}_t)_{(i)}) dx \qquad \text{(since } g'' \text{ is assumed to be non-decreasing)}
\end{aligned}
$$

$$= g''_{(i)}((\mathbf{x}_t)_{(i)})((\mathbf{x}_t)_{(i)} - \mathbf{x}^*_{(i)}).$$

In this case too we observe that $\phi'(t) \leq 0$. Thus, for every t such that $(\mathbf{x}_t)_{(i)} \neq \mathbf{x}^*_{(i)}$, $\phi'(t) \leq 0$. Also, since g_i is μ-strongly convex and has L-Lipschitz gradients $\phi(t) \in [\mu, L]$. Suppose $\phi(t) = c \geq \mu$. Then for every $s \leq t$, $\phi(t) \geq c$. Consider the Lyapunov function $\varepsilon_s = e^{2cs}((\mathbf{x}_s)_{(i)} - \mathbf{x}^*_{(i)})^2$. For $s \leq t$

$$\begin{aligned}\dot{\varepsilon}_s &= e^{2cs}\left(2c((\mathbf{x}_s)_{(i)} - \mathbf{x}^*_{(i)})^2 - 2((\mathbf{x}_s)_{(i)} - \mathbf{x}^*_{(i)})(\dot{(\mathbf{x}_s)_{(i)}})\right) \\ &= e^{2cs}\left(2c(\mathbf{x}_s)_{(i)} - \mathbf{x}^*_{(i)})^2 - 2(\mathbf{x}_s)_{(i)} - \mathbf{x}^*_{(i)})^2\phi(s)\right) \\ &\leq 0.\end{aligned}$$

Thus

$$e^{2ct}((\mathbf{x}_t)_{(i)} - \mathbf{x}^*_{(i)})^2 \varepsilon_t \leq \varepsilon_0 = ((\mathbf{x}_0)_{(i)} - \mathbf{x}^*_{(i)})^2.$$

Since $\phi(t) = c$, we can compute the following bound on $\left|g'_{(i)}((\mathbf{x}_t)_{(i)})\right|$:

$$\begin{aligned}\left|g'_{(i)}((\mathbf{x}_t)_{(i)})\right| &= c\left|(\mathbf{x}_t)_{(i)} - \mathbf{x}^*_{(i)}\right| \\ &\leq ce^{-ct}\left|(\mathbf{x}_0)_{(i)} - \mathbf{x}^*_{(i)}\right|.\end{aligned}$$

Thus if $(\mathbf{x}_s)_{(i)} \neq \mathbf{x}^*_{(i)}$, then $\left|g'_{(i)}((\mathbf{x}_t)_{(i)})\right| \leq ce^{-ct}\left|(\mathbf{x}_0)_{(i)} - \mathbf{x}^*_{(i)}\right|$ for some $c \in [\mu, L]$. However, if $(\mathbf{x}_s)_{(i)} = \mathbf{x}^*_{(i)}$, then $\left|g'_{(i)}((\mathbf{x}_t)_{(i)})\right| = 0 \leq ce^{-ct}\left|(\mathbf{x}_0)_{(i)} - \mathbf{x}^*_{(i)}\right|$. Thus for every t, $\left|g'_{(i)}((\mathbf{x}_t)_{(i)})\right| \leq ce^{-ct}\left|(\mathbf{x}_0)_{(i)} - \mathbf{x}^*_{(i)}\right|$ for some $c \in [\mu, L]$.

Now split the integral as follows:

$$\begin{aligned}\zeta &= \int_0^\infty \sqrt{\sum_{i=1}^d \left(g'_{(i)}((\mathbf{x}_t)_{(i)})\right)^2}\, dt \\ &= \underbrace{\int_0^{1/L} \sqrt{\sum_{i=1}^d \left(g'_{(i)}((\mathbf{x}_t)_{(i)})\right)^2}\, dt}_{E_1} + \underbrace{\int_{1/L}^{1/\mu} \sqrt{\sum_{i=1}^d \left(g'_{(i)}((\mathbf{x}_t)_{(i)})\right)^2}\, dt}_{E_2} + \underbrace{\int_{1/\mu}^{\infty} \sqrt{\sum_{i=1}^d \left(g'_{(i)}((\mathbf{x}_t)_{(i)})\right)^2}\, dt}_{E_3}.\end{aligned}$$

To bound E_1 observe that for $t \in [0, 1/L]$ and $c \in [\mu, L]$, $ce^{-ct} \leq Le^{-Lt}$. Thus

$$\begin{aligned}E_1 &\leq \int_0^{1/L} Le^{-Lt}\sqrt{\sum_{i=1}^d \left((\mathbf{x}_0)_{(i)} - \mathbf{x}^*_{(i)}\right)^2}\, dt \\ &= \left(1 - \frac{1}{e}\right)\|\mathbf{x}_0 \; - \; \Pi_{\mathbf{X}^*}(\mathbf{x}_0)\|_2.\end{aligned}$$

Similarly for E_3 observe that for $t \in [1/\mu, \infty)$ and $c \in [\mu, L]$, $ce^{-ct} \leq \mu e^{-\mu t}$. Thus

$$\begin{aligned} E_3 &\leq \int_{1/\mu}^{\infty} \mu e^{-\mu t} \sqrt{\sum_{i=1}^{d} \left((\mathbf{x}_0)_{(i)} - \mathbf{x}^*_{(i)} \right)^2} \, dt \\ &= \left(\frac{1}{e} \right) \left\| \mathbf{x}_0 \ - \ \Pi_{\mathbf{X}^*}(\mathbf{x}_0) \right\|_2 . \end{aligned}$$

To bound E_2, we will further split the integral. Define $\alpha_i = (\mathbf{x}_0)_{(i)} - \mathbf{x}^*_{(i)}$. Observe that for some fixed $t > 0$ and $\max_{c \geq 0} ce^{-ct} = (1/te)$. Thus for $t \in [2^{k-1}/L, 2^k/L]$ and $c \geq 0$, $ce^{-ct} \leq \frac{L}{2^{k-1}e}$. Then

$$\begin{aligned} E_2 &\leq \sum_{k=1}^{r} \int_{2^{k-1}/L}^{2^k/L} \sqrt{\sum_{i=1}^{d} \left(g'_{(i)}((\mathbf{x}_t)_{(i)}) \right)^2} \, dt \\ &\leq \sum_{k=1}^{r} \int_{2^{k-1}/L}^{2^k/L} \left(\frac{L}{2^{k-1}e} \right) \left\| \mathbf{x}_0 \ - \ \Pi_{\mathbf{X}^*}(\mathbf{x}_0) \right\|_2 \\ &= \sum_{k=1}^{r} \left(\frac{2^k}{L} - \frac{2^{k-1}}{L} \right) \left(\frac{L}{2^{k-1}e} \right) \left\| \mathbf{x}_0 \ - \ \Pi_{\mathbf{X}^*}(\mathbf{x}_0) \right\|_2 \\ &= r \left\| \mathbf{x}_0 \ - \ \Pi_{\mathbf{X}^*}(\mathbf{x}_0) \right\|_2 / e \\ &\leq (\log_2(2\kappa)/e) \left\| \mathbf{x}_0 \ - \ \Pi_{\mathbf{X}^*}(\mathbf{x}_0) \right\|_2 \\ &\leq (1 + \log \kappa) \left\| \mathbf{x}_0 \ - \ \Pi_{\mathbf{X}^*}(\mathbf{x}_0) \right\|_2 . \end{aligned}$$

Resubstituting the bounds for E_1, E_2, E_3, we get

$$\zeta \leq (2 + \log \kappa) \left\| \mathbf{x}_0 \ - \ \Pi_{\mathbf{X}^*}(\mathbf{x}_0) \right\|_2 .$$

This completes the proof. ■

Appendix D. Proofs of Results in Section 3

Each proof in this section is organized in a separate subsection.

D.1 Proof of Lemma 14

First, we show that $f(\mathbf{x}_t)$ is non-increasing with respect to t. For any $s \geq 0$ observe that by LG,

$$\begin{aligned} f(\mathbf{x}_{s+1}) &= f(\mathbf{x}_s - \eta \nabla f(\mathbf{x}_s)) \\ &\leq f(\mathbf{x}_s) + \langle \nabla f(\mathbf{x}_s), -\eta \nabla f(\mathbf{x}_s) \rangle + \frac{L}{2} \left\| \eta \nabla f(\mathbf{x}_s) \right\|_2^2 \\ &= f(\mathbf{x}_s) - \eta(1 - \eta L/2) \left\| \nabla f(\mathbf{x}_s) \right\|_2^2 \\ &\leq f(\mathbf{x}_s) - \frac{\eta \left\| \nabla f(\mathbf{x}_s) \right\|_2^2}{2}, \end{aligned}$$

for $\eta \leq 1/L$. Thus for any s, $f(\mathbf{x}_{s+1}) \leq f(\mathbf{x}_s)$, as was to be shown. Next, fix any iterate t. We will show that $\|\mathbf{x}_s - \mathbf{x}_t\|_2^2$ is non-increasing in s for $s \leq t$. This would show self-contractedness for $s_3 = t$, for any t, concluding the proof. Consider any $s < t$, then

$$\begin{aligned}
\|\mathbf{x}_{s+1} - \mathbf{x}_t\|_2^2 &= \|\mathbf{x}_s - \eta \nabla f(\mathbf{x}_s) - \mathbf{x}_t\|_2^2 \\
&= \|\mathbf{x}_s - \mathbf{x}_t\|_2^2 + 2\left\langle \eta \nabla f(\mathbf{x}_s), \mathbf{x}_t - \mathbf{x}_s \right\rangle + \eta^2 \|\nabla f(\mathbf{x}_s)\|_2^2 \\
&\overset{\text{(i)}}{\leq} \|\mathbf{x}_s - \mathbf{x}_t\|_2^2 + 2\eta (f(\mathbf{x}_t) - f(\mathbf{x}_s)) + \eta^2 \|\nabla f(\mathbf{x}_s)\|_2^2 \\
&\overset{\text{(ii)}}{\leq} \|\mathbf{x}_s - \mathbf{x}_t\|_2^2 + 2\eta (f(\mathbf{x}_{s+1}) - f(\mathbf{x}_s)) + \eta^2 \|\nabla f(\mathbf{x}_s)\|_2^2 \\
&\overset{\text{LG}}{\leq} \|\mathbf{x}_s - \mathbf{x}_t\|_2^2 + 2\eta \left(\left\langle \nabla f(\mathbf{x}_s), -\eta \nabla f(\mathbf{x}_s) \right\rangle + \frac{\eta^2 L}{2} \|\nabla f(\mathbf{x}_s)\|_2^2 \right) + \eta^2 \|\nabla f(\mathbf{x}_s)\|_2^2 \\
&= \|\mathbf{x}_s - \mathbf{x}_t\|_2^2 + \eta^2 (\eta L - 1) \|\nabla f(\mathbf{x}_s)\|_2^2 \\
&\leq \|\mathbf{x}_s - \mathbf{x}_t\|_2^2,
\end{aligned}$$

for $\eta \leq 1/L$. Above, inequality (i) holds because of convexity and inequality (ii) holds as we have shown that $f(\mathbf{x}_t)$ is non-increasing in t, and $s + 1 \leq t$. Thus, $\|\mathbf{x}_s - \mathbf{x}_t\|_2^2$ is non-increasing in the iterates s, as was to be shown. ■

As indicated after the statement of Lemma 14, if f is convex and has L-Lipschitz gradients, then GD with $\eta \in (0, 2/L]$ is a descent method (we show it below); however for GD to be self-contracted, one needs further restriction on the step-size. To see the latter, consider $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$. Here we have $L = 2$. Let us set $\eta = 7/4L = 7/8$. Let $x_0 = 8$, so that $x_1 = 8 - (7/8) \cdot (2 \cdot 8) = -6$ and $x_2 = -6 - (7/8) \cdot (2 \cdot -6) = 4.5$. However,

$$|x_2 - x_1| > |x_2 - x_0|,$$

and so (x_0, x_1, x_2) cannot be part of a self-contracted curve.

The fact that GD is a descent method if $\eta \in (0, 1/L]$ is a well-known fact, but perhaps a bit less known is that this can be shown for $\eta \in (0, 2/L]$ (using standard techniques). We show it here for completeness. For any $\mathbf{x}^* \in \mathbf{X}^*$,

$$\begin{aligned}
\|\mathbf{x}_{s+1} - \mathbf{x}^*\|_2^2 &= \|\mathbf{x}_s - \eta \nabla f(\mathbf{x}_s) - \mathbf{x}^*\|_2^2 \\
&= \|\mathbf{x}_s - \mathbf{x}^*\|_2^2 - 2\left\langle \eta \nabla f(\mathbf{x}_s), \mathbf{x}_s - \mathbf{x}^* \right\rangle + \eta^2 \|\nabla f(\mathbf{x}_s)\|_2^2 \\
&\overset{\text{(i)}}{=} \|\mathbf{x}_s - \mathbf{x}^*\|_2^2 - 2\left\langle \eta \nabla f(\mathbf{x}_s) - \eta \nabla f(\mathbf{x}^*), \mathbf{x}_s - \mathbf{x}^* \right\rangle + \eta^2 \|\nabla f(\mathbf{x}_s)\|_2^2. \\
&\overset{\text{(ii)}}{\leq} \|\mathbf{x}_s - \mathbf{x}^*\|_2^2 - 2\eta \|\nabla f(\mathbf{x}_s)\|_2^2 / L + \eta^2 \|\nabla f(\mathbf{x}_s)\|_2^2. \\
&\leq \|\mathbf{x}_s - \mathbf{x}^*\|_2^2,
\end{aligned}$$

since $\eta \leq 2/L$ (with strict inequality in the last step if $\eta < 2/L$ and $\nabla f(\mathbf{x}_s) \neq 0$). Equality (i) holds because $\nabla f(\mathbf{x}^*) = 0$ and inequality (ii) holds since LG+convexity implies $\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 / L$. This is shown formally by Zhou (2018, Lemma 4).

D.2 Proof of Theorem 15

The GF result is due to Manselli and Pucci (1991). We analyze GD curves by using Lemma 14 to extend the techniques introduced by Daniilidis et al. (2015, Theorem 3.1) and Manselli and Pucci (1991) for analyzing GF curves.

We will assume $d \geq 2$ since in the case $d = 1$ the self-contracted curve is the shortest path. Some definitions are in order:

- The projection of any set K to a line u will be denoted as $\Pi_{\mathbf{u}}(K)$.
- Length of a one dimensional object (for example the projection of a set K to a line u) will be denoted as $\ell(\cdot)$ (for example $\ell(\Pi_{\mathbf{u}}(K))$.
- Mean width of a convex set K:
$$W(K) := (\sigma_d)^{-1} \int_{\mathbf{u} \in \mathbb{S}^{d-1}} \ell(\Pi_{\mathbf{u}}(K)) \, d\mathbf{u},$$
where σ_d is the volume of $\mathbb{S}^{d-1}$, with respect to the Lebesgue measure.
- For $k \in \mathbb{N}_0$, $\Gamma(k)$ is the set of iterates after iteration k (inclusive): $\Gamma(k) := \{\mathbf{x}_k, \mathbf{x}_{k+1}, \ldots\}$.
- The convex closure of the set $\Gamma(k)$ will be denoted as $\Omega(k)$.

Note that since the GD curve is self-contracted and converges to $\mathbf{x}_\infty$, we have for all $t \in \mathbb{N}_0$,

$$\left\| \mathbf{x}_t - \mathbf{x}_\infty \right\|_2 \leq \left\| \mathbf{x}_0 - \mathbf{x}_\infty \right\|_2 .$$

Thus, all iterates $\mathbf{x}_0, \mathbf{x}_1, \ldots$ stay within a ball of radius $\left\| \mathbf{x}_0 - \mathbf{x}_\infty \right\|_2$ centered at $\mathbf{x}_\infty$. The mean width of this path can be at most the diameter of the ball, that is, $W(\Omega(0)) \leq 2 \left\| \mathbf{x}_0 - \mathbf{x}_\infty \right\|_2$. We will be showing that if $\eta \leq 1/L$

$$\zeta_\eta \leq 28^{2d^2} \cdot W(\Omega(0)), \tag{28}$$

and if $\eta \leq 1/2L\sqrt{d}$

$$\zeta_\eta \leq 2^{(4d \log d) - 1} \cdot W(\Omega(0)), \tag{29}$$

which will lead to the bound in the theorem since $W(\Omega(0)) \leq 2 \left\| \mathbf{x}_0 - \mathbf{x}_\infty \right\|_2$ and $d \geq 2$. Both these bounds will be shown by setting up a recurrence of the form.

$$W(\Omega(k+1)) + \epsilon \left\| \mathbf{x}_{k+1} - \mathbf{x}_k \right\|_2 \leq W(\Omega(k)). \tag{30}$$

for two different values of ϵ.

By telescoping to T iterations, this would lead to

$$\sum_{k=0}^{T} \left\| \mathbf{x}_{k+1} - \mathbf{x}_k \right\|_2 \leq \frac{1}{\epsilon} \left(W(\Omega(0)) - W(\Omega(T+1)) \right) \leq \frac{W(\Omega(0))}{\epsilon}.$$

Since the right hand side is the same for every T, indeed we would obtain

$$\zeta_\eta = \sum_{k=0}^{\infty} \left\| \mathbf{x}_{k+1} - \mathbf{x}_k \right\|_2 \leq \frac{W(\Omega(0))}{\epsilon},$$

which would complete the proof. It remains to prove Equation (30) with the appropriate values of ϵ that would lead to Equations (28) and (29).

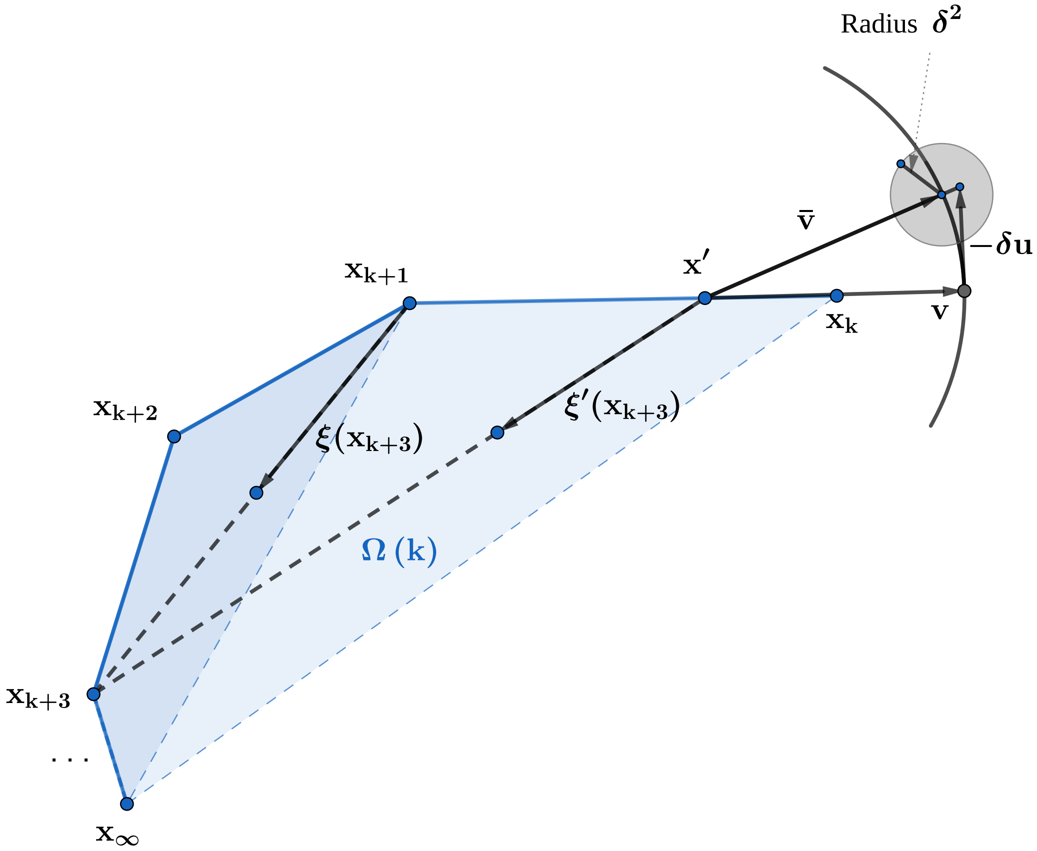


Figure 9: Illustration of some entities appearing in the proof of Theorem 15.

D.2.1 Proof of Equation (28)

We will show that if $\eta \leq 1/L$, Equation (30) is true with $\epsilon = (1/28)^{2d^2}$. Define the following entities (see Figure 9):

$$\begin{aligned}
\mathbf{x}' &:= \frac{\mathbf{x}_{k+1}}{3} + \frac{2\mathbf{x}_k}{3}, \\
\mathbf{v} &:= \frac{\mathbf{x}_k - \mathbf{x}_{k+1}}{\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2}, \\
\xi'(\mathbf{y}) &:= \frac{\mathbf{y} - \mathbf{x}'}{\|\mathbf{y} - \mathbf{x}'\|_2}, \text{ for } \mathbf{y} \neq \mathbf{x}', \\
\xi(\mathbf{y}) &:= \frac{\mathbf{y} - \mathbf{x}_{k+1}}{\|\mathbf{y} - \mathbf{x}_{k+1}\|_2}, \text{ for } \mathbf{y} \neq \mathbf{x}_{k+1}.
\end{aligned}$$

Also denote $\mathbf{v}^\perp$ as the orthogonal hyperplane to $\mathbf{v}$. Define $\delta = (1/27)^d$. We will prove that for some unit vector $\mathbf{u} \in \mathbf{v}^\perp$, the unit vector

$$\bar{\mathbf{v}} = \frac{\mathbf{v} - \delta\mathbf{u}}{\|\mathbf{v} - \delta\mathbf{u}\|_2}, \tag{31}$$

has at most a fixed negative inner product with any unit vector $\xi'(\mathbf{x}_t)$, namely:

$$\left\langle \bar{\mathbf{v}}, \xi'(\mathbf{x}_t) \right\rangle \le -\delta^2. \tag{32}$$

As we see later, this will allow us to bound the mean width integral for $\Omega(k+1)$, in terms of the mean width integral for $\Omega(k)$ in order to prove Equation (30). Note that since δ is a small constant, $\bar{\mathbf{v}}$ is very close to $\mathbf{v}$. To motivate the truth of (32), note the following fact about $\mathbf{v}$ itself. For all $\mathbf{y} \in \Gamma(k+2)$,

$$\left\langle \mathbf{v}, \mathbf{y} - \mathbf{x}' \right\rangle \le -\frac{\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2}{6} < 0. \qquad \left(\equiv \left\langle \mathbf{v}, \xi'(\mathbf{y}) \right\rangle \le -\frac{\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2}{6\left\|\mathbf{y} - \mathbf{x}'\right\|_2} < 0.\right) \tag{33}$$

This is true by the self-contractedness property. To see this, think of $\mathbf{x}'$ as the origin , and $\mathbf{x}_k$ as the 'positive' direction. Since $\|\mathbf{y} - \mathbf{x}_{k+1}\|_2 \le \|\mathbf{y} - \mathbf{x}_k\|_2$, the projection of $\mathbf{y}$ onto the segment $[\mathbf{x}_k, \mathbf{x}_{k+1}]$ lies towards the negative side and farther than the mid-point. However, $\mathbf{x}'$ lies towards $\mathbf{x}_k$, and hence the positive side. Thus, the projection of $\mathbf{y} - \mathbf{x}'$ points in the opposite direction as $\mathbf{v}$ and has a magnitude at least the distance between $\mathbf{x}'$ and the mid-point: $\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 / 6$. Thus,

$$\left\langle \mathbf{v}, \mathbf{y} - \mathbf{x}' \right\rangle \le - \|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 / 6.$$

The algebra above suggests that if the projection of $\mathbf{y}$ onto the segment $[\mathbf{x}_k, \mathbf{x}_{k+1}]$ is only slightly negative, then $\|\mathbf{y} - \mathbf{x}'\|_2$ is large and most of the component is in the $\mathbf{v}^\perp$ direction. In this case, we need to only find a small vector (namely $\delta\mathbf{u}$) in the perpendicular direction that has negative inner product with $\mathbf{y}-\mathbf{x}'$. This motivates the definition of $\bar{\mathbf{v}}$ in (31). Note however that this vector $\delta\mathbf{u}$ needs to uniformly have a negative inner product with respect to every $\xi'(\mathbf{y})$. To show that this is possible, we will use the self-contractedness property to argue that all the $\xi'(\mathbf{y})$ lie nearly in a hemisphere. In what follows, we formalize these ideas.

First let us divide the unit vectors $\xi'(\mathbf{x}_t)$ into two sets: points that have a small component in the direction opposite to $\mathbf{v}$ and points that lie mostly in $\mathbf{v}^\perp$. Define,

$$\Gamma' := \{\mathbf{y} \in \Gamma(k+1) : \left\langle \mathbf{v}, \xi'(\mathbf{y}) \right\rangle \le -2\delta\}.$$

Note that for $\mathbf{y} \in \Gamma'$,

$$\begin{aligned} \left\langle \bar{\mathbf{v}}, \xi'(\mathbf{y}) \right\rangle &= \left\langle \frac{\mathbf{v} - \delta\mathbf{u}}{\|\mathbf{v} - \delta\mathbf{u}\|_2}, \xi'(\mathbf{y}) \right\rangle \\ &\le \frac{-2\delta}{\sqrt{1+\delta^2}} + \delta \qquad \text{(by Cauchy-Schwarz since } \|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1) \\ &\le -\delta^2, \end{aligned}$$

since $0 \le \delta \le 1/27$. Thus Equation (32) is satisfied for $\mathbf{y} \in \Gamma'$. For $\mathbf{y} \in \Gamma \setminus \Gamma'$, we note three properties that will be used later. First, from (33), and the definition of Γ'

$$-2\delta < \left\langle \mathbf{v}, \xi'(\mathbf{y}) \right\rangle \le -\frac{\mathbf{x}_k - \mathbf{x}_{k+1}}{6\left\|\mathbf{y} - \mathbf{x}'\right\|_2}.$$

On cross multiplying

$$\left\|\mathbf{y}-\mathbf{x}'\right\|_2 > \frac{\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2}{12\delta}, \tag{34}$$

which says that $\|\mathbf{y}-\mathbf{x}'\|_2$ is large as motivated earlier. Thus,

$$\|\mathbf{y}-\mathbf{x}\|_2 \geq \left\|\mathbf{y}-\mathbf{x}'\right\|_2 - \left\|\mathbf{x}'-\mathbf{x}\right\|_2 \geq \left(\frac{1}{12\delta} - \frac{2}{3}\right)\left\|\mathbf{x}-\mathbf{x}'\right\|_2 = \frac{1-8\delta\left\|\mathbf{x}-\mathbf{x}'\right\|_2}{12\delta} \tag{35}$$

Also, $-2\delta < \langle \mathbf{v}, \xi'(\mathbf{y})\rangle < 0$, so that,

$$\left|\left\langle \mathbf{v}, \xi'(\mathbf{y})\right\rangle\right| < 2\delta. \tag{36}$$

Let the component of $\xi'(\mathbf{y})$ in $\mathbf{v}^\perp$ be $\xi'_\perp(\mathbf{y})$. Thus,

$$\left\|\xi'_\perp(\mathbf{y})\right\|_2^2 = 1 - (\left\langle \mathbf{v}, \xi'(\mathbf{y})\right\rangle)^2 \geq 1 - 4\delta^2. \tag{37}$$

Using these facts, the goal now will be to show that for all $\mathbf{y} \in \Gamma \setminus \Gamma'$, $\xi'(\mathbf{y})$'s are almost in a hemisphere so that there we can find a common vector $\mathbf{u}$ as desired which has a high negative inner product for (32). The idea is that because of (37), $\xi'_\perp(\mathbf{y})$ is almost perpendicular to $\mathbf{v}$ and so $\xi'(\mathbf{y})$ looks almost like $\xi(\mathbf{y})$. Observe that for $\xi(\mathbf{y})$ the hemisphere property is easy to see: for all $\mathbf{y}, \mathbf{z} \in \Gamma(k+1) \setminus \Gamma'$ such that $\mathbf{z}$ comes after $\mathbf{y}$ in the path, using self-contractedness (Lemma 14), we know that $\|\mathbf{y}-\mathbf{z}\|_2 \leq \|\mathbf{x}_{k+1}-\mathbf{z}\|_2$, and thus in the triangle formed by $\mathbf{x}_{k+1}, \mathbf{y}, \mathbf{z}$, the segment between $\mathbf{y}$ and $\mathbf{z}$ is not the longest side. This means that the angle at $\mathbf{x}_{k+1}$ is acute so that

$$\langle \xi(\mathbf{y}), \xi(\mathbf{z})\rangle \geq 0. \tag{38}$$

Hence all vectors $\{\xi(\mathbf{y}) : \mathbf{y} \in \Gamma(k+1) \setminus \Gamma'\}$ belong in the same hemisphere. To show a similar result for $\xi'(\mathbf{y})$, we first bound $\|\xi(\mathbf{y}) - \xi'(\mathbf{y})\|_2$:

$$\begin{aligned}
\left\|\xi(\mathbf{y}) - \xi'(\mathbf{y})\right\|_2 &= \left\|\frac{\mathbf{y}-\mathbf{x}_{k+1}}{\|\mathbf{y}-\mathbf{x}_{k+1}\|_2} - \frac{\mathbf{y}-\mathbf{x}'}{\|\mathbf{y}-\mathbf{x}'\|_2}\right\|_2 \\
&\leq \left\|\frac{\mathbf{y}-\mathbf{x}_{k+1}}{\|\mathbf{y}-\mathbf{x}_{k+1}\|_2} - \frac{\mathbf{y}-\mathbf{x}'}{\|\mathbf{y}-\mathbf{x}_{k+1}\|_2}\right\|_2 + \left\|\frac{\mathbf{y}-\mathbf{x}'}{\|\mathbf{y}-\mathbf{x}_{k+1}\|_2} - \frac{\mathbf{y}-\mathbf{x}'}{\|\mathbf{y}-\mathbf{x}'\|_2}\right\|_2 \\
&= \frac{\|\mathbf{x}_{k+1}-\mathbf{x}'\|_2}{\|\mathbf{y}-\mathbf{x}_{k+1}\|_2} + \frac{\|\mathbf{y}-\mathbf{x}'\|_2 - \|\mathbf{y}-\mathbf{x}_{k+1}\|_2}{\|\mathbf{y}-\mathbf{x}_{k+1}\|_2} \\
&\leq \frac{\|\mathbf{x}_{k+1}-\mathbf{x}'\|_2}{\|\mathbf{y}-\mathbf{x}_{k+1}\|_2} + \frac{\|\mathbf{x}'-\mathbf{x}_{k+1}\|_2}{\|\mathbf{y}-\mathbf{x}_{k+1}\|_2} \\
&= \frac{2\,\|\mathbf{x}_{k+1}-\mathbf{x}'\|_2}{\|\mathbf{y}-\mathbf{x}_{k+1}\|_2} = \frac{4\,\|\mathbf{x}_{k+1}-\mathbf{x}'\|_2}{3\,\|\mathbf{y}-\mathbf{x}\|_2} \\
&\overset{(35)}{\leq} \frac{16\delta}{1-8\delta} \\
&\leq 32\delta,
\end{aligned}$$

since $\delta \leq 1/27$. Now we consider any $\mathbf{y}, \mathbf{z} \in \Gamma(k+1) \setminus \Gamma'$. Define $\delta_{\mathbf{y}} := \xi(\mathbf{y}) - \xi'(\mathbf{y})$ and $\delta_{\mathbf{z}} := \xi(\mathbf{z}) - \xi'(\mathbf{z})$, then,

$$\begin{aligned} 0 \leq \langle \xi(\mathbf{y}), \xi(\mathbf{z}) \rangle &= \left\langle \xi'(\mathbf{y}) + \delta_{\mathbf{y}}, \xi(\mathbf{z}) \right\rangle \\ &\leq \left\langle \xi'(\mathbf{y}), \xi(\mathbf{z}) \right\rangle + \|\delta_{\mathbf{y}}\|_2 && \text{(Cauchy-Schwarz)} \\ &\leq \left\langle \xi'(\mathbf{y}), \xi'(\mathbf{z}) + \delta_{\mathbf{z}} \right\rangle + 32\delta \\ &\leq \left\langle \xi'(\mathbf{y}), \xi'(\mathbf{z}) \right\rangle + 64\delta. \end{aligned}$$

Thus $\langle \xi'(\mathbf{y}), \xi'(\mathbf{z}) \rangle \geq -64\delta$. Further, from (36)

$$\begin{aligned} \left\langle \xi'_\perp(\mathbf{y}), \xi'_\perp(\mathbf{z}) \right\rangle &= \left\langle \xi'(\mathbf{y}), \xi'(\mathbf{z}) \right\rangle - (\left\langle \mathbf{v}, \xi'(\mathbf{y}) \right\rangle)(\left\langle \mathbf{v}, \xi'(\mathbf{z}) \right\rangle) \\ &\geq -64\delta - 4\delta^2 \\ &\geq -65\delta, \end{aligned}$$

since $\delta \leq 1/27$. From (37), $\|\xi'_\perp(\mathbf{y})\|_2 \cdot \|\xi'_\perp(\mathbf{y})\|_2 \geq 1 - 4\delta^2$, so that the set of vectors $S = \{\widehat{\xi'_\perp(\mathbf{y})} : \mathbf{y} \in \Gamma(k+1) \setminus \Gamma'\}$ ($\widehat{\mathbf{a}}$ denotes the unit vector in the direction of $\mathbf{a}$) satisfies: for all $\mathbf{y}, \mathbf{z} \in S$,

$$\langle \mathbf{y}, \mathbf{z} \rangle \geq \frac{-65\delta}{1 - 4\delta^2} \geq -66\delta = -66\left(\frac{1}{27}\right)^d \geq -\left(\frac{1}{3}\right)^d, \tag{39}$$

for $d \geq 2$. As motivated earlier, this is a set of vectors that is almost in a hemisphere. At this point, we invoke the following lemma proved by Daniilidis et al. (2015).

Lemma 24 (Lemma 3.2, (Daniilidis et al., 2015)) *Let $\Sigma \subset \mathbb{S}^{d-1}$ be a set satisfying*

$$\langle \mathbf{x}, \mathbf{y} \rangle \geq -\left(\frac{1}{3}\right)^{d+1} \quad \textit{for all } \mathbf{x}, \mathbf{y} \in \Sigma.$$

Then there exists a $\mathbf{u} \in \mathbb{S}^{d-1}$ such that

$$\langle \mathbf{u}, \mathbf{y} \rangle \geq \left(\frac{1}{3}\right)^{2d+1} \quad \textit{for all } \mathbf{y} \in \Sigma.$$

The proof of the above lemma uses a packing argument. We use the lemma for the set S identified above (Equation (39)). Note that all vectors in S lie in $\mathbb{S}^{d-1} \cap \mathbf{v}^\perp$, which can be identified as a shell in $d-1$ dimensions, homomorphic to $\mathbb{S}^{d-2}$. Thus there exists a vector $\mathbf{u} \in \mathbb{S}^{d-1} \cap \mathbf{v}$ such that for all $\mathbf{y} \in \Gamma(k+1) \setminus \Gamma'$,

$$\left\langle \mathbf{u}, \widehat{\xi'_\perp(\mathbf{y})} \right\rangle \geq \left(\frac{1}{3}\right)^{2(d-1)+1} = \left(\frac{1}{3}\right)^{2d-1}. \tag{40}$$

We pick this $\mathbf{u}$ to define $\bar{\mathbf{v}}$ in Equation (31). Thus, for $\mathbf{y} \in \Gamma(k+1) \setminus \Gamma'$,

$$\begin{aligned} \left\langle \mathbf{v} - \delta\mathbf{u}, \xi'(\mathbf{y}) \right\rangle &\overset{(33)}{\leq} -\delta \left\langle \mathbf{u}, \xi'(\mathbf{y}) \right\rangle \\ &= -\delta \left\langle \mathbf{u}, \xi'_\perp(\mathbf{y}) \right\rangle \end{aligned}$$

$$\begin{aligned} &= -\delta \left\|\xi'_\perp(\mathbf{y})\right\|_2 \left\langle \mathbf{u}, \widehat{\xi'_\perp(\mathbf{y})} \right\rangle \\ &\overset{(37)}{\le} -\delta\sqrt{1-4\delta^2}\left\langle \mathbf{u}, \widehat{\xi'_\perp(\mathbf{y})}\right\rangle \\ &\overset{(40)}{\le} -\delta\sqrt{1-4\delta^2}\left(\frac{1}{3}\right)^{2d-1} \\ &\le -\delta\left(\frac{1}{3}\right)^{2d}, \end{aligned}$$

since $\delta \le 1/27$. Finally, $\left\|\mathbf{v}-\delta\mathbf{u}\right\|_2 = \sqrt{1+\delta^2} \le 3$ so that

$$\left\langle \bar{\mathbf{v}}, \xi'(\mathbf{y})\right\rangle = \frac{\langle \mathbf{v}-\delta\mathbf{u}, \xi'(\mathbf{y})\rangle}{\left\|\mathbf{v}-\delta\mathbf{u}\right\|_2} \le \frac{-\delta\left(\frac{1}{3}\right)^{2d}}{3} \le -\delta\left(\frac{1}{3}\right)^{3d} = -\delta^2,$$

which gives us (32) as needed.

Finally, we use this identified $\bar{\mathbf{v}}$ to prove the recurrence (30). Consider the part of the shell $\mathbb{S}^{d-1}$ δ^2-close to $\bar{\mathbf{v}}$:

$$\mathbb{S}' := \{\mathbf{v}' \in \mathbb{S}^{d-1} : \left\|\mathbf{v}'-\bar{\mathbf{v}}\right\|_2 \le \delta^2\},$$

then

$$\begin{aligned} \sigma_d W(\Omega(k+1)) &= \int_{\mathbf{u}\in\mathbb{S}^{d-1}} \ell(\Pi_{\mathbf{u}}(\Omega(k+1)))\ d\mathbf{u} \\ &= \int_{\mathbf{u}\in\mathbb{S}'} \ell(\Pi_{\mathbf{u}}(\Omega(k+1)))d\mathbf{u} + \int_{\mathbf{u}\notin\mathbb{S}'} \ell(\Pi_{\mathbf{u}}(\Omega(k+1)))\ d\mathbf{u} \\ &\le \int_{\mathbf{u}\in\mathbb{S}'} \ell(\Pi_{\mathbf{u}}(\Omega(k+1)))d\mathbf{u} + \int_{\mathbf{u}\notin\mathbb{S}'} \ell(\Pi_{\mathbf{u}}(\Omega(k)))\ d\mathbf{u}, \end{aligned} \tag{41}$$

since $\Omega(k+1) \subset \Omega(k)$. For the first integral above, note that for $\mathbf{u} \in \mathbb{S}'$:

1. Inner product with the vector $\mathbf{x}_k - \mathbf{x}'$ is high:

$$\begin{aligned} \left\langle \mathbf{x}_k - \mathbf{x}', \mathbf{u}\right\rangle &= \left\|\mathbf{x}_k - \mathbf{x}'\right\|_2 \langle \mathbf{v}, \mathbf{u}\rangle \\ &= \left(\frac{\left\|\mathbf{x}_k - \mathbf{x}_{k+1}\right\|_2}{3}\right)\langle \mathbf{v}, \mathbf{u}\rangle \\ &\ge \left(\frac{\left\|\mathbf{x}_k - \mathbf{x}_{k+1}\right\|_2}{3}\right)(1-\delta-\delta^2) \\ &\ge \frac{\left\|\mathbf{x}_k - \mathbf{x}_{k+1}\right\|_2}{4}. \end{aligned}$$

   since $1-\delta-\delta^2 \ge 3/4$ for $\delta \le 1/27$.

2. Inner product with the vector $\mathbf{y} - \mathbf{x}'$ for every $\mathbf{y} \in \Gamma(k+1)$ is non-positive:

$$\begin{aligned} \left\langle \mathbf{y} - \mathbf{x}', \mathbf{u}\right\rangle &= \left\|\mathbf{y} - \mathbf{x}'\right\|_2 \left\langle \xi'(\mathbf{y}), \mathbf{u}\right\rangle \\ &= \left\|\mathbf{y} - \mathbf{x}'\right\|_2 \left\langle \xi'(\mathbf{y}), \mathbf{v}\right\rangle + \left\|\mathbf{y} - \mathbf{x}'\right\|_2 \left\langle \xi'(\mathbf{y}), \mathbf{v}-\mathbf{u}\right\rangle \end{aligned}$$

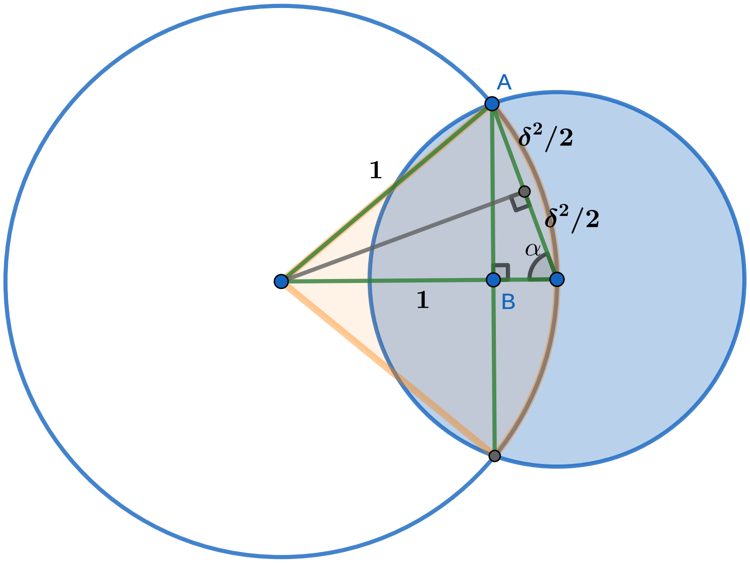


Figure 10: Two dimensional illustration of the intersection of a shell of radius 1 (unshaded circle above) with a shell of radius δ^2 (shaded circle above).

$$
\begin{aligned}
&\overset{(32)}{\leq} \left\|\mathbf{y}-\mathbf{x}'\right\|_2(-\delta^2) + \left\|\mathbf{y}-\mathbf{x}'\right\|_2 \left\langle \xi'(\mathbf{y}), \mathbf{v}-\mathbf{u}\right\rangle \\
&\leq -\left\|\mathbf{y}-\mathbf{x}'\right\|_2 \delta^2 + \left\|\mathbf{y}-\mathbf{x}'\right\|_2 \left\|\mathbf{v}-\mathbf{u}\right\|_2 \\
&\leq -\left\|\mathbf{y}-\mathbf{x}'\right\|_2 \delta^2 + \left\|\mathbf{y}-\mathbf{x}'\right\|_2 (\delta^2) \\
&\leq 0.
\end{aligned}
$$

Indeed, this means that for any point in the convex hull of $\Gamma(k+1)$ the same is true—that is for $\mathbf{y} \in \Omega(k+1)$, $\langle \mathbf{y}-\mathbf{x}', \mathbf{u}\rangle \leq 0$.

Using these two facts, we have the following lower bound on the length of $\Pi_{\mathbf{u}}(\Omega(k))$ for any $\mathbf{u} \in \mathbb{S}'$,

$$
\begin{aligned}
\ell(\Pi_{\mathbf{u}}(\Omega(k))) &\geq \left\langle \mathbf{x}_k - \mathbf{x}', \mathbf{u}\right\rangle + \ell(\Pi_{\mathbf{u}}(\Omega(k+1))) \\
&\geq \frac{\left\|\mathbf{x}_k - \mathbf{x}_{k+1}\right\|_2}{4} + \ell(\Pi_{\mathbf{u}}(\Omega(k+1))).
\end{aligned}
$$

Thus,

$$
\int_{\mathbf{u}\in\mathbb{S}'} \ell(\Pi_{\mathbf{u}}(\Omega(k+1)))\, d\mathbf{u} \leq \int_{\mathbf{u}\in\mathbb{S}'} \ell(\Pi_{\mathbf{u}}(\Omega(k)))\, d\mathbf{u} - \int_{\mathbf{u}\in\mathbb{S}'} \left(\frac{\left\|\mathbf{x}_k - \mathbf{x}_{k+1}\right\|_2}{4}\right)\, d\mathbf{u}
$$

Continuing from (41), and substituting the above inequality,

$$
\sigma_d W(\Omega(k+1)) \leq -\int_{\mathbf{u}\in\mathbb{S}'} \left(\frac{\left\|\mathbf{x}_k - \mathbf{x}_{k+1}\right\|_2}{4}\right)\, d\mathbf{u} + \int_{\mathbf{u}\in\mathbb{S}^{d-1}} \ell(\Pi_{\mathbf{u}}(\Omega(k)))\, d\mathbf{u}
$$

$$= -\frac{\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 \cdot \text{Volume}(\mathbb{S}')}{4} + \int_{\mathbf{u} \in \mathbb{S}^{d-1}} \ell(\Pi_{\mathbf{u}}(\Omega(k)))\ d\mathbf{u}.$$

On simplifying, this leads to the bound

$$W(\Omega(k+1)) \leq -\underbrace{\left(\frac{\text{Volume}(\mathbb{S}')}{4\sigma_d}\right)}_{=:\ \epsilon,\text{ as needed for (30)}} \|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 + W(\Omega(k)),$$

where Volume($\cdot$) is defined with respect to the Lebesgue measure in $d-1$ dimensions. To compute ϵ, we find this volume. Note that $\mathbb{S}'$ is a sector whose boundary is the intersection between $\mathbb{S}^{d-1}$ and a shell of radius δ^2 with center on the surface of $\mathbb{S}^{d-1}$. This boundary defines a shell in $(d-1)$ dimensions. See Figure 10. In two dimensions the intersection is just two points, but for general d-dimensions, it is a shell in $(d-1)$-dimensions. The radius of this shell γ is the length of the segment AB which can be calculated with simple trigonometric calculations.

$$\gamma = \text{length(AB)} = \delta^2 \sin\alpha = \delta^2\sqrt{1-\cos^2\alpha} = \delta^2\left(1 - \frac{\delta^4}{4}\right) \geq \left(\frac{1}{28}\right)^{2d}.$$

The volume of the sphere defined by this $(d-1)$-dimensional shell lower bounds the volume of $\mathbb{S}'$. To illustrate in two dimensions (see Figure 10), think of this volume as the length of the orange arc that lies in the shaded circle. This length is lower bounded by the length of the diameter 2γ. In general, for d-dimensions, this would be the volume of the $(d-1)$-dimensional sphere of radius γ. Using the formula of the volume of a $(d-1)$-dimensional sphere (Γ below denotes the gamma function),

$$\text{Volume}(\mathbb{S}') \geq \frac{((\pi\gamma^2)^{d-1/2}}{\Gamma\left(\frac{d-1}{2}+1\right)}.$$

Also σ_d is the volume of a d-dimensional shell, given by

$$\sigma_d = \frac{d\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)}.$$

Thus,

$$\epsilon = \frac{\text{Volume}(\mathbb{S}')}{4\sigma_d} \geq \frac{\gamma^{d-1}\cdot\Gamma\left(\frac{d}{2}+1\right)}{4\sqrt{\pi}\cdot\Gamma\left(\frac{d-1}{2}+1\right)} \geq \gamma^d = \left(\frac{1}{28}\right)^{2d^2},$$

as was needed to be shown to prove the recurrence (30).

D.2.2 Proof of Equation (29)

We will show that for $\eta \leq 1/2L\sqrt{d}$, Equation (30) is true with $\epsilon = 1/2^{(4d\log d)-1}$. We use some of the notation introduced in Section D.2.1.

The step-size constraint ensures that gradients at two consecutive iterates have a high inner product (or small angle). For some $k \geq 0$, we have by LG

$$\|\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k+1})\|_2 \leq L\,\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 = \eta L\,\|\nabla f(\mathbf{x}_k)\|_2$$

Squaring both sides and rearranging leads to

$$\|\nabla f(\mathbf{x}_{k+1})\|_2^2 + (1-\eta^2 L^2)\|\nabla f(\mathbf{x}_k)\|_2^2 \leq 2\langle \nabla f(\mathbf{x}_{k+1}), \nabla f(\mathbf{x}_k)\rangle = 2\|\nabla f(\mathbf{x}_k)\|_2 \|\nabla f(\mathbf{x}_{k+1})\|_2 \cos\theta,$$

where θ is the angle between $\nabla f(\mathbf{x}_k)$ and $\nabla f(\mathbf{x}_{k+1})$. Further observe that

$$\|\nabla f(\mathbf{x}_{k+1})\|_2^2 + (1-\eta^2 L^2)\|\nabla f(\mathbf{x}_k)\|_2^2 \geq 2\sqrt{(1-\eta^2 L^2)}\,\|\nabla f(\mathbf{x}_{k+1})\|_2 \|\nabla f(\mathbf{x}_k)\|_2 .$$

Putting it together, we obtain

$$\sqrt{1-\eta^2 L^2} \leq \cos\theta,$$

which leads to the following dimension dependent lower bound

$$\sqrt{1-(1/4d)} \leq \cos\theta. \tag{42}$$

As shown in Equation (38), $\langle \xi(\mathbf{y}), \xi(\mathbf{z})\rangle \geq 0$ for $\mathbf{y}, \mathbf{z} \in \Gamma(k+2)$. Using this fact, we want to show that there exists a single unit vector w such that $\langle \mathbf{w}, \xi(\mathbf{y})\rangle$ is large for all $\mathbf{y} \in \Gamma(k+2)$. To obtain this, we use a result due to (Santaló, 1946, Theorem 1). To apply Santaló's theorem, consider the set $K = \{\xi(\mathbf{y}) : \mathbf{y} \in \Gamma(k+2)\} \subset \mathbb{S}^{d-1}$ (thus $n = d-1$). By Equation (38), the spherical diameter D of K has cosine at least 0. We wish to lower bound the cosine of the spherical radius R. The result of Santaló has three case-wise conclusions, but each of them assert that if $\cos D \geq 0$, then

$$\frac{d\cos^2 R - 1}{\text{(positive quantity)}} \geq \cos D \geq 0.$$

A fortiori this implies that $d\cos^2 R - 1 \geq 0$ or $\cos R \geq \sqrt{1/d}$. Using the definition of spherical radius, we conclude that there exists a $\mathbf{w} \in \mathbb{S}^{d-1}$ such that for all $\mathbf{y} \in \Gamma(k+2)$

$$\langle \mathbf{w}, \xi(\mathbf{y})\rangle \geq \sqrt{1/d}. \tag{43}$$

For $\mathbf{y} = \mathbf{x}_{k+2} = \mathbf{x}_{k+1} - \eta\nabla f(\mathbf{x}_{k+1})$, we have $\langle \mathbf{w}, \nabla f(\mathbf{x}_{k+1}\rangle \geq \sqrt{1/d}$. We use this fact to show that $\langle \mathbf{w}, \xi(\mathbf{x}_k)\rangle$ is negative, as follows. Let $\angle(\mathbf{u}, \mathbf{v})$ denote $\arccos(\langle \widehat{\mathbf{u}}, \widehat{\mathbf{v}}\rangle)$, where $\widehat{\mathbf{u}}$ and $\widehat{\mathbf{v}}$ are unit vectors in the direction of $\mathbf{u}$ and $\mathbf{v}$ respectively (so that arccos of their inner product gives us the angle). Then

$$\begin{aligned}
\langle \mathbf{w}, \xi(\mathbf{x}_k)\rangle &= \cos(\angle(\mathbf{w}, \mathbf{x}_k - \mathbf{x}_{k+1})) \qquad (44)\\
&= -\cos(\angle(\mathbf{w}, \nabla f(\mathbf{x}_k)))\\
&\leq -\cos(\angle(\mathbf{w}, \nabla f(\mathbf{x}_{k+1})) + \angle(\nabla f(\mathbf{x}_k), \nabla f(\mathbf{x}_{k+1})))\\
&\leq -\cos(\angle(\mathbf{w}, \nabla f(\mathbf{x}_{k+1})))\cos(\angle(\nabla f(\mathbf{x}_k), \nabla f(\mathbf{x}_{k+1})))\\
&\qquad + \sin(\angle(\mathbf{w}, \nabla f(\mathbf{x}_{k+1})))\sin(\angle(\nabla f(\mathbf{x}_k), \nabla f(\mathbf{x}_{k+1})))\\
&\leq -\sqrt{1/d}\sqrt{1-(1/4d)} + \sqrt{1-(1/d)}\sqrt{1/4d}\\
&\leq -\sqrt{1/4d}, \qquad (45)
\end{aligned}$$

for any $d \geq 2$. Define $\mathbb{S}^\perp$ to be a shell in $d-1$ dimensions of unit vectors orthogonal to $\mathbf{w}$, ie $\mathbb{S}^\perp = \{\mathbf{u} : \|\mathbf{u}\|_2 = 1, \langle \mathbf{u}, \mathbf{w}\rangle = 0\}$. Now consider a set of unit vectors *close* to $\mathbf{w}$ given by

$$\mathbb{S}' = \{\mathbf{u} = \lambda\mathbf{w} + \sqrt{1-\lambda^2}\ \mathbf{w}^\perp : |\lambda| \in [\sqrt{1-(1/4d)}, 1], \mathbf{w}^\perp \in \mathbb{S}^\perp]\}.$$

We will relate the mean width integral of $\Omega(k)$ and $\Omega(k+1)$ by splitting across $\mathbb{S}'$ and its complement.

$$\begin{aligned}\sigma_d W(\Omega(k+1)) &= \int_{\mathbf{u}\in\mathbb{S}^{d-1}} \ell(\Pi_{\mathbf{u}}(\Omega(k+1)))\ d\mathbf{u}\\ &= \int_{\mathbf{u}\in\mathbb{S}'} \ell(\Pi_{\mathbf{u}}(\Omega(k+1)))\ d\mathbf{u} + \int_{\mathbf{u}\notin\mathbb{S}'} \ell(\Pi_{\mathbf{u}}(\Omega(k+1)))\ d\mathbf{u}\\ &\leq \int_{\mathbf{u}\in\mathbb{S}'} \ell(\Pi_{\mathbf{u}}(\Omega(k+1)))\ d\mathbf{u} + \int_{\mathbf{u}\notin\mathbb{S}'} \ell(\Pi_{\mathbf{u}}(\Omega(k)))\ d\mathbf{u}, \qquad (46)\end{aligned}$$

since $\Omega(k+1) \subset \Omega(k)$. Thus we reduce the second integral to the corresponding integral in the mean width calculation of $\Omega(k)$. The first part of the integral leads to a negative term which we upper bound to obtain Equation (30). Pick any $\mathbf{u} = \lambda\mathbf{w} + \sqrt{1-\lambda^2}\ \mathbf{w}^{\perp} \in \mathbb{S}'$ and $\mathbf{y} \in \Gamma^{k+2}$. Suppose $\lambda \in [\sqrt{1-(1/4d)}, 1]$, $\langle \mathbf{w}, \xi(\mathbf{y})\rangle \geq \sqrt{1-\lambda^2}$. Thus $\langle \mathbf{u}, \xi(\mathbf{y})\rangle \geq 0$. Similarly, $\langle \mathbf{w}, \xi(\mathbf{x}_k)\rangle \leq -\sqrt{1-\lambda^2}$, and so $\langle \mathbf{u}, \xi(\mathbf{x}_k)\rangle \leq 0$. Thus in such directions u,

$$\ell(\Pi_{\mathbf{u}}(\Omega(k))) - \ell(\Pi_{\mathbf{u}}(\Omega(k+1))) \geq \left|\langle \mathbf{x}_k - \mathbf{x}_{k+1}, \mathbf{u}\rangle\right|.$$

This is also true for the corresponding $-\lambda \in [\sqrt{1-(1/4d)}, 1]$ (since $\ell(\Pi_{\mathbf{u}}(\cdot)) = \ell(\Pi_{-\mathbf{u}}(\cdot))$). Thus

$$\int_{\mathbf{u}\in\mathbb{S}'} \ell(\Pi_{\mathbf{u}}(\Omega(k)))\ d\mathbf{u} \geq \int_{\mathbf{u}\in\mathbb{S}'} \ell(\Pi_{\mathbf{u}}(\Omega(k+1)))\ d\mathbf{u} + \underbrace{\int_{\mathbf{u}\in\mathbb{S}'} \left|\langle \mathbf{x}_k - \mathbf{x}_{k+1}, \mathbf{u}\rangle\right| d\mathbf{u}}_{Z}.$$

We now lower bound the second term. To do so, we perform the integration over all values of λ and $\mathbf{v}$ that determine $\mathbf{u}$. Note that if $\mathbf{u} = \lambda\mathbf{w} + \sqrt{1-\lambda^2}\mathbf{v}$, $d\mathbf{u} = (\sqrt{1-\lambda^2})^{d-2} d\mathbf{v} d\lambda$. Then

$$\begin{aligned}Z &= 2\int_{\lambda\in[\sqrt{1-(1/4d)},1]}\int_{\mathbf{v}\in\mathbb{S}^{\perp}} (\sqrt{1-\lambda^2})^{d-2}\left\langle \mathbf{x}_k - \mathbf{x}_{k+1}, \mathbf{u}\right\rangle\ d\mathbf{v}d\lambda\\ &= 2\int_{\lambda\in[\sqrt{1-(1/4d)},1]}\int_{\mathbf{v}\in\mathbb{S}^{\perp}} (\sqrt{1-\lambda^2})^{d-2}(\lambda\left\langle \mathbf{x}_k - \mathbf{x}_{k+1}, \mathbf{w}\right\rangle + \sqrt{1-\lambda^2}\left\langle \mathbf{x}_k - \mathbf{x}_{k+1}, \mathbf{v}\right\rangle)\ d\mathbf{v}d\lambda\\ &= 2\int_{\lambda\in[\sqrt{1-(1/4d)},1]}\int_{\mathbf{v}\in\mathbb{S}^{\perp}} (\sqrt{1-\lambda^2})^{d-2}(\lambda\left\langle \mathbf{x}_k - \mathbf{x}_{k+1}, \mathbf{w}\right\rangle)\ d\mathbf{v}d\lambda \qquad \left(\because \int_{\mathbf{v}\in\mathbb{S}^{\perp}} \mathbf{v}d\mathbf{v} = 0\right)\\ &\geq \left(\frac{2\sigma_{d-1}\left\|\mathbf{x}_k - \mathbf{x}_{k+1}\right\|_2}{\sqrt{4d}}\right)\int_{\lambda\in[\sqrt{1-(1/4d)},1]} \lambda(\sqrt{1-\lambda^2})^{d-2}d\lambda\\ &= \frac{2\sigma_{d-1}\left\|\mathbf{x}_k - \mathbf{x}_{k+1}\right\|_2}{(\sqrt{4d})^d}.\end{aligned}$$

Using this value of Z with Equation (46), we obtain

$$W(\Omega(k+1)) \leq W(\Omega(k)) - \epsilon'\left\|\mathbf{x}_k - \mathbf{x}_{k+1}\right\|_2,$$

where

$$\epsilon' = \frac{Z}{d} = \frac{2\sigma_{d-1}}{\sigma_d(\sqrt{4d})^d} \geq \frac{1}{d(\sqrt{4d})^d} \geq \frac{1}{2^{(4d\log d)-1}}.$$

This implies that Equation (30) holds for $\epsilon = 1/2^{(4d\log d)-1}$, completing the proof. ■

D.3 Proof of Theorem 16

The separability ensures that we are solving d different optimization problems. For every index i, let $\mathbf{X}_i^*$ be the optimal set with respect to g_i. For such a one dimensional quasiconvex function, we showed in Lemma 14 that x follows a self-contracted curve. Thus, it cannot go in the opposite direction of the minima. By continuity in one dimension it clearly cannot overshoot. The length of this direct path is dist $\left((\mathbf{x}_0)_{(i)}, \mathbf{X}_i^*\right)$. Now observe that,

$$\begin{aligned}\int_0^\infty \|\dot{\mathbf{x}}_t\|_2 \ dt &= \int_0^\infty \sqrt{\sum_{j=1}^d (\dot{\mathbf{x}}_t)_j^2} \ dt \\ &\leq \int_0^\infty \sum_{j=1}^d |\left(\dot{\mathbf{x}}_t\right)_j| \ dt \\ &= \sum_{j=1}^d \left|(\mathbf{x}_0)_j - \mathbf{X}_j^*\right| \\ &\leq \sqrt{d} \ \text{dist}\left(\mathbf{x}_0, \mathbf{X}^*\right).\end{aligned}$$

For GD, we have a similar proof. First, notice that the direction of the update is towards $(\mathbf{x}_j^* - (\mathbf{x}_k)_j)$. By quasiconvexity and since $g_j((\mathbf{x}_k)_j) \geq g_j((\mathbf{x}_0)_j)$, $(-\nabla g_j((\mathbf{x}_k)_j)(\mathbf{x}_j^* - (\mathbf{x}_k)_j)) \geq 0$. The only thing we need to show that the GD curve does not overshoot. For an index j let $\mathbf{x}_j^*$ be the closest minimum in $\mathbf{X}_j^*$ to $(\mathbf{x}_0)_j$. Then by LG, $|\nabla g_j((\mathbf{x}_k)_j| \leq L((\mathbf{x}_k)_j - \mathbf{x}_j^*)$. Thus $|\eta \nabla g_j((\mathbf{x}_k)_j| \leq \left|(\mathbf{x}_k)_j - \mathbf{x}_j^*\right|$, and the update cannot cross $\mathbf{x}_j^*$. Consequently for every j: $\sum_{k=0}^\infty ((\mathbf{x}_k)_j - (\mathbf{x}_{k+1})_j) = ((\mathbf{x}_0)_j - \mathbf{x}_j^*)$. Then

$$\begin{aligned}\sum_{k=0}^\infty \|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 &\leq \sum_{k=0}^\infty \|\mathbf{x}_k - \mathbf{x}_{k+1}\|_1 \\ &= \sum_{k=0}^\infty \sum_{j=1}^d |(\mathbf{x}_k)_j - (\mathbf{x}_{k+1})_j| \\ &= \sum_{j=1}^d \sum_{k=0}^\infty |(\mathbf{x}_k)_j - (\mathbf{x}_{k+1})_j| \\ &= \sum_{j=1}^d \left|(\mathbf{x}_0)_j - \mathbf{x}_j^*\right| \\ &\leq \sqrt{d} \left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2,\end{aligned}$$

where in the last inequality we observed that for any vector $\mathbf{u} \in \mathbb{R}^d$, $\|\mathbf{u}\|_1 \leq \sqrt{d} \ \|\mathbf{u}\|_2$. ■

Appendix E. Proof of Theorem 17

We first provide a broad structure of the proof, and then prove the individual claims in separate subsections. The PKL bound is proved first (Section E.1—E.4), and then it is shown that the same function used in the PKL lower bound also leads to a lower in the linear convergence case (Section E.5).

For any dimension $d \geq 6$, we will construct a function $f : \mathbb{R}^d \to \mathbb{R}$ that will be separable over its parameter $\mathbf{x} = (\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \dots \mathbf{x}_{(d)}) \in \mathbb{R}^d$, as per Equation (13):

$$f(\mathbf{x}) = \sum_{i=1}^{d} g(\mathbf{x}_{(i)}).$$

f will be constructed so that its condition number ν will be bounded as $\nu \leq 3d^2$. For this f we will exhibit an $\mathbf{x}_0$ such that GF or GD have a path length lower bounded as $\zeta \text{ (or } \zeta_\eta) \geq \frac{c\sqrt{d}}{\log d} \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*)$ for $c = 1/6$ in the GF case (see Equation (51) and $c = 1/16$ in the GD case (see Equation (53)).

To motivate this construction, we first prove that such a construction will lead to the statement of the theorem. Suppose the prescribed condition number bound κ is such that $\kappa > 3d^2$. Then the condition number ν of the function we constructed is bounded by $\nu \leq \kappa$ and hence the function is in $\mathcal{F}_\kappa$. Then

$$\zeta \text{ (or } \zeta_\eta) \geq \frac{c\sqrt{d}}{\log d} \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*) \geq \min \left\{ \frac{c\sqrt{d}}{\log d}, \frac{c\kappa^{1/4}}{\log \kappa} \right\} \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*),$$

completing the proof for $\kappa > 3d^2$.

On the other hand suppose $\kappa \leq 3d^2$. Since $\kappa \geq 216$, $\sqrt{\kappa} - \sqrt{\kappa/2} > \sqrt{3}$, and so there exists a $\kappa/2 \leq \nu \leq \kappa$ such that $\sqrt{\nu/3}$ is an integer, say d'. Note that $d \geq d' \geq 6$ because $d' = \sqrt{\nu/3} \geq \sqrt{\kappa/6} \geq 6$ (this is the only part of the proof that uses $\kappa \geq 216$). Since f is separable, we can simply ignore $d - d'$ components of $\mathbb{R}^d$ and instead write $f(\mathbf{x}) = \sum_{i=1}^{d'} g(\mathbf{x}_{(i)})$. Via the same construction, the path length and condition number of f will depend on d' as follows:

$$\text{the condition number is at most } \nu \leq \kappa, \text{ so that } f \in \mathcal{F}_\kappa,$$

and the path length is at least

$$\zeta \text{ (or } \zeta_\eta) \geq \frac{c\sqrt{d'}}{\log d'} \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*).$$

Since $216 \leq \kappa \leq 2\nu \leq 6d'^2$, $e \leq 6 \leq \sqrt{\kappa/6} = d'$. Again, since the function $\sqrt{x}/(\log x)$ is increasing in x for $x \geq e^2$, we conclude

$$\begin{aligned}
\frac{\sqrt{d'}}{\log d'} &\geq \frac{(\kappa/6)^{1/4}}{\log(\sqrt{\kappa/6})} \\
&\geq \frac{2(\kappa/6)^{1/4}}{\log \kappa} \\
&\geq \frac{\kappa^{1/4}}{\log \kappa}.
\end{aligned}$$

This leads to the path length bound

$$\zeta \text{ (or } \zeta_\eta) \geq \frac{c\sqrt{d'}}{\log d'} \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*) \geq \min\left\{\frac{c\sqrt{d}}{\log d}, \frac{c\kappa^{1/4}}{\log \kappa}\right\} \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*),$$

completing the proof for $\kappa \leq 3d^2$.

We now exhibit the pathological function g that defines f and prove some properties about it.

E.1 Construction of g

For any dimension $d \geq 6$ we will exhibit the g such that for $f(\mathbf{x}) = \sum_{i=1}^d g(\mathbf{x}_{(i)})$, (a) the condition number ν of f is bounded as $\nu \leq 3d^2$ and (b) the path length for some initial point $\mathbf{x}_0$ (different for GF and GD) is lower bounded by

$$\zeta \geq \frac{\sqrt{d}}{6\log d} \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*) \text{ that is } c = 1/6,$$

for GF and

$$\zeta_\eta \geq \frac{\sqrt{d}}{16\log d} \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*) \text{ that is } c = 1/16,$$

for GD with some $\eta \in [1/2L, 1/L]$. As argued before this will prove the theorem statement.

Define $\delta = 1/d$ and note that since $d \geq 6$, $\delta \leq 0.2$. Define the component function $g : \mathbb{R} \to \mathbb{R}$ as follows:

$$g(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x^2 & \text{if } x \in [0, 0.5] \\ 0.5 - (1-x)^2 & \text{if } x \in [0.5, 1-\delta] \\ (0.5 - \delta^2) + 2\delta(x - (1-\delta)) & \text{if } x \in [1-\delta, \gamma] \\ \alpha + \beta x^2 & \text{if } x \geq \gamma, \end{cases} \tag{47}$$

where

$$\gamma = 1 - \delta + 6\log(1/2\delta),$$
$$\beta = \frac{\delta}{\gamma},$$
$$\alpha = (0.5 - \delta^2) + 2\delta(\gamma - (1-\delta)) - \beta\gamma^2.$$

(These precise values of α, β, γ are required for the function to satisfy differentiability at $x = \gamma$ and PKL at all points. For a reader interested in the broad idea and not the fine details we summarize the rationale behind setting these values: γ is set so that GF or GD with the initial point (to be defined shortly) does not ever access the region $x \geq \gamma$. Yet, to ensure that the function is PKL everywhere, we have quadratic growth away from γ. α, β are set such that the function remains differentiable at $x = \gamma$.)

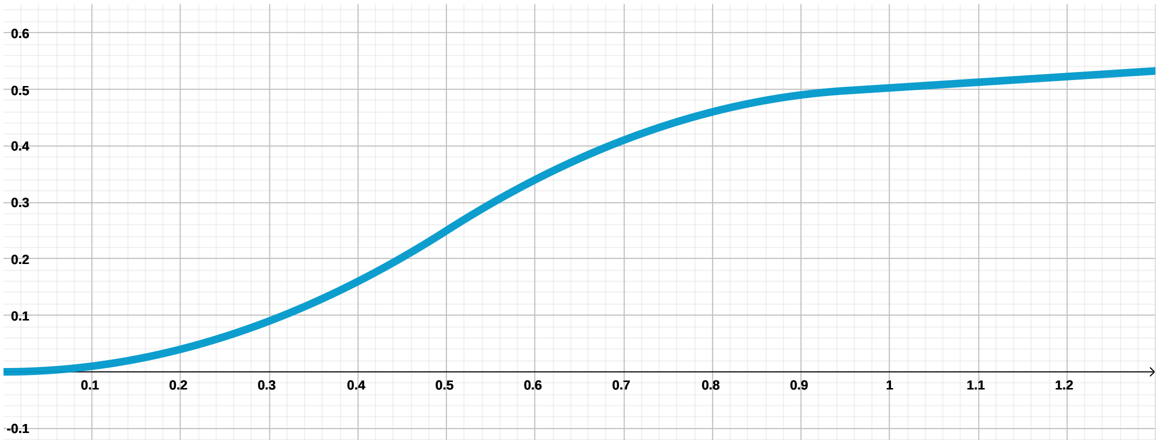


Figure 11: $g(\cdot)$ as defined in Equation (47).

g is plotted in Figure 11. g is everywhere continuously differentiable with the following gradient:

$$\nabla g(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 2x & \text{if } x \in [0, 0.5] \\ 2(1-x) & \text{if } x \in [0.5, 1-\delta] \\ 2\delta & \text{if } x \in [1-\delta, \gamma] \\ 2\beta x & \text{if } x \geq \gamma. \end{cases}$$

Thus f is also continuously differentiable. We consider gradient flow and gradient descent on f. We can write the update equations for each component separately:

$$\text{GF}: \qquad (\dot{\mathbf{x}})_{(i)} = -\nabla g(\mathbf{x}_{(i)}), \tag{48}$$

$$\text{GD}: \qquad (\mathbf{x}^{+})_{(i)} = -\eta \nabla g(\mathbf{x}_{(i)}). \tag{49}$$

We prove a path length lower bound starting from specific initialization points $\mathbf{x}_0$ which are introduced next.

E.2 Identification of $\mathbf{x}_0$

We will need slightly different initialization points for GF and GD to ease computation (the main principle is the same). For GF we will set the following initialization point $\mathbf{x}_0$:

$$(\mathbf{x}_0)_{(i)} = \begin{cases} 0.5 & \text{if } i = 1 \\ (1-\delta) + \delta(i-2)\log(1/2\delta) & \text{if } i > 1. \end{cases}$$

For GD, for ease of computation set $\eta \in [1/4, 1/2]$ ($\equiv [1/2L, 1/L]$) such that

$$k_1 := \log_{1+2\eta}(1/2\delta) = \frac{\log(1/2\delta)}{\log(1+2\eta)}$$

is a natural number. This is possible for $d \geq 6$ since

$$\log(1/2\delta) \;\geq\; 1.7 \text{ and } \left(\frac{1}{\log(2)}, \frac{1}{\log(1.5)}\right) \supset (1.5, 2.4) \text{ so that } \log(1/2\delta)(2.4-1.5) \geq 1.$$

Observe that $k_1 \leq 3\log(1/2\delta)$. Given this η and k_1, define the following initialization point $\mathbf{x}_0$ for GD:

$$(\mathbf{x}_0)_{(i)} = \begin{cases} 0.5 & \text{if } i = 1 \\ (1-\delta) + 2\eta k_1 \delta(i-2) & \text{if } i > 1. \end{cases}$$

The GD and GF bounds follow the same technique but the precise computations are slightly different. Thus we write the GD and GF bounds separately in the following subsections. The main idea behind the staggering us discussed in the main paper and illustrated in Figure 5. Because of the staggering, in every consecutive iterate, a single component goes from value 0.5 to 0.0 leading a large path length.

E.3 GF Analysis

We make the following observations about the function f and the initial point $\mathbf{x}_0$.

(1.1) The distance between the initial and the optimal set is bounded as,

$$\begin{aligned}
\operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*) = \|\mathbf{x}_0 - \mathbf{0}\|_2 &\leq \sqrt{\sum_{i=1}^{d} (1 + (i\delta\log(1/2\delta)))^2} \\
&\leq \sqrt{\sum_{i=1}^{d} (1 + (i\log(d)/d))^2} \\
&\leq \sqrt{2\sum_{i=1}^{d} (1 + (i\log(d)/d)^2)} \\
&= \sqrt{2d + 2\sum_{i=1}^{d} (i\log(d)/d)^2} \\
&= \sqrt{2d + 2\left(\frac{d(d+1)(2d+1)\log^2 d}{6d^2}\right)} \\
&\leq \sqrt{2d + \left(d\log^2 d\right)} && \text{(since } d \geq 6\text{)} \\
&\leq \sqrt{d\log^2 d + d\log^2 d} && \text{(since } d \geq 6\text{)} \\
&= \sqrt{2d}\ \log d. && (50)
\end{aligned}$$

(1.2) $f^* = 0$.

(1.3) The gradients of f are L-Lipschitz with $L = 2$. To see this, first notice that the gradients of g are 2-Lipschitz since they are 2-Lipschitz in each of the pieces in the definition (47), and the derivatives are continuous. Then

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle = \sum_{i=1}^{d} (\nabla g(\mathbf{x}_{(i)}) - \nabla g(\mathbf{y}_{(i)}))(\mathbf{x}_{(i)} - \mathbf{y}_{(i)})$$

$$\begin{aligned} &\leq \sum_{i=1}^{d} 2(\mathbf{x}_{(i)} - \mathbf{y}_{(i)})^2 && (\nabla g \text{ is 2-Lipschitz}) \\ &= 2\left\|\mathbf{x} - \mathbf{y}\right\|_2^2 . \end{aligned}$$

(1.4) f is μ-PKL for $\mu = 2/3d^2$. To see this, first observe that

$$\frac{\left\|\nabla f(\mathbf{x})\right\|_2^2}{2(f(\mathbf{x}) - f^*)} = \frac{\left\|\nabla f(\mathbf{x})\right\|_2^2}{2f(\mathbf{x})} \geq \min_i \left(\min_{\mathbf{x}_{(i)} \in (0, \max_i (\mathbf{x}_0)_{(i)}]} \frac{(\nabla g(\mathbf{x}_{(i)}))^2}{2g(\mathbf{x}_{(i)})} \right) = \min_{x \in (0, \max_i (\mathbf{x}_0)_{(i)}]} \frac{(\nabla g(x))^2}{2g(x)}.$$

We bound the final quantity for each piece in the definition (47) where $g(x) \neq 0$:

- $x \in (0, 0.5]$. $g(x) = x^2$, $(\nabla g(x))^2 = 4x^2$. Thus $(\nabla g(x))^2/2g(x) = 2 \geq \mu$.
- $x \in [0.5, 1-\delta]$. $g(x) \leq 1$, $(\nabla g(x))^2 \geq (2\delta)^2 = 4\delta^2$. Thus $(\nabla g(x))^2/2g(x) \geq 2\delta^2 \geq \mu$.
- $x \in [1-\delta, \gamma]$. Here, $(\nabla g(x))^2 = 4\delta^2$, and

$$\begin{aligned} \max_{x \in [1-\delta, \gamma]} g(x) = g(\gamma) = &= (0.5 - \delta^2) + 12\delta \log(1/2\delta) \\ &\leq 0.5 + 12 \log(d/2)/d \\ &\leq 3 && (\text{for all } x,\ 12\log(x/2)/x \leq 2.5). \end{aligned}$$

  Thus, $(\nabla g(x))^2/2g(x) \geq 2\delta^2/3 = \mu$
- $x \in [\gamma, \infty)$. $(\nabla g(x))^2 = 4\beta^2 x^2$, and $g(x) = (\alpha + \beta x)^2$. The ratio is minimized at $x = \gamma$, where $g(x) = (0.5 - \delta^2) + 12\delta \log(1/2\delta)$:

$$\begin{aligned} \frac{4\beta^2\gamma^2}{g(\gamma)} &= \frac{2\delta^2}{0.5 + 12\log(d/2)/d} \\ &\geq \frac{2\delta^2}{3} && (\text{for all } x,\ 12\log(x/2)/x \leq 2.5) \\ &= \mu. \end{aligned}$$

(1.5) The condition number of f, $\nu = L/\mu \leq 3d^2$.

Consider the time interval $[0, t_1]$ where t_1 is given by,

$$t_1 = \frac{\log\left(1/2\delta\right)}{2}.$$

We make the following computations to determine the value of $\mathbf{x}_{t_1}$:

(2.1) $(\mathbf{x}_{t_1})_{(1)}$: The flow for $(\mathbf{x}_t)_{(1)}$ in the interval $[0, 0.5]$ is given as $(\mathbf{x}_t)_{(1)} = 0.5e^{-2t}$. Thus

$$(\mathbf{x}_{t_1})_{(1)} = 0.5e^{-\log(1/2\delta)} = \delta.$$

(2.2) $(\mathbf{x}_{t_1})_{(2)}$: The flow for $(\mathbf{x}_t)_{(2)}$ in the interval $[0.5, 1-\delta]$ is given as $(\mathbf{x}_t)_{(2)} = 1 - \delta e^{2t}$. As computed below, $(\mathbf{x}_t)_{(2)}$ decreases from $(1-\delta)$ to 0.5 for $t \in [0, t_1]$, and achieves the value 0.5 at t_1:

$$(\mathbf{x}_{t_1})_{(2)} = 1 - \frac{\delta}{2\delta} = 0.5.$$

(2.3) $(\mathbf{x}_{t_1})_{(i)}$, for $i \geq 3$: The flow for $(\mathbf{x}_t)_{(i)}$ in the interval $[1-\delta, \infty)$ is given as $(\mathbf{x}_t)_{(i)} = (1-\delta) + \delta(i-2)\log\left(1+2\delta^2\right) - 2\delta t$. Thus,

$$\begin{aligned}(\mathbf{x}_{t_1})_{(i)} &= (1-\delta) + \delta(i-2)\log\left(1/2\delta\right) - 2\delta t_1 \\ &= (1-\delta) + \delta(i-3)\log\left(1/2\delta\right).\end{aligned}$$

Given this, first we lower bound the path length for the interval $[0, t_1]$, the path length is at least:

$$\|\mathbf{x}_0 - \mathbf{x}_{t_1}\|_2 \geq (\mathbf{x}_0)_{(2)} - (\mathbf{x}_{t_1})_{(2)} = 0.5 - \delta \geq 0.3.$$

Next we perform the same computations for the interval $[t_1, 2t_1]$. In observations (2.1), (2.2), (2.3) we obtained

$$(\mathbf{x}_{t_1})_{(i)} = \begin{cases} \delta & \text{if } i = 1 \\ 0.5 & \text{if } i = 2 \\ (1-\delta) + \delta(i-3)\log\left(1/2\delta\right) & \text{if } i > 2. \end{cases}$$

Compare this to $\mathbf{x}_0$. Observe that $(\mathbf{x}_{t_1})_{(1)} \leq \delta$, and that for $i > 1$, $(\mathbf{x}_{t_1})_{(i)} = (\mathbf{x}_0)_{(i-1)}$. Thus, for the time interval $[t_1, 2t_1]$, $((\mathbf{x}_t)_{(2)}, (\mathbf{x}_t)_{(3)}, \ldots (\mathbf{x}_t)_{(d)})$ follow the exact same dynamics as did $((\mathbf{x}_t)_{(1)}, (\mathbf{x}_t)_{(2)}, \ldots (\mathbf{x}_t)_{(d-1)})$ for the time interval $[0, t_1]$. Continuing in this fashion, more generally for every $s \in \{0, 1, \ldots d-2\}$, the same computations for the interval $[st_1, (s+1)t_1]$ lead to a path length which is at least:

$$\left\|\mathbf{x}_{st_1} - \mathbf{x}_{(s+1)t_1}\right\|_2 \geq (\mathbf{x}_{st_1})_{(s+2)} - (\mathbf{x}_{(s+1)t_1})_{(s+2)} = 0.5 - \delta \geq 0.3.$$

Adding up all path length lower bounds for $s \in \{0, 1, \ldots, d-2\}$ we obtain a lower bound on the overall path length:

$$\begin{aligned}\zeta &\geq 0.3(d-1) \\ &\geq \left(\frac{\sqrt{d}}{6\log d}\right)\left(\sqrt{2d}\log d\right) && \text{(since } d \geq 6\text{)} \\ &\geq \left(\frac{\sqrt{d}}{6\log d}\right)\operatorname{dist}\left(\mathbf{x}_0, \mathbf{X}^*\right), && (51)\end{aligned}$$

where the last inequality uses Equation (50). This concludes the proof for the PKL lower bound in the GF case.

E.4 GD Analysis

Consider the iterates $k \in \{1, 2, \ldots k_1\}$. As noted previously, k_1 is given by

$$k_1 = \log_{1+2\eta}(1/2\delta).$$

First, we make the following observations analogous to the ones made in the GF case (observations 1.1–1.5). The details can be found in the GF analysis.

(3.1) The distance between the initial and the optimal set is bounded as,

$$\begin{aligned} \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*) = \|\mathbf{x}_0 - \mathbf{0}\|_2 &\leq \sqrt{\sum_{i=1}^{d} (1 + (2i\delta k_1))^2} \\ &= \sqrt{2d + 36 \sum_{i=1}^{d} (i \log(d)/d)^2} && (\text{since } k_1 \leq 3\log(1/2\delta)) \\ &= \sqrt{2d + 36\left(\frac{d(d+1)(2d+1)\log^2 d}{6d^2}\right)} \\ &\leq \sqrt{2d + \left(13d\log^2 d\right)} && (\text{since } d \geq 6) \\ &\leq 4\sqrt{d}\ \log d. && (52) \end{aligned}$$

(3.2) $f^* = 0$.

(3.3) The gradients of f are L-Lipschitz with $L = 2$. This is the same observation as (1.3).

(3.4) f is μ-PKL for $\mu = 2/3d^2$. This is the same observation as (1.4).

(3.5) The condition number of f, $\nu = L/\mu \leq 3d^2$.

We make the following computations to determine the value of $\mathbf{x}_{k_1}$:

(4.1) $(\mathbf{x}_{k_1})_{(1)}$:

$$(\mathbf{x}_{k_1})_{(1)} \leq (\mathbf{x}_1)_{(1)} = 0.5 - \eta \leq 0.25.$$

(4.2) $(\mathbf{x}_{k_1})_{(2)}$: The iterates for $(\mathbf{x}_k)_{(2)}$ for $k \in \{1, 2, \ldots k_1\}$ are $(\mathbf{x}_k)_{(2)} = 1 - (1 + 2\eta)^k\delta$. As computed below, $(\mathbf{x}_k)_{(2)}$ decreases from $(1 - \delta)$ to 0.5 for $k \in \{1, 2, \ldots k_1\}$ and achieves the value 0.5 at k_1:

$$(\mathbf{x}_{k_1})_{(2)} = 1 - \frac{\delta}{2\delta} = 0.5.$$

(4.3) $(\mathbf{x}_{k_1})_{(i)}$, for $i \geq 3$: The updates for $(\mathbf{x}_k)_{(i)}$ in the interval $[1 - \delta, \infty)$ are given as $(\mathbf{x}_k)_{(i)} = (1 - \delta) + 2\eta\delta(i - 2)k_1 - 2\eta\delta k$. Thus,

$$\begin{aligned} (\mathbf{x}_{k_1})_{(i)} &= (1 - \delta) + 2\eta\delta(i - 2)k_1 - 2\eta\delta k_1 \\ &= (1 - \delta) + 2\eta\delta(i - 3)k_1. \end{aligned}$$

Given this, first we lower bound the path length for the iterates $\{1, 2, \dots k_1\}$. The path length is at least:

$$\|\mathbf{x}_0 - \mathbf{x}_{k_1}\|_2 \geq (\mathbf{x}_0)_{(2)} - (\mathbf{x}_{k_1})_{(2)} = 0.5 - \delta \geq 0.3.$$

Next we perform the same computations for the interval $[t_1, 2t_1]$. Through observations (4.1), (4.2), (4.3) we obtained $(\mathbf{x}_{k_1})_{(1)} \in [0, 0.5 - \eta]$ and

$$(\mathbf{x}_{k_1})_{(i)} = \begin{cases} 0.5 & \text{if } i = 2 \\ (1-\delta) + 2\eta\delta(i-3)k_1 & \text{if } i > 2. \end{cases}$$

Compare this to $\mathbf{x}_0$. Observe that for $i > 1$, $(\mathbf{x}_{k_1})_{(i)} = (\mathbf{x}_0)_{(i-1)}$. Thus, for the iterates $\{k_1 + 1, \dots 2k_1\}$, $((\mathbf{x}_k)_{(2)}, (\mathbf{x}_k)_{(3)}, \dots (\mathbf{x}_k)_{(d)})$ follow the exact same dynamics as did $((\mathbf{x}_k)_{(1)}, (\mathbf{x}_k)_{(2)}, \dots (\mathbf{x}_k)_{(d-1)})$ for the time interval $\{1, 2, \dots k_1\}$. Continuing in this fashion, more generally for every $s \in \{0, 1, \dots d-2\}$, the same computations for the interval $\{sk_1 + 1, sk_1 + 2, \dots (s+1)k_1\}$ lead to a path length bound at least as large as:

$$\left\|\mathbf{x}_{sk_1} - \mathbf{x}_{(s+1)k_1}\right\|_2 \geq (\mathbf{x}_{sk_1})_{(s+2)} - (\mathbf{x}_{(s+1)k_1})_{(s+2)} = 0.5 - \delta \geq 0.3.$$

Adding up all path length lower bounds for $s \in \{0, 1, \dots, d-2\}$ we obtain a lower bound on the overall path length:

$$\begin{aligned} \zeta_\eta &\geq 0.3(d-1) \\ &\geq \left(\frac{\sqrt{d}}{16 \log d}\right)\left(4\sqrt{d}\ \log d\right) \\ &\geq \left(\frac{\sqrt{d}}{16 \log d}\right) \operatorname{dist}\left(\mathbf{x}_0, \mathbf{X}^*\right), \end{aligned} \tag{53}$$

where the last inequality uses the bound (52). This concludes the proof for the PKL lower bound in the GD case.

E.5 Lower Bound Under Linear Convergence

We compute the linear convergence constant c for the function f and relate it to the path length lower bounds shown in Section E.3 (GF) and Section E.4 (GD).

GF analysis. Suppose $\{x_t\}_{t \in \mathbb{R}_0^+}$ follows the dynamics $\dot{x} = -\nabla g(x)$ with some initial point. Note that the optimal set is $X^* = [-\infty, 0]$. Then for any $x > 0$,

$$\frac{\dot{x}}{\operatorname{dist}(x, X^*)} = \frac{\dot{x}}{x} = -\frac{\nabla g(x)}{x} \geq -\min(2, 2\delta, 2\delta/\gamma, 2\beta).$$

This corresponds to the definition $\nabla g(x)$ in different regions. The minimum above is given by $2\delta/\gamma$ since $\delta \leq 1$ and $\gamma > 1$ for $d = 6$. We compute $2\delta/\gamma = 2d^{-1}(1 + 6\log(d/2))^{-1} \geq (4d \log d)^{-1}$. Thus linear convergence holds for every component of f with $A = 1$ and $c = (4d \log d)^{-1}$; consequently linear convergence also holds for f with constants $(1, c)$. For the $\mathbf{x}_0$ chosen in Section E.2, we showed a lower bound on the path length in terms of d (Section E.3) that can be translated into a lower bound in terms of c as follows:

$$\zeta \geq \left(\frac{\sqrt{d}}{6\log d}\right) \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*) \geq \left(\frac{\sqrt{1/c}}{12\log^{1.5}(1/c)}\right) \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*).$$

This concludes the analysis in terms of the linear convergence constant c for GF. Since $d \geq 6$, this construction admits $c \in (0, (4 \cdot 6\log 6)^{-1}) \supseteq (0, 0.023)$.

GD analysis. As identified in Section E.2, $\eta \geq 1/4$. Consider any $x > 0$ (corresponding to one component of the iterates) and note that the optimal set is $X^* = [-\infty, 0]$. Then for an update $x^+ \leftarrow x - \eta\nabla g(x)$ we have

$$\frac{x^+ - x}{\operatorname{dist}(x, X^*)} = \frac{-\eta\nabla g(x)}{x} \geq -\eta(4d\log d)^{-1} \geq (16d\log d)^{-1}.$$

(The lower bound for $\frac{\nabla g(x)}{x}$ is shown in the GF computation above and the final inequality uses $\eta \geq 1/4$.) Thus linear convergence holds for every component of f with $c = (16d\log d)^{-1}$; consequently linear convergence also holds for f with this value of c. For the $\mathbf{x}_0$ chosen in Section E.2, we showed a lower bound on the path length in terms of d (Section E.4) that can be translated into a lower bound in terms of c as follows:

$$\zeta \geq \left(\frac{\sqrt{d}}{16\log d}\right) \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*) \geq \left(\frac{\sqrt{1/c}}{64\log^{1.5}(1/c)}\right) \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*).$$

This concludes the analysis in terms of the linear convergence constant c for GD. Since $d \geq 6$, this construction admits $c \in (0, (4 \cdot 6\log 6)^{-1}) \supseteq (0, 0.0058)$.

■

Appendix F. Proof of Theorem 18

For any dimension d, we will construct a separable quadratic function in d dimensions:

$$f(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{d} a_i \mathbf{x}_{(i)}^2,$$

where $a_i = \omega^{d-i}$ for $\omega = 11$. Since f is a separable quadratic function, it is also a separable quasiconvex function. The condition number of f is $\nu = \omega^{d-1}$, so that $\log\nu = (d-1)(\log\omega)$ and in particular $\sqrt{\log\nu} \leq \sqrt{d(\log\omega)}$. Also $\mathbf{X}^* = \{\mathbf{0}\}$. In what follows we will identify an $\mathbf{x}_0$ such that the path length for GF satisfies $\zeta \geq 0.7\sqrt{d}\operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*)$ and the path length for GD for $\eta = 1/2a_1 = 1/2L$ satisfies $\zeta_\eta \geq 0.5\sqrt{d}\operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*)$. Before we illustrate this construction, we prove that such a construction would lead to the results in Equations (14) and (15).

For the given value of κ and d, we consider two cases. Suppose $\log\kappa \geq (d-1)(\log\omega)$. Then, the above function with condition number ν satisfies $\nu \geq \kappa$ so that $f \in \mathcal{Q}_\kappa$. The construction ensures that

$$\zeta \text{ (or } \zeta_\eta) \geq c\sqrt{d}\operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*),$$

for appropriate values of c which implies Equations (14) and (15).

On the other hand, suppose $\log\kappa \leq (d-1)(\log\omega)$. Then we identify the largest $2 \leq d' \leq d$ such that $\log\kappa \geq (d'-1)(\log\omega)$ (since $\kappa \geq 5$, this is possible). Note that this means $\log\kappa \leq d'(\log\omega)$. Next we instantiate the construction for dimension d' instead of d, that is

$$f(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{d'} a_i \mathbf{x}_{(i)}^2,$$

with $a_i = \omega^{d'-i}$. For this f, the condition number $\omega^{d'-1}$ is smaller than κ so that $f \in \mathcal{Q}_\kappa$. The path length is at least

$$\begin{aligned} \zeta \text{ (or } \zeta_\eta) &\geq c\sqrt{d'}\ \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*) \\ &\geq c\sqrt{\log\kappa/\log\omega}\ \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*) \\ &= (c/\sqrt{\log\omega})\sqrt{\log\kappa}\ \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*). \end{aligned}$$

As we will prove, in the GF case, $c = 0.7$ so that $c/\sqrt{\log 11} \geq 0.45$, and in the GD case, $c = 0.5$ so that $c/\sqrt{\log 11} \geq 0.3$. This leads to the bounds in Equations (14) and (15).

Now we analyze the path length for f in d dimensions and show the bound ζ (or ζ_η) $\geq c\sqrt{d}\ \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*)$ for a specific $\mathbf{x}_0$.

F.1 GF Analysis

Define $\mathbf{x}_0 = \mathbf{1}_d$ so that $\operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*) = \sqrt{d}$. Set $\delta = 0.07$. Observe that the GF dynamics lead to $(\mathbf{x}_t)_{(i)} = e^{-a_i t}$. Consider the time steps $\{t_i\}_{i=0}^d$ where $t_0 = 0$ and $t_i = \log(1/\delta)/a_i$. Observe the following:

1. For every $i \in [d]$, $(\mathbf{x}_{t_i})_{(i)} = \delta$.

2. For every $i \in [d-1]$,

$$\begin{aligned} (\mathbf{x}_{t_i})_{(i+1)} &= e^{-a_{i+1}t_i} \\ &= e^{-\log(1/\delta)/\omega} \\ &= \delta^{1/\omega}. \end{aligned}$$

Thus splitting the path length integral in these time steps,

$$\begin{aligned} \zeta &= \int_0^\infty \|\dot{\mathbf{x}}_t\|_2\ dt \\ &\geq \sum_{i=1}^d \int_{t_{i-1}}^{t_i} \|\dot{\mathbf{x}}_t\|_2\ dt \\ &\geq \sum_{i=1}^d \left\|\mathbf{x}_{t_i} - \mathbf{x}_{t_{i-1}}\right\|_2 \\ &\geq \sum_{i=1}^d (\mathbf{x}_{t_{i-1}} - \mathbf{x}_{t_i})_{(i)} \end{aligned}$$

$$
\begin{aligned}
&\geq \sum_{i=1}^{d} (\delta^{1/\omega} - \delta) \\
&= d(\delta^{1/\omega} - \delta) \\
&= \sqrt{d}\ (\delta^{1/\omega} - \delta) \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*) \\
&\geq 0.7\sqrt{d} \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*) && \text{(plugging in values of } \delta \text{ and } \omega\text{).}
\end{aligned}
$$

F.2 GD Analysis

Define $\mathbf{x}_0 = \mathbf{1}_d$ so that $\operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*) = \sqrt{d}$. Set $\delta = e^{-3}$. Let $\eta = 1/2a_1 = 1/2L$. Observe that the GD iterates are $(\mathbf{x}_k)_{(i)} = (1 - \eta a_i)^k$. Consider the iterates $\{k_i\}_{i=0}^{d}$ where $k_0 = 0$ and $k_i = a_1 \log(1/\delta)/a_i$ which are integers. Observe the following:

1. For every $i \in [d]$,

$$
\begin{aligned}
(\mathbf{x}_{t_i})_{(i)} &= (1 - \eta a_i)^{k_i} \\
&\leq e^{-k_i \eta a_i} \\
&\leq \sqrt{\delta}.
\end{aligned}
$$

2. For every $i \in [d-1]$,

$$
\begin{aligned}
(\mathbf{x}_{t_i})_{(i+1)} &= (1 - \eta a_{i+1})^{k_i} \\
&\geq e^{-2a_{i+1}k_i} && (1 - x \geq e^{-2x} \text{ for } x \leq 0.5) \\
&= e^{-\log(1/\delta)/\omega} \\
&= \delta^{1/\omega}.
\end{aligned}
$$

Thus splitting the path length sum in these iterates,

$$
\begin{aligned}
\zeta_\eta &= \sum_{k=0}^{\infty} \|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 \\
&\geq \sum_{i=0}^{d-1} \sum_{k=k_i}^{k_{i+1}-1} \|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 \\
&\geq \sum_{i=0}^{d-1} \left\|\mathbf{x}_{k_i} - \mathbf{x}_{k_{i+1}}\right\|_2 \\
&\geq \sum_{i=0}^{d-1} (\mathbf{x}_{k_i} - \mathbf{x}_{k_{i+1}})_{(i+1)} \\
&\geq \sum_{i=0}^{d-1} (\delta^{1/\omega} - \sqrt{\delta}) \\
&= d(\delta^{1/\omega} - \sqrt{\delta}) \\
&= \sqrt{d}\ (\delta^{1/\omega} - \sqrt{\delta}) \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*)
\end{aligned}
$$

$$\geq 0.5\sqrt{d}\ \mathrm{dist}\left(\mathbf{x}_0, \mathbf{X}^*\right),$$

plugging in values of δ and ω. ■

Appendix G. Lower Bound Simulations

In this section, we empirically simulate the lower bound constructions shown in Section 4 and draw some insights into the tightness of these bounds.

G.1 PKL

In this simulation, we assess the PKL lower bound construction of Theorem 17. For the constructed lower bound function f in the proof, it is not entirely evident if the computations are tight and that the final path length bound is $\Omega(\kappa^{1/4}/\log\kappa)$ (perhaps it has a larger dependence like $\Omega(\sqrt{\kappa})$). To assess this, we simulate GD with the f described in Section E.1 and numerically compute the path length. By varying the dimension $d \in [e^2, e^9]$, we obtain a range of values for the condition number of f: $\kappa \in [28.6, 3.29 \cdot 10^7]$. For each of these functions with different κ values, the step-size for GD and the initialization point are set as described in Section E.2. The path length is computed by adding the lengths of all the updates, until $\|\mathbf{x}_k\|_2 \leq 10^{-6}$. At this point the remaining distance to the origin $\|\mathbf{x}_k\|_2$ is added to the path length computation. Finally, μ is computed as the 'effective' PKL constant from the iterates actually seen while running the simulation:

$$\mu = \max_{t\in\{0,1,\ldots k\}} \frac{\|\nabla f(\mathbf{x}_t)\|_2^2}{2f(\mathbf{x}_t)}.$$

Consequently, since $L = 2$, we obtain $\kappa = 2/\mu$. The observed path length ratio $\zeta_\eta/\mathrm{dist}\left(\mathbf{x}_0, \mathbf{X}^*\right)$ is plotted against the effective κ in Figure 12.

The X-axis in the figure corresponds to $\kappa \in [28.6, 3.29 \cdot 10^7]$. From the figure, we observe that the path length ratio is $\approx 3\kappa^{1/4}/\log\kappa$ across various values of the effective κ. Note that for the given range of κ, $\log\log\kappa \in [1.21, 2.85]$. Thus if we were to modify the X-axis with a $\log\log\kappa$ factor, the dependence would no longer be linear. We infer visually that $\Omega(\kappa^{1/4}/\log\kappa)$ is the right dependence on κ including $\log\log\kappa$ factors. However, the constants in the lower bound can be improved since Theorem 17 only shows $\kappa \geq \kappa^{1/4}/16\log\kappa$.

G.2 Quadratics

For quadratics, we showed a $\Theta(\log(\sqrt{\kappa}))$ upper bound (Theorem 10) and lower bound (Theorem 18). We simulate the lower bound example described in the proof of Theorem 18 to assess the constants in these bounds. Consider gradient descent and gradient flow with

$$f(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{d} a_i \mathbf{x}_{(i)}^2,$$

where $d = 150$ and a_i is set similar to the construction in the proof of Theorem 18: for every $i \in [d]$, $a_i = \omega^{d-i}$. To vary the condition number $\kappa = \omega^{d-1}$, we vary $\omega \in [1.00008, 2]$ (instead

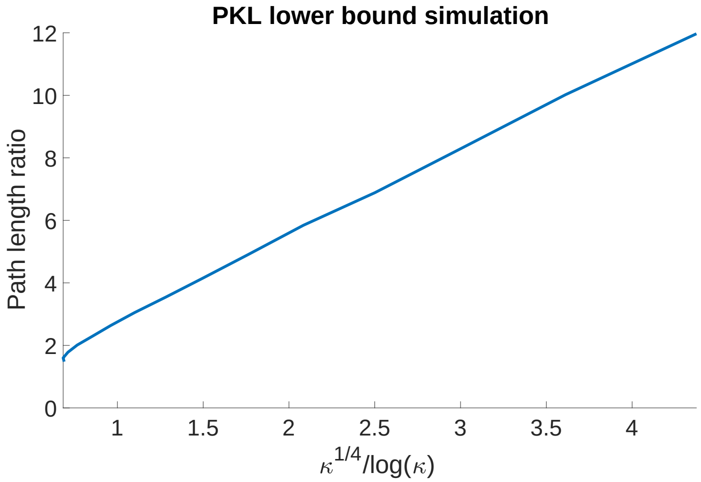


Figure 12: Dependence of the path length ratio $\zeta_\eta / \operatorname{dist}(\mathbf{x}_0, \mathbf{X}^*)$ on the condition number for the PKL lower bound construction in the proof of Theorem 17. We infer that the dependence is $\Omega(\kappa^{1/4}/\log\kappa)$, including $\log\log\kappa$ factors.

of varying d as we did in the proof of Theorem 18). This leads to $\kappa \in [9.553{\cdot}10^4, 6.2807{\cdot}10^{11}]$. For the initialization point, we set $(\mathbf{x}_0)_i = 1$ for every $i \in [d]$.

The GF path length is computed by performing the path length integral numerically (since we know $\mathbf{x}_t$ at each point in the case of quadratics). We use MATLAB's numerical integration function `integral`, with the `AbsTol` parameter set to 10^{-50}. The GD path length is computed by simulating GD with step size $\eta = 1/2a_1$ and adding the lengths of all the updates, until every component except the last one is smaller than 10^{-2}, ie $\max_{i \neq d}(\mathbf{x}_k)_{(i)} < 10^{-2}$. At this point the remaining distance to the origin $\|\mathbf{x}_k\|_2$ is added to the path length computation. We use this heuristic since GD convergence is quite slow when κ is very large (the asymptotic convergence rate with respect to κ is $\mathcal{O}(\kappa)$). Further, we limit the largest κ in the GD plot to around $2 \cdot 10^7$. The results are plotted in Figure 13.

Further, to verify that our lower bound construction is non-trivial, we compare the path length of the construction in Theorem 18 with a randomized construction that has the a_i values sampled as follows: $a_1 = 1, a_d = 1/\kappa, a_i \sim \text{Unif}(1/\kappa, 1)$, where $\kappa = \omega^{d-1}$ for $\omega \in [1.00008, 2]$ as in the previous construction. For the initialization point, every $(\mathbf{x}_0)_i$ is sampled independently from $\text{Unif}(0,1)$ and then normalized so that $\|\mathbf{x}_0\|_2 = \sqrt{d}$. The results are plotted in Figure 14.

We do not show the results with averaging across multiple runs for the randomized construction, but we observed behavior similar to Figure 14 across runs. Thus the randomized quadratic construction does not have a $\sqrt{\log\kappa}$ dependence for its path length. This indicates that the lower bound construction in Theorem 18 is non-trivial. At the same time, Figure 13 shows that the upper bound overestimates the path length of this lower bound construction. Given these results, we conjecture that the constants in the upper bound can be improved.

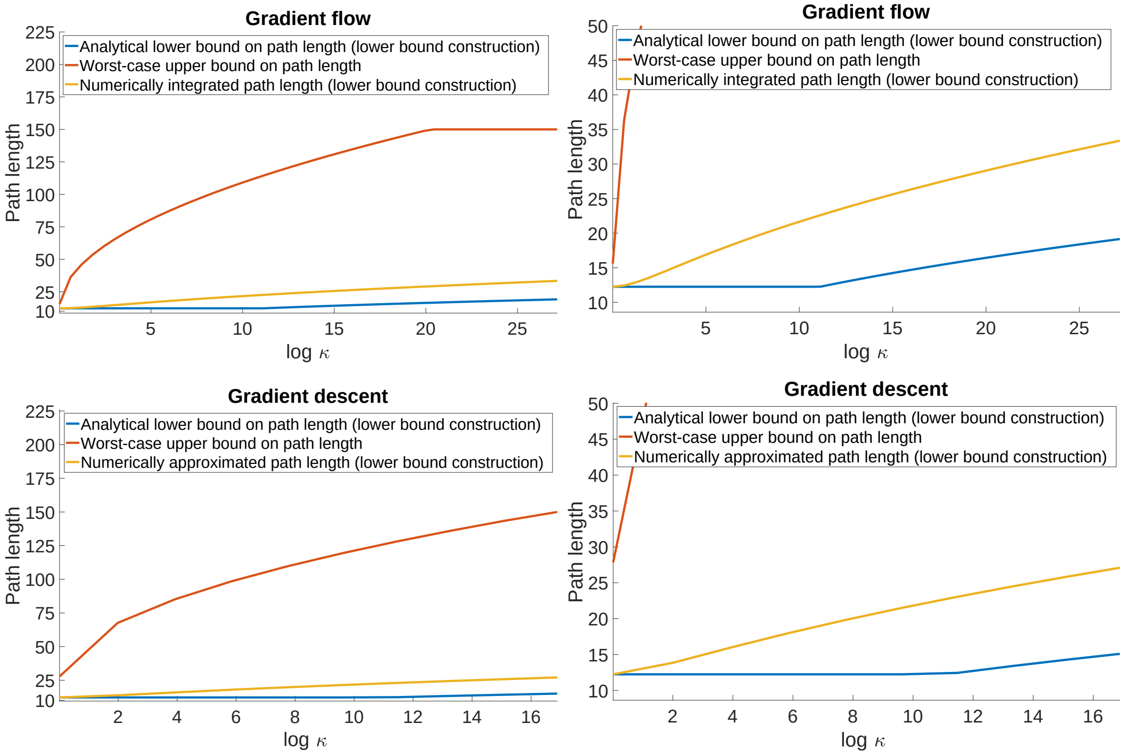


Figure 13: Path length of GF (top) and GD (bottom) for quadratic objectives. The lower limit on the X-axis is the same for all plots. The plots on the right are 'zoomed in' versions of the plots on the left; they focus on a smaller range for the Y-axis. The lower bound refers to the specific construction in Theorem 18, described in Section G.2. The shortest path has length $\sqrt{d} = \sqrt{150} \approx 12.25$, and the κ-independent upper bound on the path length is d.

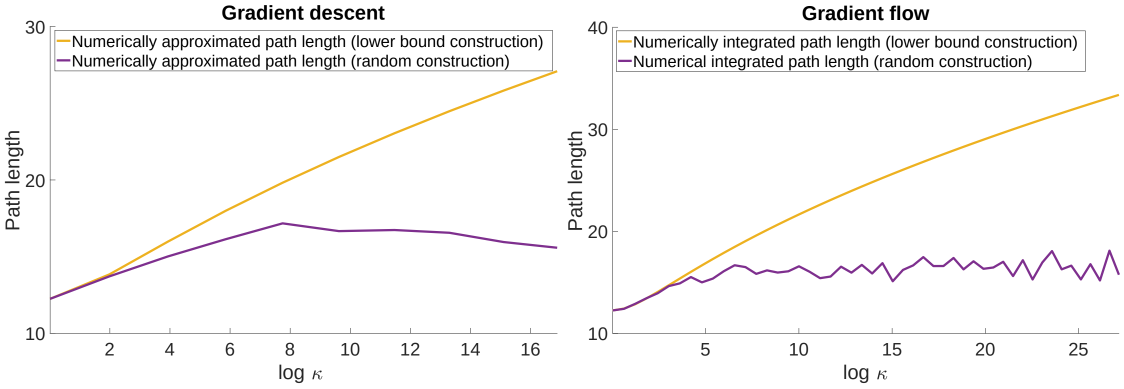


Figure 14: The lower bound construction of Theorem 18 has larger path length than the randomized construction defined in Appendix G.2.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252, 2019.

CJ Argue, Sébastien Bubeck, Michael B Cohen, Anupam Gupta, and Yin Tat Lee. A nearly-linear bound for chasing nested convex bodies. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 117–122. SIAM, 2019.

Hedy Attouch and Juan Peypouquet. The rate of convergence of nesterov's accelerated forward-backward method is actually faster than 1/kˆ2. *SIAM Journal on Optimization*, 26(3):1824–1834, 2016.

Mordecai Avriel, Walter E Diewert, Siegfried Schaible, and Israel Zang. *Generalized concavity*. SIAM, 2010.

Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statististics*, 45(1): 77–120, 02 2017.

Jérôme Bolte and Edouard Pauwels. Curiosities and counterexamples in smooth convex optimization. *arXiv preprint arXiv:2001.07999*, 2020.

Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

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

Sébastien Bubeck, Yuanzhi Li, Haipeng Luo, and Chen-Yu Wei. Improved path-length regret bounds for bandits. In *Conference on Learning Theory*, pages 508–528, 2019.

Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

Aris Daniilidis, Olivier Ley, and Stéphane Sabourau. Asymptotic behaviour of self-contracted planar curves and gradient orbits of convex functions. *Journal de mathématiques pures et appliquées*, 94(2):183–199, 2010.

Aris Daniilidis, Guy David, Estibalitz Durand-Cartagena, and Antoine Lemenant. Rectifiability of self-contracted curves in the Euclidean space and applications. *The Journal of Geometric Analysis*, 25(2):1211–1239, 2015.

Aris Daniilidis, Robert Deville, Estibalitz Durand-Cartagena, and Ludovic Rifford. Self-contracted curves in Riemannian manifolds. *Journal of Mathematical Analysis and Applications*, 457(2):1333–1352, 2018.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685, 2019.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

Estibalitz Durand-Cartagena and Antoine Lemenant. Self-contracted curves are gradient flows of convex functions. *Proceedings of the American Mathematical Society*, 147(6): 2517–2531, 2019.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476, 2018.

Werner Fenchel. Convex cones, sets, and functions. Lecture Notes. Princeton University, 1953.

Robert M Freund, Paul Grigas, and Rahul Mazumder. Condition number analysis of logistic regression, and its implications for standard first-order solution methods. *arXiv preprint arXiv:1810.08727*, 2018.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *European Control Conference*, pages 310–315. IEEE, 2015.

Cristóbal Guzmán and Arkadi Nemirovski. On lower complexity bounds for large-scale smooth convex optimization. *Journal of Complexity*, 31(1):1–14, 2015.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234, 2016.

Trevor Hastie, Jonathan Taylor, Robert Tibshirani, and Guenther Walther. Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics*, 1:1–29, 2007.

M. Hohenwarter, M. Borcherds, G. Ancsin, B. Bencze, M. Blossier, A. Delobelle, C. Denizet, J. Éliás, Á Fekete, L. Gál, Z. Konečný, Z. Kovács, S. Lizelfelner, B. Parisse, and G. Sturr. GeoGebra 4.4, December 2013. `http://www.geogebra.org`.

Hamed Karimi, Julie Nutini, and Mark 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*, pages 795–811. Springer, 2016.

Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. In *Annales de L'Institut Fourier*, volume 48, pages 769–783, 1998.

Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, 176(1-2):311–337, 2019.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, pages 6389–6399, 2018.

Stanislaw Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les Équations aux Dérivées Partielles*, 117:87–89, 1963.

Paolo Manselli and Carlo Pucci. Maximum length of steepest descent curves for quasi-convex functions. *Geometriae Dedicata*, 38(2):211–227, 1991.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *Annals of Statististics*, 46(6A):2747–2774, 12 2018.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pages 4951–4960, 2019.

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

Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

LA Santaló. Convex regions on the n-dimensional spherical surface. *Annals of Mathematics*, pages 448–459, 1946.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2019.

Eugene Stepanov and Yana Teplitskaya. Self-contracted curves have finite length. *Journal of the London Mathematical Society*, 96(2):455–481, 2017.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pages 1263–1291, 2018.

Xingyu Zhou. On the Fenchel duality between strong convexity and Lipschitz continuous gradient. *arXiv preprint arXiv:1803.06573*, 2018.