SIAM J. MATH. DATA SCI.
Vol. 6, No. 3, pp. 602–626

Gradient Descent in the Absence of Global Lipschitz Continuity of the Gradients[*]

Vivak Patel[†] and Albert S. Berahas[‡]

Abstract. Gradient descent (GD) is a collection of continuous optimization methods that have achieved immeasurable success in practice. Owing to data science applications, GD with diminishing step sizes has become a prominent variant. While this variant of GD has been well studied in the literature for objectives with globally Lipschitz continuous gradients or by requiring bounded iterates, objectives from data science problems do not satisfy such assumptions. Thus, in this work, we provide a novel global convergence analysis of GD with diminishing step sizes for differentiable nonconvex functions whose gradients are only locally Lipschitz continuous. Through our analysis, we generalize what is known about gradient descent with diminishing step sizes, including interesting topological facts, and we elucidate the varied behaviors that can occur in the previously overlooked divergence regime. Thus, we provide a general global convergence analysis of GD with diminishing step sizes under realistic conditions for data science problems.

Key words. gradient descent, diminishing step sizes, convergence, divergence

MSC codes. 90C26, 68T99, 68W40

DOI. 10.1137/22M1527210

1. Introduction.

Proposed nearly two centuries ago [14, 15, 28, 37], gradient descent is a set of canonical continuous optimization methods that have achieved immeasurable success in a plethora of applications (e.g., [9, 19, 29]). Owing to their prominence and utility in data science, gradient descent methods have continued to grow in variety, and their theory has received renewed interest by the optimization and data science communities for problems in this area (e.g., [20, 22, 35, 46]). In particular, gradient descent with prescheduled step sizes has become popular owing to the additional expense of using line search techniques for data science problems. Correspondingly, the theory of gradient descent with prescheduled step sizes has grown in a number of interesting directions, including new local convergence rate analyses (e.g., [20, 33]) and saddle-point avoidance analyses (e.g., [21, 30, 36]).

That said, the more fundamental global convergence analysis of gradient descent with pre-scheduled step sizes has lagged owing to two challenges. First, gradient descent with pre-scheduled step sizes does not guarantee a monotonic reduction in the objective function (cf. Armijo's method [1]), which is the key ingredient used to analyze such methods via

[*] Received by the editors October 6, 2022; accepted for publication (in revised form) February 15, 2024; published electronically July 5, 2024.
https://doi.org/10.1137/22M1527210
Funding: This work is supported by the Wisconsin Alumni Research Foundation and the Office of Naval Research.

[†] Department of Statistics, University of Wisconsin – Madison, Madison, WI 53706 USA (vivak.patel@wisc.edu, http://vivakpatel.org).

[‡] Department of Industrial and Operations Engineering, University of Michigan – Ann Arbor, Ann Arbor, MI 48109 USA (albertberahas@gmail.com, https://aberahas.engin.umich.edu).

Zoutendjik's approach [60]. Second, because of the nonconvexity of common data science problems,[1] the analysis of gradient descent cannot leverage uniform continuity of the gradient function or global Lipschitz continuity of the gradient function or presuppose that its iterates remain in a bounded region for a function with a locally Lipschitz continuous gradient,[2] which are instrumental assumptions for overcoming the previous challenge [51, 52]. As a result of the latter challenge, typical analysis approaches for global convergence of gradient descent fall short (see subsection 2.3 for an overview). In fact, even the new vogue for analysis in machine learning, *the continuous approach* [4, 5, 39, 41], falls short because this approach requires compactness of the image space of the iterates in [4, Theorem 3.2], boundedness of iterates [23, Theorem 2], or global Lipschitz continuity of the gradient of the objective function [41, Assumption 1].[3] To summarize, to the best of our knowledge, existing global convergence analyses of gradient descent with diminishing step sizes do not apply to canonical, nonconvex, differentiable data science problems.

To address this shortcoming, we generalize recently developed techniques for the analysis of stochastic gradient descent [48, 49, 50] to analyze gradient descent with diminishing step sizes for nonconvex optimization problems that are bounded from below and whose gradient is *locally Lipschitz continuous*, which are more realistic assumptions for canonical data science problems [49, section A]. Our analysis has several important contributions.

1. First, we present a novel upper-bound model, which can be used under milder assumptions that are appropriate for data science problems (see subsection 2.3 for a discussion and Lemma 3.1 for the result). This upper-bound model is directly useful in analyzing many other algorithms for unconstrained optimization, and the strategies used to prove the result seem useful for analyzing algorithms for constrained optimization.
2. Second, our analysis provides counterexamples to what is known about gradient descent with diminishing step sizes. Specifically, previous results (e.g., [6, Proposition 1.2.4]) showed that, under a global Lipschitz continuity assumption on the gradient, the iterates tend to a region where the gradient is zero; the objective function converges to a finite limit; and, if the iterates remain bounded, then the iterates converge to a stationary point. Our analysis, under the more realistic local Lipschitz continuity assumption on the gradient, offers a correction to this view—that the gradient function can remain bounded away from zero and the objective function can diverge (see explicitly constructed examples in section 4).
3. Our analysis addresses a preliminary question about gradient descent and nonconvexity: Given a relatively arbitrary objective function, can its nonconvexity cause gradient

[1] Canonical data science problems such as Poisson regression, linear three-or-more–layer feed-forward networks, and linear three-or-more time horizon recurrent networks fail to possess globally Lipschitz continuous gradients or uniformly continuous gradients when trained using standard loss functions [49, section 1].

[2] If the iterates remain in a bounded region, then compactness and the local Lipschitz continuity condition would imply a global Lipschitz continuous constant in the bounded region.

[3] As of the submission of this work, the continuous approach has received a great boon owing to the work of [32]. Roughly, if continuous gradient descent trajectories are bounded and a clever generalization of the Kudryka–Łojasiewicz inequality holds (which is shown to hold for a broad class of objective functions), then gradient descent with a sufficiently small step size will generate bounded iterates [32, Corollary 1]. While it is true that the boundedness of continuous gradient descent trajectories assumption retains the flavor of boundedness of the iterates, it is a noteworthy improvement. We discuss this again in subsection 3.3.

descent to behave erratically in a region? Despite the general nonconvexity allowed by our assumptions, we show that the limit supremum and limit infimum of the objective function evaluated at the iterates must tend to each other if the iterates remain in a region for long enough, even if they eventually escape; we also show that the limit of the gradient function evaluated at the iterates must tend to zero if the iterates remain in a region for long enough, even if they eventually escape (see Theorem 3.10). A more interesting question is whether such a statement holds uniformly over important subsets of nonconvex objective functions of the ones considered here.[4] Our analysis at least gives hope that such a statement may be true.

4. Our analysis adds several topological insights to what is known (e.g., [6, Proposition 1.2.4]). Primarily, we show that the subsequential limit points of the iterates are a connected set that is either a singleton or infinite. Moreover, if the set is infinite, we conclude that it cannot contain an open set.

Thus, to the best of our knowledge, our results provide a more general and complete global convergence/divergence analysis of gradient descent with diminishing step sizes under realistic assumptions for nonconvex, differentiable optimization problems that arise in data science.[5]

The remainder of this work is organized as follows. In section 2, we specify the class of nonconvex optimization problems of interest and the precise form of gradient descent with diminishing step sizes. In section 3, we analyze the behavior of gradient descent with diminishing step sizes. In section 4, we construct examples that elucidate the possible behaviors of gradient descent with diminishing step sizes in the divergence regime. Final remarks are given in section 5.

2. Gradient descent.

We begin by introducing the general class of optimization problems that we consider in this work. Then, we specify the precise form of gradient descent with diminishing step sizes. With the problem class and procedure specified, we describe relevant analysis approaches in the literature.

2.1. Optimization problem.

To cover a variety of canonical problems in data science [49, section A], consider the optimization problem

$$\min_{x \in \mathbb{R}^p} F(x) \tag{2.1}$$

under the following assumptions.

[4] Such functions must be beyond those that have globally Lipschitz continuous gradients and still be valid for data science problems. One promising set of function classes is that of L-smooth adaptable functions or relatively smooth functions, in which the error between the objective function at a point and a first-order Taylor approximation at another point is controlled by a global constant and, roughly, a Bregman distance between the point of interest and the approximation point [2, 8, 40, 53]. Because each such function class is determined by the choice of Bregman distance function, determining the right function class is still an open question. Another promising set of function classes is that of generalized smooth functions, in which the local Lipschitz rank is allowed to grow at different rates [38]. The case in which the Lipschitz rank can grow quadratically with respect to the gradient norm seems promising for data science problems [54, Table 2]. While the correct function class is still being carefully constructed, promising classes are being developed and analyzed.

[5] We again reference the excellent work of [32] for a complementary discussion about the behavior of gradient descent with either diminishing or constant step sizes for definable objective functions when gradient trajectories are bounded.

Assumption 2.1. The objective function, $F:\mathbb{R}^p \to \mathbb{R}$, is bounded from below by a constant $F_{l.b.}$.

Assumption 2.2. The gradient function $\dot{F}(x) = \nabla F(z)|_{z=x}$ exists for all $x \in \mathbb{R}^p$ and is locally Lipschitz continuous.

For our context, we use the following definition of local Lipschitz continuity.

Definition 2.3. *A function $G:\mathbb{R}^p \to \mathbb{R}^p$ is locally Lipschitz continuous if, for every $x \in \mathbb{R}^p$, there exists an open ball of x, $\mathcal{N}$ and there exists $L \ge 0$ such that, for all $y, z \in \mathcal{N}$,*

$$\|G(y) - G(z)\|_2 \le L \|y - z\|_2 . \tag{2.2}$$

Equivalently, G is locally Lipschitz continuous if, for every compact set $\mathcal{C} \subset \mathbb{R}^p$, there exists $L \ge 0$ such that (2.2) holds for all $y, z \in \mathcal{C}$. This well-known statement is shown in Lemma SM1.1 of the supplementary material.

To give an example of the broad applicability of Assumption 2.2, any optimization problem whose objective function is twice continuously differentiable immediately satisfies Assumption 2.2. This well-known statement is given formally in Lemma SM2.1.

2.2. Gradient descent with diminishing step sizes. Now, suppose we apply gradient descent with diminishing step sizes to solve (2.1). Specifically, given $x_0 \in \mathbb{R}^p$, we generate a sequence $\{x_k : k \in \mathbb{N}\}$ according to

$$x_{k+1} = x_k - M_k \dot{F}(x_k), \tag{2.3}$$

where M_k satisfies some of the following properties.

Property 2.4. $\{M_k : k+1 \in \mathbb{N}\} \subset \mathbb{R}^{p\times p}$ are symmetric positive definite matrices.

Property 2.5. $\sum_{k=0}^{\infty} \lambda_{\min}(M_k)$ diverges, where $\lambda_{\min}(M_k)$ denotes the smallest eigenvalue of M_k.

Property 2.6. $\lim_{k\to\infty} \lambda_{\max}(M_k) = 0$, where $\lambda_{\max}(M_k)$ denotes the largest eigenvalue of M_k.

Properties 2.4, 2.5, and 2.6 are a matrix-valued generalization of classical diminishing step size requirements [6, Proposition 1.2.4]. Moreover, Properties 2.4, 2.5, and 2.6 are enough to show that the objective function evaluated at the iterates converges and to show that the limit infimum of the norm of the gradient function evaluated at the iterates converges to zero (see Theorem 3.6). To show that the gradient function converges to zero, these properties will be augmented with the following.

Property 2.7. There exists $\kappa \ge 1$ such that $\lambda_{\max}(M_k)/\lambda_{\min}(M_k) \le \kappa$ for all $k+1 \in \mathbb{N}$.

Of interest, Properties 2.4, 2.5, and 2.6 can potentially account for adaptive step-size selection procedures that exist in the literature, namely, those that do not make use of objective function information. For example, Properties 2.4, 2.5, and 2.6 can apply to the method of [6] (with $\lambda = 1$), which combines incremental nonlinear least squares, the Gauss–Newton method, and the extended Kalman filter. However, especially in the nonlinear case, Property 2.5 would be difficult to verify without assuming something akin to what is called persistent excitation in the control literature [7, 10, 31, 47]. Indeed, in the objective-free first-order optimization (e.g., AdaGrad-type methods), this persistent excitation condition often manifests through a combination of assumptions about the optimization problem (e.g., bounded gradients) and the diagonal or identity-scaling choice of $\{M_k : k+1 \in \mathbb{N}\}$ [18, 24, 25, 26, 27, 56, 57].

2.3. Important analysis approaches in the literature.

With the problem and algorithm established, we briefly review two important analysis frameworks in the literature with respect to simple objective functions satisfying Assumptions 2.1 and 2.2: $|x|^3$ and $\exp(x)$. Note that these two examples are essential components in verifying that canonical data science problems have neither globally Lipschitz continuous gradients nor uniformly continuous gradients [see [49], section A].

In one analysis framework for trust region methods (e.g., [42, 34]), continuity of the gradient function, properties of the algorithm, and evaluations of the objective function are needed to show that the limit infimum of the gradient function evaluated at the iterates is zero. Furthermore, assuming uniform continuity of the gradient function allows for the conclusion that the limit of the gradient function evaluated at the iterates is zero [11, 12, 34, 42, 58, 59].[6] While continuity of the gradient function certainly holds for our two example objectives, neither of them satisfy uniform continuity of the gradient function. Moreover, in our context, gradient function information is not combined with objective function information to ensure sufficient decay at each step, which limits our ability to use the assumption of continuity of the gradient in place of Assumption 2.2.[7]

In the other analysis framework espoused by [6, Proposition 1.2.4], [45, Theorem 3.2], and [3, Lemma 10.4], the essential ingredient is a global upper-bound model for the objective function,

$$F(y) \leq F(x) + \dot{F}(x)^\intercal (y-x) + \frac{L}{2}\|y-x\|_2^2 \quad \text{for all } y, x \in \mathbb{R}^p, \tag{2.4}$$

where L is a fixed constant that arises from the assumption that the gradient function is *globally Lipschitz continuous* (i.e., L is the same regardless of $x \in \mathbb{R}^p$ and $\mathcal{N}$ in Definition 2.3). Indeed, this global upper-bound model is commonly used in recent analyses, both deterministic and stochastic [16, 17, 18, 24, 25, 26, 27, 56, 57]. This global upper-bound model is actualized by replacing y with x_{k+1} and x with x_k and rewriting the right-hand side strictly in terms of quantities depending on x_k. Then, the upper-bound model is manipulated to show that the objective function is decreasing. Unfortunately, such a global upper-bound model does not apply to the two simple example objective functions, which renders such analyses inapplicable to common data science problems.

In [6, Exercise 1.2.5], this global upper-bound model is relaxed to the case where such an L exists for every level set of the objective function and assumes every level set is bounded. In this case, this relaxed upper-bound model can then be used to establish that, if a gradient descent procedure remains in a level set, then the objective function converges to a finite value and the gradient function converges to zero. Indeed, this relaxed upper-bound model

[6] In [11, AF.2], uniform continuity implies the needed property.

[7] This raises the question of how much objective function information is really needed in order to ensure similar results as trust-region without substantially increasing computational costs for data science problems.

can account for $|x|^3$, but it cannot account for $\exp(x)$ or our example in section 4, which has bounded level sets, yet the iterates never remain in any level set. Hence, even this relaxation cannot account for the types of problems that satisfy Assumptions 2.1 and 2.2.

Our approach can be viewed as a generalization of [6, Exercise 1.2.5] because we can use Assumption 2.2 to write a valid upper-bound model for any two points in $\mathbb{R}^p$, even though we only assume local Lipschitz continuity of the gradient (see Lemma 3.1 and Example 3.2). We now introduce this analysis approach.

3. Global convergence analysis.

Here, we study the global convergence of gradient descent, (2.3), with diminishing step sizes satisfying Properties 2.4, 2.5, and 2.6 on a general class of nonconvex functions as defined by Assumptions 2.1 and 2.2. Our main conclusion is that, despite the allowed nonconvexity of a problem, the objective function and gradient function at the iterates are either stabilizing or the iterates must continually tend further away from the origin. Thus, if we somehow know that the iterates remain bounded, then they must converge to a stationary point.

To prove these claims, our main innovation is to analyze the gradient descent procedure under a stopping time framework, which is a theoretical construction that allows us to analyze the procedure without modifying it. We enumerate the steps in our analysis here.

1. In subsection 3.1, we establish a novel upper-bound model based on stopping times to relate the optimality gaps of two arbitrary points even under local Lipschitz continuity of the gradient function (see Lemma 3.1). We then simplify this statement when we substitute the two arbitrary points with consecutive iterates generated by the gradient descent procedure with diminishing step sizes (see Corollary 3.4).
2. In subsection 3.2, we apply Zoutendjik's analysis approach [60]. We show that the limit supremum and limit infimum of the objective function evaluated at the iterates must tend to each other if the iterates remain in a region for long enough (even if they eventually escape). We also show that the limit infimum of the gradient function evaluated at the iterates must tend to zero if the iterates remain in a region for long enough (even if they eventually escape).
3. In subsection 3.3, we strengthen the preceding statement using Property 2.7: We show that the limit of the gradient function evaluated at the iterates tends to zero if the iterates remain in a region for long enough (even if they eventually escape).
4. In subsection 3.4, we establish topological properties of the iterates when their subsequential limits are a bounded set. In particular, we establish the well-known results that the limit points of the iterates converge to a closed set where the gradient function is zero, and we establish—to the best of our knowledge—the novel result that this set is connected and cannot contain an open set (see Theorem 3.10). In other words, when it converges, gradient descent with diminishing step sizes tends to either a single point or an infinite set that must, in a sense, lack volume. Moreover, gradient descent with diminishing step sizes cannot have a cycle, nor can it converge to a limit cycle with a finite number of points.

We turn our attention to the divergence regime in section 4.

3.1. A relationship for the optimality gap.

We now establish an upper-bound inequality for the optimality gap between two points in $\mathbb{R}^p$ under local Lipschitz continuity (see subsection

2.3). To establish this result, we make use of a technique from probability theory that analyzes stochastic processes under stopping times. For the deterministic equivalent, we define, for an arbitrary point $x \in \mathbb{R}^p$ and $R \geq 0$,

$$\pi_x(R) = \begin{cases} 1, & \|x\|_2 \leq R, \\ 0, & \text{otherwise}. \end{cases} \tag{3.1}$$

Lemma 3.1. *Suppose that $F : \mathbb{R}^p \to \mathbb{R}$ satisfies Assumptions 2.1 and 2.2. Then, for all $R \geq 0$, there exists a constant $C_R > 0$ such that, for all $x, y \in \mathbb{R}^p$,*

$$[F(y) - F_{l.b.}]\pi_y(R)\pi_x(R) \leq \left[F(x) - F_{l.b.} - \dot{F}(x)^\intercal (y-x) + C_R \|y-x\|_2^2\right]\pi_x(R). \tag{3.2}$$

At first glance, we might think that Lemma 3.1 can be proved by combining (2.4) with $L \geq 0$ specific to the radius $R > 0$ of interest to show that

$$[F(y) - F_{l.b.}]\pi_y(R)\pi_x(R) \leq \left[F(x) - F_{l.b.} - \dot{F}(x)^\intercal (y-x) + \frac{L}{2}\|y-x\|_2^2\right]\pi_y(R)\pi_x(R) \tag{3.3}$$

and then using $\pi_y(R)\pi_x(R) \leq \pi_x(R)$ to upper bound the right-hand side to conclude that

$$[F(y) - F_{l.b.}]\pi_y(R)\pi_x(R) \leq \left[F(x) - F_{l.b.} - \dot{F}(x)^\intercal (y-x) + \frac{L}{2}\|y-x\|_2^2\right]\pi_x(R). \tag{3.4}$$

Unfortunately, it is the last step that can be problematic because the right-hand side can become negative, which produces a false inequality. The following example illustrates the issue.

Example 3.2. Consider

$$F(x) = \begin{cases} 10(1-x), & x \leq 1, \\ \dfrac{1}{10x-9} - 1, & x > 1, \end{cases} \quad \text{for which} \quad \dot{F}(x) = \begin{cases} -10, & x \leq 1, \\ \dfrac{-10}{(10x-9)^2}, & x > 1, \end{cases} \tag{3.5}$$

which is bounded from below and for which $\dot{F}(x)$ is globally Lipschitz continuous. If we now set $R = 1$, $x = 1$, then we see that $L = 0$ on $[-1, 1]$ and $\pi_x(1) = 1$. If we now select $y = 11$ (which would be the iterate generated by a gradient descent procedure at $x = 1$ with step size 1), then $\pi_y(1) = 0$. Plugging this into (3.4), $0 = (F(11)+1)0 \leq (0 + 1 - 100)1 = -99$, which is false. Hence, proving Lemma 3.1 requires a little more care, as we show below.

Proof. First, for any $R \geq 0$, define L_R to be the Lipschitz constant for the gradient in the closed ball of radius R around the point $0 \in \mathbb{R}^p$, which is well defined by Assumption 2.2 and Lemma SM1.1. For any fixed $\delta > 0$, it readily follows that $L_R \leq L_{R+\delta}$. Second, let $L(y, x)$ be the Lipschitz constant of the gradient in a closed ball of radius $\|y-x\|_2$ around the point x. Finally, for any $R \geq 0$, define G_R to be the maximum $\|\dot{F}(x)\|_2$ for all x in a closed ball of radius R around the point $0 \in \mathbb{R}^p$. Now, let $y, x \in \mathbb{R}^p$ be arbitrary.

By Taylor's remainder theorem,

$$\begin{aligned} F(y) - F_{l.b.} = F(x) - F_{l.b.} + \dot{F}(x)^\intercal (y-x) \\ + \int_0^1 \left[\dot{F}(x + t(y-x)) - \dot{F}(x)\right]^\intercal (y-x)dt. \end{aligned} \tag{3.6}$$

By applying Assumption 2.2 to the last term,

$$F(y) - F_{l.b.} \leq F(x) - F_{l.b.} + \dot{F}(x)^\intercal (y-x) + \frac{L(y,x)}{2} \|y-x\|_2^2. \tag{3.7}$$

Note that, to understand why we must keep going at this point in the proof, see Remark 3.7. We now introduce $\pi_y(R)$ and $\pi_x(R)$ into (3.7). That is,

$$\begin{aligned} &[F(y) - F_{l.b.}]\pi_y(R)\pi_x(R) \\ &\quad \leq \left[F(x) - F_{l.b.} - \dot{F}(x)^\intercal (y-x) + \frac{L(y,x)}{2} \|y-x\|_2^2 \right] \pi_y(R)\pi_x(R). \end{aligned} \tag{3.8}$$

If $\pi_y(R)\pi_x(R) = 1$, then $\|y\|_2 \leq R$ and $\|x\|_2 \leq R$. Thus, $L(y,x) \leq L_R \leq L_{R+\delta}$. When $\pi_y(R)\pi_x(R) = 0$, then both sides are trivially zero. Therefore,

$$\begin{aligned} &[F(y) - F_{l.b.}]\pi_y(R)\pi_x(R) \\ &\quad \leq \left[F(x) - F_{l.b.} - \dot{F}(x)^\intercal (y-x) + \frac{L_{R+\delta}}{2} \|y-x\|_2^2 \right] \pi_y(R)\pi_x(R). \end{aligned} \tag{3.9}$$

Now, we want $\pi_x(R)$ alone on the right-hand side. So, we simply add and subtract a term involving $\pi_x(R)$ and study the difference term. That is,

$$\begin{aligned} &[F(y) - F_{l.b.}]\pi_y(R)\pi_x(R) \\ &\quad \leq \left[F(x) - F_{l.b.} - \dot{F}(x)^\intercal (y-x) + \frac{L_{R+\delta}}{2} \|y-x\|_2^2 \right] \pi_x(R) \\ &\qquad + \left[F(x) - F_{l.b.} - \dot{F}(x)^\intercal (y-x) + \frac{L_{R+\delta}}{2} \|y-x\|_2^2 \right] [\pi_y(R)\pi_x(R) - \pi_x(R)]. \end{aligned} \tag{3.10}$$

We now have two cases to upper bound the last term of (3.10). Note that $\pi_y(R)\pi_x(R) - \pi_x(R) \leq 0$.

Case 1. If

$$F(x) - F_{l.b.} - \dot{F}(x)^\intercal (y-x) + \frac{L_{R+\delta}}{2} \|y-x\|_2^2 \geq 0, \tag{3.11}$$

then

$$\left[F(x) - F_{l.b.} - \dot{F}(x)^\intercal (y-x) + \frac{L_{R+\delta}}{2} \|y-x\|_2^2 \right] [\pi_y(R)\pi_x(R) - \pi_x(R)] \leq 0. \tag{3.12}$$

Hence, in this case, we can upper bound the last term in (3.10) by any nonnegative term.

Case 2. If

$$F(x) - F_{l.b.} - \dot{F}(x)^\intercal (y-x) + \frac{L_{R+\delta}}{2} \|y-x\|_2^2 < 0, \tag{3.13}$$

then, using $\pi_y(R)\pi_x(R) \leq \pi_x(R)$,

$$\left[F(x) - F_{l.b.} - \dot{F}(x)^\intercal (y-x) + \frac{L_{R+\delta}}{2} \|y-x\|_2^2 \right] [\pi_y(R)\pi_x(R) - \pi_x(R)] \geq 0. \tag{3.14}$$

Thus, we need only to find a lower bound for the first term in the product to upper bound the entire term. Specifically,

$$-\left\|\dot{F}(x)\right\|_2 \|y-x\|_2 \leq F(x) - F_{l.b.} - \dot{F}(x)^\intercal(y-x) + \frac{L_{R+\delta}}{2}\|y-x\|_2^2. \tag{3.15}$$

Now, when $\pi_y(R)\pi_x(R) < \pi_x(R)$, $\|x\|_2 \leq R$. Moreover, if (3.13) holds, then $R+\delta < \|y\|_2$. To see this, suppose that (3.13) holds and $\|y\|_2 \leq R+\delta$. Then, $L(y,x) \leq L_{R+\delta}$. If we now apply (3.7) and this inequality,

$$0 \leq F(y) - F_{l.b.} \leq F(x) - F_{l.b.} - \dot{F}(x)^\intercal(y-x) + \frac{L_{R+\delta}}{2}\|y-x\|_2^2, \tag{3.16}$$

which contradicts (3.13). Hence, in this case, $R+\delta < \|y\|_2$.

Using the triangle inequality, $R+\delta < \|y\|_2 \leq \|x\|_2 + \|y-x\|_2 \leq R + \|y-x\|_2$. That is, $1 \leq \|y-x\|_2/\delta \leq \|y-x\|_2^2/\delta^2$.

Hence,

$$\left[F(x) - F_{l.b.} - \dot{F}(x)^\intercal(y-x) + \frac{L_{R+\delta}}{2}\|y-x\|_2^2\right][\pi_y(R)\pi_x(R) - \pi_x(R)]$$

$$\leq \delta\left\|\dot{F}(x)\right\|_2 \frac{\|y-x\|_2}{\delta}[\pi_x(R) - \pi_y(R)\pi_x(R)] \tag{3.17}$$

$$\leq \delta\left\|\dot{F}(x)\right\|_2 \frac{\|y-x\|_2^2}{\delta^2}[\pi_x(R) - \pi_y(R)\pi_x(R)] \tag{3.18}$$

$$\leq \frac{G_R}{\delta}\|y-x\|_2^2\pi_x(R), \tag{3.19}$$

where, in the last line, we have used $\pi_x(R) - \pi_y(R)\pi_x(R) \leq \pi_x(R)$ because these are $\{0,1\}$-valued quantities.

Putting these two cases together in (3.10), we conclude that

$$\begin{aligned}&[F(y) - F_{l.b.}]\pi_y(R)\pi_x(R)\\ &\leq \left[F(x) - F_{l.b.} - \dot{F}(x)^\intercal(y-x) + \left(\frac{L_{R+\delta}}{2} + \frac{G_R}{\delta}\right)\|y-x\|_2^2\right]\pi_x(R).\end{aligned} \tag{3.20}$$

Letting $C_R = L_{R+\delta}/2 + G_R/\delta$, the conclusion follows. ■

Remark 3.3. If we replace (y,x) with (x_{k+1},x_k) in the preceding result, it might be tempting to choose a δ that minimizes C_R and then to use a standard approach to find a complexity result. However, this complexity result would only hold if all of the iterates remained within a radius R of 0, which, under Assumptions 2.1 and 2.2, cannot be guaranteed a priori, as shown by our construction in section 4. Thus, a complexity result would only be appropriate if some additional information is known to guarantee a single Lipschitz constant (e.g., by knowing that the iterates remain bounded), in which case we would directly make use of (2.4) and would have no use for Lemma 3.1.

We now apply Lemma 3.1 to the iterate sequence generated by gradient descent. To do so, we will make use of the following notation:

$$\chi_k^0(R) = \begin{cases} 1, & \|x_j\|_2 \leq R,\ j=0,\ldots,k, \\ 0, & \text{otherwise.} \end{cases} \tag{3.21}$$

That is, $\chi_k^0(R) = \pi_{x_0}(R)\pi_{x_1}(R)\cdots\pi_{x_k}(R)$. With this notation, we have the following simplification of Lemma 3.1 when applied to gradient descent.

Corollary 3.4. *Suppose that $F : \mathbb{R}^p \to \mathbb{R}$ satisfies Assumptions 2.1 and 2.2. Let $x_0 \in \mathbb{R}^p$, and let $\{x_k : k \in \mathbb{N}\}$ be generated by (2.3) satisfying Properties 2.4 and 2.6. Then, for all $R \geq 0$, there exists $K \in \mathbb{N}$ such that, for all $k \geq K$,*

$$[F(x_{k+1}) - F_{l.b.}]\chi_{k+1}^0(R) \leq \left[F(x_k) - F_{l.b.} - \frac{1}{2}\lambda_{\min}(M_k)\left\|\dot{F}(x_k)\right\|_2^2\right]\chi_k^0(R). \tag{3.22}$$

Proof. By Lemma 3.1, there exists $C_R > 0$ such that, for any $k+1 \in \mathbb{N}$,

$$\begin{aligned} &[F(x_{k+1}) - F_{l.b.}]\pi_{x_{k+1}}(R)\pi_{x_k}(R) \\ &\quad\leq \left[F(x_k) - F_{l.b.} - \dot{F}(x_k)^\intercal M_k \dot{F}(x_k) + C_R\left\|M_k\dot{F}(x_k)\right\|_2^2\right]\pi_{x_k}(R), \end{aligned} \tag{3.23}$$

where we have made use of (2.3) to replace $x_{k+1} - x_k$. If we now multiply both sides by the nonnegative quantity $\pi_{x_0}(R)\cdots\pi_{x_{k-1}}(R)$, then

$$\begin{aligned} &[F(x_{k+1}) - F_{l.b.}]\chi_{k+1}^0(R) \\ &\quad\leq \left[F(x_k) - F_{l.b.} - \dot{F}(x_k)^\intercal M_k \dot{F}(x_k) + C_R\left\|M_k\dot{F}(x_k)\right\|_2^2\right]\chi_k^0(R). \end{aligned} \tag{3.24}$$

The result follows if we show that there exists $K \in \mathbb{N}$ such that, for all $k \geq K$,

$$-\dot{F}(x_k)^\intercal M_k \dot{F}(x_k) + C_R\left\|M_k\dot{F}(x_k)\right\|_2^2 \leq -\frac{1}{2}\lambda_{\min}(M_k)\left\|\dot{F}(x_k)\right\|_2^2. \tag{3.25}$$

To this end, we prove that, if M is symmetric positive definite with $\lambda_{\max}(M) < 1/(2C_R)$, then, for any $v \in \mathbb{R}^p$ with unit norm, $-v^\intercal Mv + C_R v^\intercal MMv \leq -\frac{1}{2}\lambda_{\min}(M)$. Let $0 < \lambda_{\min}(M) = \lambda_p \leq \lambda_{p-1} \leq \cdots \leq \lambda_2 \leq \lambda_1 = \lambda_{\max}(M) < 1/(2C_R)$, where λ_ℓ denote the eigenvalues of M. Using the Schur decomposition, there exists an orthogonal matrix Q such that $-v^\intercal Mv + C_R v^\intercal MMv = \sum_{\ell=1}^p(-\lambda_\ell + C_R\lambda_\ell^2)w_\ell^2$, where w_ℓ is the ℓ^{th} component of Qv (note that $\|w\|_2 = \|Qv\|_2 = \|v\|_2 = 1$). Since $\lambda_\ell < 1/(2C_R)$, it follows that $C_R\lambda_\ell^2 < \lambda_\ell/2$. Subtracting λ_ℓ from both sides, $-\lambda_\ell + C_R\lambda_\ell^2 < -\lambda_\ell/2 \leq -\lambda_{\min}(M)/2$. Thus,

$$-v^\intercal Mv + C_R v^\intercal MMv \leq -\sum_{\ell=1}^p \frac{\lambda_\ell}{2}w_\ell^2 = -\frac{1}{2}v^\intercal Mv \leq -\frac{\lambda_{\min}(M)}{2}. \tag{3.26}$$

Since $\lambda_{\max}(M_k) \to 0$, there exists a $K \in \mathbb{N}$ such that, for all $k \geq K$, $\lambda_{\max}(M_k) \leq 1/(2C_R)$. Hence, there exists a K such that, for all $k \geq K$, (3.25) holds. ■

Remark 3.5. In light of *universal gradient methods* [44], we may be interested in whether this result can be generalized to the case of assuming Hölder continuity of the gradient function. Just as with Lipschitz continuity, Hölder continuity of the gradient condition is considered either globally, as it is for universal gradient methods [44], or locally. If we can generalize the above result to the case of local Hölder continuity and continue with our analysis below, then we would see that gradient descent with diminishing step sizes is, in a sense, more universal than universal gradient methods. We anticipate that this is possible to do by two approaches. In one approach, we can mimic [50] and use Young's inequality to recover something similar to the recursion in Corollary 3.4 with an additional additive term of $\lambda_{\max}(M_k)^{1+\alpha}(1-\alpha)/2$, where $\alpha \in [0,1]$ is the Hölder constant (with 1 corresponding to Lipschitz continuity). In this case, we will need to strengthen Property 2.6 so that $\sum_k \lambda_{\max}(M_k)^{1+\alpha} < \infty$. To avoid strengthening this property, as a second approach, we anticipate using a stopping condition on the gradient, as done in [48]. We will leave this to future work.

3.2. Applying Zoutendjik's analysis approach.

We now apply the recursive relationship established in Corollary 3.4 to study the objective and gradient using Zoutendjik's analysis method [60]. Recall that our main conclusion from the next result is that the limit supremum and limit infimum of the objective function evaluated at the iterates must tend to each other if the iterates persist in a region for long enough (even if they eventually escape), and the limit infimum of the gradient function evaluated at the iterates must tend to zero under similar circumstances. We stress that these conclusions are not the same as presupposing that the iterates remain in a bounded region.

Theorem 3.6. *Suppose that $F:\mathbb{R}^p \to \mathbb{R}$ satisfies Assumptions 2.1 and 2.2. Let $x_0 \in \mathbb{R}^p$, and let $\{x_k : k \in \mathbb{N}\}$ be generated by (2.3) satisfying Properties 2.4, 2.5, and 2.6. Then, for all $R \geq 0$,*

$$\lim_{k\to\infty} F(x_k)\chi_k^0(R) \text{ exists and is finite, and } \liminf_{k\to\infty} \left\| \dot{F}(x_k) \right\|_2 \chi_k^0(R) = 0. \tag{3.27}$$

If $\sup_k \|x_k\|_2 < \infty$, then $\lim_{k\to\infty} F(x_k)$ exists and is finite, and $\liminf_{k\to\infty} \|\dot{F}(x_k)\|_2 = 0$.

Proof. Let $R \geq 0$. The conditions of Corollary 3.4 are satisfied, and its conclusion is used freely herein. For the objective function, there exists $K \in \mathbb{N}$ such that, for all $k \geq K$, $[F(x_{k+1}) - F_{l.b.}]\chi_{k+1}^0(R) \leq [F(x_k) - F_{l.b.}]\chi_k^0(R)$. Because $\{[F(x_k) - F_{l.b.}]\chi_k^0(R) : k \geq K\}$ is a nonincreasing sequence bounded from below, it converges. Now, if we further assume that $\sup_k \|x_k\|_2 < \infty$, then there exists an $R > 0$ such that $\chi_k^0(R) = 1$ for all $k+1 \in \mathbb{N}$. Hence, $\lim_{k\to\infty} F(x_k) - F_{l.b.}$ exists and is finite.

For the gradient function, applying the conclusion of Corollary 3.4 and rearranging terms, for all $k \geq K$,

$$\frac{1}{2}\lambda_{\min}(M_k) \left\| \dot{F}(x_k) \right\|_2^2 \chi_k^0(R) \leq [F(x_k) - F_{l.b.}]\chi_k^0(R) - [F(x_{k+1}) - F_{l.b.}]\chi_{k+1}^0(R). \tag{3.28}$$

Letting $j \geq K$ and using $F(x_{j+1}) - F_{l.b.} \geq 0$,

$$\sum_{k=K}^{j} \frac{1}{2}\lambda_{\min}(M_k) \left\| \dot{F}(x_k) \right\|_2^2 \chi_k^0(R) \leq [F(x_K) - F_{l.b.}]\chi_K^0(R). \tag{3.29}$$

Now, for a contradiction, suppose that there exists $c > 0$ such that $\liminf_{k\to\infty} \|\dot F(x_k)\|_2^2 \chi_k^0(R) > c$. Then, there exists a $K' > K$ such that

$$\frac{c}{2}\sum_{k=K'}^{j} \lambda_{\min}(M_k) \le \sum_{k=K'}^{j} \frac{1}{2}\lambda_{\min}(M_k)\left\|\dot F(x_k)\right\|_2^2 \chi_k^0(R) \le [F(x_K) - F_{l.b.}]\chi_K^0(R) < \infty. \tag{3.30}$$

By Property 2.5, we have a contradiction. This part of the result follows for any $R \ge 0$.

Now, if $\sup_k \|x_k\|_2 < \infty$, then there exists an $R > 0$ such that $\sup_k \|x_k\|_2 < R$. Therefore, $\chi_k^0(R) = 1$ for all $k+1 \in \mathbb{N}$, and thus, the final part of the result follows. ■

Remark 3.7. Suppose we directly attempt to use Zoutendjik's analysis approach in (3.7) with $y = x_{k+1}$ and $x = x_k$. We begin by rearranging (3.7) and summing up to $j \in \mathbb{N}$ to conclude that

$$\sum_{k=0}^{j} \dot F(x_k)^\intercal M_k \dot F(x_k) - \frac{L(x_{k+1},x_k)}{2}\left\|M_k \dot F(x_k)\right\|_2^2 \le F(x_0) - F_{l.b.}. \tag{3.31}$$

Thus, we conclude that

$$\lim_{k\to\infty} \dot F(x_k)^\intercal M_k \dot F(x_k) - \frac{L(x_{k+1},x_k)}{2}\left\|M_k \dot F(x_k)\right\|_2^2 = 0. \tag{3.32}$$

Unfortunately, this conclusion does not imply that $\dot F(x_k) \to 0$ as $k \to \infty$. For instance, suppose that, as $k \to \infty$, $L(x_{k+1},x_k) \to \infty$. If $M_k = 2L(x_{k+1},x_k)^{-1} I$ for all k, then a straightforward substitution will show that the limit is satisfied, yet $\dot F(x_k)$ does not have to be zero. Hence, using Zoutendjik's analysis method on this line of logic would not produce the desired conclusion. However, as shown in Theorem 3.6, using Zoutendjik's analysis method on the conclusion of Lemma 3.1 is fruitful.

3.3. Convergence of the gradient.

One limitation of Theorem 3.6 is that it only provides for the limit infimum of the gradient function to be zero. Here, we will use Property 2.7 to conclude that the limit of the gradient function is zero.

Theorem 3.8. *Suppose that $F : \mathbb{R}^p \to \mathbb{R}$ satisfies Assumptions 2.1 and 2.2. Let $x_0 \in \mathbb{R}^p$, and let $\{x_k : k \in \mathbb{N}\}$ be generated by (2.3) satisfying Properties 2.4, 2.5, 2.6, and 2.7. Then, for all $R \ge 0$,*

$$\lim_{k\to\infty} F(x_k)\chi_k^0(R) \text{ exists and is finite, and } \lim_{k\to\infty}\left\|\dot F(x_k)\right\|_2 \chi_k^0(R) = 0. \tag{3.33}$$

If $\sup_k \|x_k\|_2 < \infty$, then $\lim_{k\to\infty} F(x_k)$ exists and is finite, and $\lim_{k\to\infty} \|\dot F(x_k)\|_2 = 0$.

Before proving this statement, we briefly comment on conditions to guarantee $\sup_k \|x_k\|_2 < \infty$ given an arbitrary initialization x_0 and diminishing sequence $\{M_k : k+1 \in \mathbb{N}\}$. Consider the simple example of $F(\theta) = \exp(-\|x\|_2^2)$, which has globally Lipschitz continuous gradients (cf. Assumption 2.2, which is our much less restrictive assumption). For this example, any iterate sequence initialized at $x_0 \neq 0$ will diverge. Thus, to avoid divergence of the iterates, a geometric condition on the objective function seems to be necessary. Such geometric conditions have tended to be global, ranging from strong convexity to the global Polyak–Łojasiewicz condition [33]. One geometric condition, the uniform Kurdyka–Łojasiewicz condition [32], is shown to hold for definable functions on o-minimal structures and, when used in combination with an assumption of bounded continuous gradient paths, can guarantee that iterates remain bounded (for sufficiently small step sizes). While the assumption of bounded continuous gradient paths retains some flavor of bounding the iterates directly, this geometric condition and its corresponding analysis provide an important step in understanding when $\sup_k \|x_k\|_2 < \infty$. We now turn to the proof of Theorem 3.8.

Proof. By Theorem 3.6, we need only prove that, for any $R \geq 0$, $\limsup_{k\to\infty} \|\dot{F}(x_k)\|_2 \chi_k^0(R) = 0$. Fix $R \geq 0$. There are two cases.

Case 1. For some $K+1 \in \mathbb{N}$, $\chi_K^0(R) = 0$. Then, $\chi_k^0(R) = 0$ for all $k \geq K$. The result follows.

Case 2. For all $k+1 \in \mathbb{N}$, $\chi_k^0(R) = 1$. In this case, $\|x_k\|_2 \leq R$ for all $k+1 \in \mathbb{N}$. Let L_R be the Lipschitz constant in the closed ball of radius R around 0 (see Lemma SM1.1), and let G_R be the supremum of $\|\dot{F}(x)\|_2$ over all x in the closed ball of radius R around 0.

We now proceed in two steps. First, we show that, for any $\epsilon > 0$, there exists a $K' \in \mathbb{N}$ such that, for all $k \geq K'$,

$$\left| \left\|\dot{F}(x_{k+1})\right\|_2 - \left\|\dot{F}(x_k)\right\|_2 \right| < \frac{\epsilon}{4}. \tag{3.34}$$

Then, we use a proof by contradiction to show that the $\limsup_{k\to\infty} \|\dot{F}(x_k)\|_2 \not> \epsilon$.

For the first part, let $\epsilon > 0$. Now,

$$\begin{aligned}
\left| \left\|\dot{F}(x_{k+1})\right\|_2 - \left\|\dot{F}(x_k)\right\|_2 \right| &\leq \left\|\dot{F}(x_{k+1}) - \dot{F}(x_k)\right\|_2 && (3.35)\\
&\leq L_R \|x_{k+1} - x_k\|_2 && (3.36)\\
&\leq L_R \left\| M_k \dot{F}(x_k) \right\|_2 && (3.37)\\
&\leq L_R G_R \lambda_{\max}(M_k). && (3.38)
\end{aligned}$$

By Property 2.6, there exists $K' \in \mathbb{N}$ such that, for all $k \geq K'$, $L_R G_R \lambda_{\max}(M_k) < \epsilon/4$.

Suppose now that $\limsup_{k\to\infty} \|\dot{F}(x_k)\|_2 > \epsilon$. Let $u_0 = \min\{k > \max\{K, K'\} : \|\dot{F}(x_k)\|_2 > \epsilon\}$, where K is given by Corollary 3.4. By Theorem 3.6, we can now define the following three subsequences of $\mathbb{N}$ for all $i \in \mathbb{N}$:

1. $j_i = \min\{t > u_{i-1} : \|\dot{F}(x_t)\|_2 < \epsilon/2\}$.
2. $u_i = \min\{t > j_i : \|\dot{F}(x_t)\|_2 > \epsilon\}$.
3. $\ell_i = \min\{t \in [j_i, u_i) : \|\dot{F}(x_s)\|_2 > \epsilon/2, s = t+1, \ldots, u_i\}$.

Note that, by construction, $\|\dot{F}(x_{\ell_i})\|_2 \leq \epsilon/2$ and $\|\dot{F}(x_{u_i})\|_2 > \epsilon$. Hence,

$$\frac{\epsilon}{2} = \epsilon - \frac{\epsilon}{2} < \left\|\dot{F}(x_{u_i})\right\|_2 - \left\|\dot{F}(x_{\ell_i})\right\|_2 = \sum_{t=\ell_i}^{u_i - 1} \left\|\dot{F}(x_{t+1})\right\|_2 - \left\|\dot{F}(x_t)\right\|_2. \tag{3.39}$$

If we now make use of the reverse triangle inequality, local Lipschitz continuity, and (2.3), then $\epsilon/2 < \sum_{t=\ell_i}^{u_i-1} L_R \lambda_{\max}(M_t) \|\dot{F}(x_t)\|_2$ (note, the reasoning is the same as the first part of the proof).

Now, since $\ell_i > K'$ for all $i \in \mathbb{N}$ and $\|\dot F(x_{\ell_i+1})\|_2 > \epsilon/2$, $\|\dot F(x_{\ell_i})\|_2 > \epsilon/4$ by (3.34). Hence, $\epsilon/4 < \|\dot F(x_s)\|_2$ for $s = \ell_i, \ldots, u_i - 1$. Using this fact,

$$\frac{\epsilon}{2} < \sum_{t=\ell_i}^{u_i-1} L_R \lambda_{\max}(M_t) \|\dot F(x_t)\|_2 \le \frac{\epsilon}{4} \sum_{t=\ell_i}^{u_i-1} L_R \lambda_{\max}(M_t) \left(\frac{4 \left\| \dot F(x_t) \right\|_2}{\epsilon} \right)^2. \tag{3.40}$$

Simplifying and applying Property 2.7, for all $i \in \mathbb{N}$,

$$\frac{\epsilon^2}{8 L_R \kappa} < \sum_{t=\ell_i}^{u_i-1} \lambda_{\min}(M_t) \left\| \dot F(x_t) \right\|_2^2. \tag{3.41}$$

Summing both sides over $i \in \mathbb{N}$, the left-hand side diverges, while the right-hand side is bounded by (3.29). Hence, we have a contradiction, and the conclusion follows. ■

From this proof, we might question whether it is necessary to use Property 2.7 in order to replace $\lambda_{\max}(M_t)$ with $\lambda_{\min}(M_t)$ in (3.41). We provide a concrete example where our reasoning faces difficulty if Property 2.7 is not used. As the example below shows, it is possible to relax Property 2.7 if the sequence $\{M_k\}$ eventually has common invariant subspaces, but we do not pursue this here.

Example 3.9. Let $F : \mathbb{R}^2 \to \mathbb{R}$ be

$$F(x) = \frac{1}{2}(x^{(1)})^2 + \frac{1}{10}(x^{(2)})^2, \tag{3.42}$$

where $x^{(i)}$ is the i^{th} component of x. Consider now x_0 such that $x_0^{(1)} = 0$ and $x_0^{(2)} = 1$. In order to violate Property 2.7, let

$$M_k = \frac{1}{5} \begin{bmatrix} (k+1)^{-1/2} & 0 \\ 0 & (k+1)^{-1} \end{bmatrix}, \; k+1 \in \mathbb{N}. \tag{3.43}$$

When we apply gradient descent, $x_k^{(1)} = 0$ for all $k \in \mathbb{N}$ and $x_k^{(2)} > 0.8(k+1)^{-1/5}$ [43, p. 1578]. Then, $\|\dot F(x_k)\|_2 > 0.16(k+1)^{-1/5}$. Now, we have $\lambda_{\max}(M_k)\|\dot F(x_k)\|_2^2 > 0.01(k+1)^{-9/10}$, which produces a divergent series, whereas $\lambda_{\min}(M_k)$ in place of $\lambda_{\max}(M_k)$ would produce a convergent series.[8]

3.4. Topological properties of the iterates.

We now turn our attention to the asymptotic behavior of the iterates in the bounded regime. We will make use of the closure of subsequential limits (see Lemma SM3.1) and a fact about the density of subsequential limits of a decaying sequence (see Lemma SM3.2). We state the main result in Theorem 3.10.

Theorem 3.10. *Suppose that $F : \mathbb{R}^p \to \mathbb{R}$ satisfies Assumptions 2.1 and 2.2. Let $x_0 \in \mathbb{R}^p$, and let $\{x_k : k \in \mathbb{N}\}$ be generated by (2.3) satisfying Properties 2.4, 2.5, 2.6, and 2.7. If $\sup_k \|x_k\|_2 < \infty$ and we let $\mathcal{C}$ denote the subsequential limits of $\{x_k : k+1 \in \mathbb{N}\}$, then*

[8] To show convergence, we need an upper bound on the rate of convergence of $x_k^{(2)}$, which is on the order of $(k+1)^{-1/5}$ (see [43], p. 1578).

1. *$\mathcal{C}$ is closed;*
2. *For all $z \in \mathcal{C}$, $\dot{F}(z) = 0$;*
3. *$\mathcal{C}$ is connected;*
4. *$\mathcal{C}$ does not contain an open set; and*
5. *Either $|\mathcal{C}| = 1$ or $|\mathcal{C}| = \infty$.*

Proof. The first statement follows from Lemma SM3.1. For the second statement, if $z \in \mathcal{C}$, then there is a subsequence $\{x_{k_j} : j \in \mathbb{N}\}$ such that $\lim_j x_{k_j} = z$. By the continuity of $x \mapsto \dot{F}(x)$ (see Assumption 2.2), $\dot{F}(z) = \lim_j \dot{F}(x_{k_j})$. The limit on the right-hand side is zero by Theorem 3.8.

For the third statement, recall that $\mathcal{C}$ is bounded by hypothesis and $\mathcal{C}$ is closed by the first statement. Hence, $\mathcal{C}$ is compact. Suppose that $\mathcal{C}$ is not connected. Then, there are two disjoint open sets, O_1 and O_2, whose union contains $\mathcal{C}$ and whose individual intersections with $\mathcal{C}$ are nonempty. We denote the intersections of O_1 and O_2 with $\mathcal{C}$ by $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively. We now proceed in three steps. First, we verify that $\mathcal{C}_1$ and $\mathcal{C}_2$ are closed and, consequently, compact. Second, we use compactness to show that the distance between $\mathcal{C}_1$ and $\mathcal{C}_2$ is strictly larger than zero. Third, we use the diminishing step sizes and Lemma SM3.2 to derive a contradiction.

Suppose that $\mathcal{C}_1$ is not closed. Let z be a limit point of $\mathcal{C}_1$ that is not in $\mathcal{C}_1$. Then, $z \in \mathcal{C}$, which implies that $z \in \mathcal{C}_2 \subset O_2$. There is a sequence of points in $\mathcal{C}_1$ contained in an arbitrarily small neighborhood of z, which implies that $\mathcal{C}_1 \cap O_2 \neq \emptyset$, which is a contradiction. Hence, $\mathcal{C}_1$ is closed. The same argument shows $\mathcal{C}_2$ is closed.

Since $\mathcal{C}_1$ and $\mathcal{C}_2$ are closed and bounded, they are compact. Now, $(z_1, z_2) \mapsto \|z_1 - z_2\|_2$ is a continuous function. Hence, this function applied to $\mathcal{C}_1 \times \mathcal{C}_2$ must achieve its minimum at some points $z_1^* \in \mathcal{C}_1$ and $z_2^* \in \mathcal{C}_2$. If $z_1^* = z_2^*$, then $O_1 \cap O_2 \neq \emptyset$, which is a contradiction. Hence, $z_1^* \neq z_2^*$ so the distance between any points in $\mathcal{C}_1$ and $\mathcal{C}_2$ is at least $\|z_1^* - z_2^*\|_2 > 0$.

Define a function $g : \mathbb{R}^p \to \mathbb{R}_{\geq 0}$ such that $g(x) = \inf_{w \in \mathcal{C}_1} \|x - w\|_2$. Then, $g(z) = 0$ for any $z \in \mathcal{C}_1$ and $g(z) \geq \|z_1^* - z_2^*\|_2$ for $z \in \mathcal{C}_2$. Hence, $\liminf_k g(x_k) = 0$ and $\limsup_k g(x_k) \geq \|z_1^* - z_2^*\|$. We now verify that $\lim_k g(x_{k+1}) - g(x_k) = 0$ and apply Lemma SM3.2 to derive a contradiction. For any $k \in \mathbb{N}$, there exists a $w_k \in \mathcal{C}_1$ such that $g(x_k) = \|x_k - w_k\|_2$. Hence,

$$g(x_{k+1}) - g(x_k) = \inf_{w \in \mathcal{C}_1} \|x_{k+1} - w\|_2 - \|x_k - w_k\|_2 \tag{3.44}$$

$$\leq \|x_{k+1} - w_k\|_2 - \|x_k - w_k\|_2 \tag{3.45}$$

$$\leq \|x_{k+1} - x_k\|_2 \tag{3.46}$$

$$\leq \|M_k \dot{F}(x_k)\|_2. \tag{3.47}$$

Note that $\|M_k \dot{F}(x_k)\|_2 \leq \lambda_{\max}(M_k) G_R$, where $G_R = \sup_{x : \|x\|_2 \leq R} \|\dot{F}(x)\|_2$ and $R = \sup_k \|x_k\|_2 < \infty$. Since $\lambda_{\max}(M_k) \to 0$, then $g(x_{k+1}) - g(x_k) \to 0$. Hence, by Lemma SM3.2, there is a subsequence $\{g(x_{k_j}) : j \in \mathbb{N}\}$ that converges to, say, $\|z_1^* - z_2^*\|_2/2$. Consequently, $\{x_{k_j} : j \in \mathbb{N}\}$ is a bounded sequence, and it has a subsequence that converges to a point z^* such that

$\inf_{w\in\mathcal{C}_1}\|z^*-w\|_2=\|z_1^*-z_2^*\|_2/2$. Hence, z^* is a subsequential limit, but it is not in either $\mathcal{C}_1$ or $\mathcal{C}_2$, which is a contradiction. Thus, $\mathcal{C}$ is connected.

For the fourth statement, suppose that $\mathcal{C}$ contains an open set O. Let $z\in O$. Since z is a limit point of a subsequence, there exists an $k\in\mathbb{N}$ such that $x_k\in O$. By the first statement, $\dot{F}(x_k)=0$, which, by (2.3), implies that $x_j=x_k$ for all $j\geq k$. Hence, $\mathcal{C}$ is the singleton, $\{x_k\}$, which is a contradiction. Thus, $\mathcal{C}$ cannot contain an open set.

For the final statement, recall that $\mathcal{C}$ is connected. This implies that $\mathcal{C}$ cannot contain a finite number of points other than a single point. So, either $|\mathcal{C}|=1$ or $|\mathcal{C}|=\infty$. ■

4. The divergence regime.

Theorem 3.6 leaves open the possibility that the iterates can diverge. Of course, this divergence regime is possible even under the stricter assumption of global Lipschitz continuity of the gradient. Under global Lipschitz continuity of the gradient, when the iterates diverge, the objective function still converges to a finite quantity and the gradient function converges to zero [6, Proposition 1.2.4]. For example, the globally Lipschitz smooth function, $F(x)=\exp(-x^2)$, achieves its minimum as the iterates diverge, and, in this divergence regime, the objective function converges to zero and the gradient function converges to zero.

While the preceding example gives a rosy prognosis, globally Lipschitz smooth functions can experience pathological behavior—at least for a finite number of iterates. In [55, p. 62], given a finite number m and an algorithm, a continuously differentiable function on the unit interval can be constructed such that, at m test points (presumably, corresponding to m iterates of an optimization algorithm), the objective function is zero and the gradient function is -1.[9] Thus, globally Lipschitz smooth functions can experience gradient functions that are bounded away from zero for a finite amount of time but must eventually be well behaved. On the other extreme of functions that are continuously differentiable (a condition that is more general than Assumption 2.2), a function can be constructed on all of $\mathbb{R}$ such that, for a given sequence of test points, the objective at these test points is zero and the gradient remains fixed at 1 [13, Example 2.1.1]. Thus, for locally Lipschitz continuous gradient functions, which fall between the cases of globally Lipschitz continuous gradient functions and continuous gradient functions, what behavior can we expect?

To be unequivocal, under our realistic assumption of local Lipschitz continuity of the gradient function, will the objective function always converge to a finite quantity and will the gradient function always converge to zero when the iterates diverge? Unfortunately, the answer is no—that is, the assumption of local Lipschitz continuity of the gradient function will produce behaviors that are more aligned with the assumption of continuous gradient functions. In this section, we will construct several examples that show the extreme behaviors that can occur in the divergence regime when only local Lipschitz continuity is assumed. Of note, we construct an example in which catastrophic divergence can occur: The iterates diverge, the objective function diverges to infinity, and the gradient norm remains uniformly bounded away from zero. We show this construction here. Our remaining constructions are specified in Table 1. We underscore that our constructions can be used to generate objective functions on

[9] The same claim can be shown to hold using the construction in [13, Example 2.1.1].

Table 1
Summary of counterexamples for the divergence regime.

Reference	Summary
This section	A case for which the iterates of gradient descent will produce objective function values that diverge and gradient function values that are uniformly bounded away from zero.
Section SM4	A case for which the iterates of gradient descent will produce objective function values whose limit supremum is infinity and whose limit infimum is zero, while the gradient function values remain bounded away from zero.
Section SM5	A case for which the iterates of gradient descent will produce objective function values that diverge and the gradient function tends to zero.

which gradient descent can have other interesting behaviors that we do not explicitly construct here (e.g., the limit infimum and limit supremum of the gradient function being distinct).

4.1. Construction of the objective function.

Let $\{m_k : k+1 \in \mathbb{N}\}$ be a sequence of scalars such that $m_k > 0$, $\sum_k m_k = \infty$, and $m_k \to 0$ as $k \to \infty$. Define $S_0 = 0$ and $S_{k+1} = \sum_{j=0}^{k} m_k$ for all integers $k \geq 0$.

For the objective function, define $F : \mathbb{R} \to \mathbb{R}$ by

$$F(x) = \begin{cases} -x, & x \leq 0, \\ f_j(x), & x \in (S_j, S_{j+1}] \ \text{ for all } j+1 \in \mathbb{N}, \end{cases} \tag{4.1}$$

where $\{f_j : (S_j, S_{j+1}] \to \mathbb{R} : j+1 \in \mathbb{N}\}$ are defined iteratively as follows. Let

$$f_0(x) = \begin{cases} -x, & x \in \left(0, \dfrac{m_0}{16}\right), \\ \dfrac{8}{m_0}\left(x - \dfrac{m_0}{8}\right)^2 - \dfrac{3m_0}{32}, & x \in \left[\dfrac{m_0}{16}, \dfrac{3m_0}{16}\right), \\ -\dfrac{5m_0}{16} \exp\left(\dfrac{5m_0/16}{x - m_0/2} + 1\right) + \dfrac{m_0}{4}, & x \in \left[\dfrac{3m_0}{16}, \dfrac{m_0}{2}\right), \\ \dfrac{m_0}{4}, & x = \dfrac{m_0}{2}, \\ \dfrac{5m_0}{16} \exp\left(-\dfrac{5m_0/16}{x - m_0/2} + 1\right) + \dfrac{m_0}{4}, & x \in \left(\dfrac{m_0}{2}, \dfrac{13m_0}{16}\right), \\ \dfrac{-8}{m_0}\left(x - \dfrac{7m_0}{8}\right)^2 + \dfrac{19m_0}{32}, & x \in \left[\dfrac{13m_0}{16}, \dfrac{15m_0}{16}\right), \\ -x + \dfrac{3m_0}{2}, & x \in \left[\dfrac{15m_0}{16}, m_0\right], \end{cases} \tag{4.2}$$

which is plotted for a particular choice of m_0 in Figure 1. Now, for $j \in \mathbb{N}$, let $x' = x - S_j$, and let

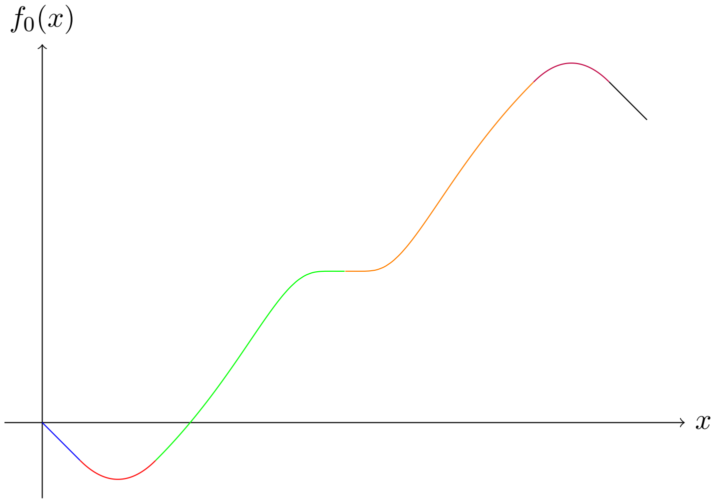

Figure 1. *Plot of $f_0(x)$ with $m_0 = 8.0$, with each component shown in a different color.*

$$f_j(x) = \begin{cases} -x' + f_{j-1}(S_j), & x' \in \left(0, \dfrac{m_j}{16}\right), \\ \dfrac{8}{m_j}\left(x' - \dfrac{m_j}{8}\right)^2 - \dfrac{3m_j}{32} + f_{j-1}(S_j), & x' \in \left[\dfrac{m_j}{16}, \dfrac{3m_j}{16}\right), \\ -\dfrac{5m_j}{16}\exp\left(\dfrac{5m_j/16}{x' - m_j/2} + 1\right) + \dfrac{m_j}{4} + f_{j-1}(S_j), & x' \in \left[\dfrac{3m_j}{16}, \dfrac{m_j}{2}\right), \\ \dfrac{m_j}{4} + f_{j-1}(S_j), & x' = \dfrac{m_j}{2}, \\ \dfrac{5m_j}{16}\exp\left(\dfrac{-5m_j/16}{x' - m_j/2} + 1\right) + \dfrac{m_j}{4} + f_{j-1}(S_j), & x' \in \left(\dfrac{m_j}{2}, \dfrac{13m_j}{16}\right), \\ \dfrac{-8}{m_j}\left(x' - \dfrac{7m_j}{8}\right)^2 + \dfrac{19m_j}{32} + f_{j-1}(S_j), & x' \in \left[\dfrac{13m_j}{16}, \dfrac{15m_j}{16}\right), \\ -x' + \dfrac{3m_j}{2} + f_{j-1}(S_j), & x' \in \left[\dfrac{15m_j}{16}, m_j\right]. \end{cases} \tag{4.3}$$

4.2. Properties of the objective function. We show that $F : \mathbb{R} \to \mathbb{R}$, as defined in (4.1), (4.2), and (4.3), satisfies Assumptions 2.1 and 2.2. We begin by proving that each component, $f_j : (S_j, S_{j+1}] \to \mathbb{R}$, satisfies Assumptions 2.1 and 2.2 on its domain.

Remark 4.1. Below, we define the continuous extension of f_j on $[S_j, S_{j+1}]$ by the value of $f_j(x)$ on $(S_j, S_{j+1}]$ and by $\lim_{x \downarrow S_j} f_j(x)$ for the point $x = S_j$. Moreover, at the ends of the interval, we use differentiability and the corresponding notation $\dot{f}_j$ to mean the one-sided derivatives.

Proposition 4.2. *The continuous extension $f_0 : (S_0, S_1] \to \mathbb{R}$ (as defined in (4.2)) to $[S_0, S_1]$ is continuous on its domain, bounded from below by $-3m_0/32$, and differentiable on its domain with $\dot{f}_0(S_0) = \dot{f}_0(S_1) = -1$, and its derivative is locally Lipschitz continuous. Similarly, the* *continuous extension of $f_j:(S_j,S_{j+1}]\to\mathbb{R}$ (as defined in (4.3)) to $[S_j,S_{j+1}]$ is continuous on its domain, bounded from below by $f_{j-1}(S_j)-3m_j/32$, and differentiable on its domain with $\dot f_j(S_j)=-1$ and $\dot f_j(S_{j+1})=-1$, and its derivative is locally Lipschitz continuous.*

Proof. We only look at an arbitrary $j\in\mathbb{N}$ as the proof is identical for f_0. Moreover, since f_j is equal to its reflection across the vertical axis $x=S_j+m_j/2$ followed by a reflection over the horizontal axis $y=m_j/4+f_{j-1}(S_j)$, it is enough to show continuity and differentiability on $[S_j,S_j+m_j/2]$.

To establish continuity, we need to show that the left-sided limits of f_j agree with the function value at $S_j+\delta m_j/16$ for $\delta=1,3,8$. Starting with $\delta=1$, $\lim_{x\uparrow S_j+m_j/16}-x+S_j+f_{j-1}(S_j)=-m_j/16+f_{j-1}(S_j)$. By direct substitution,

$$f_j(S_j+m_j/16)=\frac{8}{m_j}\left(\frac{m_j}{16}\right)^2-\frac{3m_j}{32}+f_{j-1}(S_j)=\frac{m_j}{32}-\frac{3m_j}{32}+f_{j-1}(S_j). \tag{4.4}$$

Hence, the left limit agrees with the function value at $\delta=1$. For $\delta=3$, the symmetry and continuity of the quadratic function implies that $\lim_{x\uparrow S_j+3m_j/16}f_j(x)=-m_j/16+f_{j-1}(S_j)$. By direct substitution,

$$f_j(S_j+3m_j/16)=-\frac{5m_j}{16}\exp\left(\frac{5m_j/16}{-5m_j/16}+1\right)+\frac{4m_j}{16}+f_{j-1}(S_j)=-\frac{m_j}{16}+f_{j-1}(S_j). \tag{4.5}$$

Hence, the left limit agrees with the function value at $\delta=3$. For $\delta=8$,

$$\lim_{x\uparrow S_j+m_j/2}-\frac{5m_j}{16}\exp\left(\frac{5m_j/16}{x-S_j-m_j/2}+1\right)+\frac{m_j}{4}+f_{j-1}(S_j)=\frac{m_j}{4}+f_{j-1}(S_j), \tag{4.6}$$

which is just $f_j(S_j+m_j/2)$. Hence, the continuous extension of f_j is continuous on $[S_j,S_{j+1}]$.

We now compute the derivatives of the components of f_j on $[S_j,S_{j+1}]$ with the convention of assigning the one-sided derivative to the component function that includes its end point. Let $x'=x-S_j$.

$$\dot f_j(x)=\begin{cases}-1, & x'\in\left[0,\frac{m_j}{16}\right),\\ \frac{16}{m_j}\left(x'-\frac{m_j}{8}\right), & x'\in\left[\frac{m_j}{16},\frac{3m_j}{16}\right),\\ \left(\frac{5m_j/16}{x'-m_j/2}\right)^2\exp\left(\frac{5m_j/16}{x'-m_j/2}+1\right), & x'\in\left[\frac{3m_j}{16},\frac{m_j}{2}\right),\\ \left(\frac{5m_j/16}{x'-m_j/2}\right)^2\exp\left(\frac{-5m_j/16}{x'-m_j/2}+1\right), & x'\in\left(\frac{m_j}{2},\frac{13m_j}{16}\right),\\ -\frac{16}{m_j}\left(x'-\frac{7m_j}{8}\right), & x'\in\left[\frac{13m_j}{16},\frac{15m_j}{16}\right),\\ -1, & x'\in\left[\frac{15m_j}{16},m_j\right].\end{cases} \tag{4.7}$$

In order to extend these component derivatives to the continuous extension of f_j, we need to verify that the left-hand limits of the derivatives agree with the right-side derivatives at

$S_j + \delta m_j/16$ for $\delta = 1, 3$ of (4.7), and we need to verify that the left-hand limit of the derivative is 0 at $\delta = 8$ of (4.7). Starting with $\delta = 1$, the left-hand limit is -1 and, by direct calculation,

$$\frac{16}{m_j}\left(S_j + \frac{m_j}{16} - S_j - \frac{m_j}{8}\right) = -1. \tag{4.8}$$

For $\delta = 3$, the left-hand limit is

$$\lim_{x \uparrow S_j + 3m_j/16} \frac{16}{m_j}\left(x - S_j - \frac{m_j}{8}\right) = 1, \tag{4.9}$$

and a direct evaluation of the third component (4.7) is

$$\left(\frac{5m_j/16}{S_j + 3m_j/16 - S_j - m_j/2}\right)^2 \exp\left(\frac{5m_j/16}{S_j + 3m_j/16 - S_j - m_j/2} + 1\right) = 1. \tag{4.10}$$

For $\delta = 8$, we need to check that the left-hand limit is zero, which can be confirmed by checking that the argument of the exponential term goes to $-\infty$ as $x \uparrow S_j + 8m_j/16$.

Overall, the derivative of the continuous extension of f_j is well defined at every point on its interval, is continuous, and is given by (with $x' = x - S_j$)

$$\dot{f}_j(x) = \begin{cases} -1, & x' \in \left[0, \frac{m_j}{16}\right), \\ \frac{16}{m_j}\left(x' - \frac{m_j}{8}\right), & x' \in \left[\frac{m_j}{16}, \frac{3m_j}{16}\right), \\ \left(\frac{5m_j/16}{x' - m_j/2}\right)^2 \exp\left(\frac{5m_j/16}{x' - m_j/2} + 1\right), & x' \in \left[\frac{3m_j}{16}, \frac{m_j}{2}\right), \\ 0, & x' = m_j/2, \\ \left(\frac{5m_j/16}{x' - m_j/2}\right)^2 \exp\left(\frac{-5m_j/16}{x' - m_j/2} + 1\right), & x' \in \left(\frac{m_j}{2}, \frac{13m_j}{16}\right), \\ -\frac{16}{m_j}\left(x' - \frac{7m_j}{8}\right), & x' \in \left[\frac{13m_j}{16}, \frac{15m_j}{16}\right), \\ -1, & x' \in \left[\frac{15m_j}{16}, m_j\right]. \end{cases} \tag{4.11}$$

Using this derivative, we can calculate the lower bound for the function. By the derivative of the extension of f_j, (4.11), we see that the function is decreasing only on $[S_j, S_j + m_j/8]$ and $[S_j + 7m_j/8, S_{j+1}]$. Moreover, $f_j(S_j + m_j/8) = -3m_j/32 + f_{j-1}(S_j)$ and $f_j(S_{j+1}) = m_j/2 + f_{j-1}(S_j)$. Thus, the lower bound of the extension of f_j is as stated.

Our last step is to verify the local Lipschitz continuity of the derivative. It is easy to verify that, within its interval, the components are twice continuously differentiable. As a result, we can use Lemma SM2.1. Similarly, if we define the second derivative at $S_j + m_j/2$ to be 0, we can verify that the objective is twice continuously differentiable at $S_j + m_j/2$. Then, we can use Lemma SM2.1 again. To conclude, we need to examine what happens around the points $S_j + \delta m_j/16$ for $\delta = 1, 3$.

Starting with $\delta = 1$, consider the points $S_j + m_j/16 - \epsilon_1 m_j/16$ and $S_j + m_j/16 + \epsilon_2 m_j/16$ for $\epsilon_1, \epsilon_2 > 0$ sufficiently small. Then, the difference in the derivatives at these points divided by the distance between the points is

$$\frac{\left|\frac{16}{m_j}\left(-\frac{m_j}{16} + \epsilon_2 \frac{m_j}{16}\right) + 1\right|}{(\epsilon_2 + \epsilon_1) m_j/16} = \frac{\epsilon_2}{(\epsilon_2 + \epsilon_1) m_j/16} \leq \frac{16}{m_j}. \tag{4.12}$$

Therefore, we conclude that the derivative is locally Lipschitz near $S_j + m_j/16$. For $\delta = 3$, we compute the same ratio at the points $S_j + 3m_j/16 - \epsilon_1 m_j/16$ and $S_j + 3m_j/16 + 5\epsilon_2 m_j/16$ for $\epsilon_1, \epsilon_2 \in [0, 1/4]$, where at most either ϵ_1 or ϵ_2 is zero. The ratio of the difference in the derivatives and the points is

$$\frac{\left|\frac{1}{(1-\epsilon_2)^2} \exp\left(\frac{\epsilon_2}{\epsilon_2 - 1}\right) - 1 + \epsilon_1\right|}{(5\epsilon_2 + \epsilon_1) m_j/16} \leq \frac{\epsilon_2 + \epsilon_2^2/2 + \epsilon_1}{(5\epsilon_2 + \epsilon_1) m_j/16} \leq \frac{16}{m_j}. \tag{4.13}$$

Therefore, we conclude that the derivative is locally Lipschitz near $S_j + 3m_j/16$. ■

With this calculation complete, we can now verify that F satisfies Assumptions 2.1 and 2.2.

Proposition 4.3. *The function $F : \mathbb{R} \to \mathbb{R}$ is continuous and differentiable on its domain; the function F is lower bounded; the derivative of the function F is locally Lipschitz continuous; $F(S_j) = S_j/2$; and $\dot{F}(S_j) = -1$ for all $j + 1 \in \mathbb{N}$. Also, the derivative of the function F is not globally Lipschitz continuous.*

Proof. By Proposition 4.2, in order to verify the continuity of F on $\mathbb{R}$, it is enough to check its continuity at the points $x = S_j$ for all j. Since $F(S_j) = f_{j-1}(S_j)$, we must check that the right-side limit of F at $x = S_j$ converges to $f_{j-1}(S_j)$. That is,

$$\lim_{x \downarrow S_j} F(x) = \lim_{x \downarrow S_j} f_j(x) = \lim_{x \downarrow S_j} -x + S_j + f_{j-1}(S_j) = f_{j-1}(S_j). \tag{4.14}$$

Thus, F is continuous at S_j for each $j \in \mathbb{N}$. We check $x = S_0 = 0$ as well. $F(0) = 0$ by definition. Moreover,

$$\lim_{x \downarrow 0} F(x) = \lim_{x \downarrow 0} f_0(x) = \lim_{x \downarrow 0} -x = 0. \tag{4.15}$$

Hence, F is continuous on its domain.

Similarly, by Proposition 4.2, to verify the differentiability of F on $\mathbb{R}$, it is enough to verify the differentiability of F at $x = S_j$ for all j. By Proposition 4.2, it follows that the derivative at each S_j is -1 from the left and the right for each $j \in \mathbb{N}$. Hence, the derivative exists at S_j for each $j + 1 \in \mathbb{N}$. Moreover, since the derivative of F is constant in a small neighborhood of S_j for each $j + 1 \in \mathbb{N}$, it is Lipschitz continuous in this region. Finally, $\dot{F}(S_j) = -1$ for all $j + 1 \in \mathbb{N}$.

To show that F is lower bounded, we will first calculate the values of $F(S_j)$. We proceed by induction. For the base case, $F(S_0) = 0 = S_0/2$. Suppose that the statement holds up to j. Then, $F(S_j) = S_j/2$. By construction, $f_{j-1}(S_j) = F(S_j) = S_j/2$. Now,

$$F(S_{j+1}) = f_j(S_{j+1}) = -S_{j+1} + S_j + \frac{3m_j}{2} + f_{j-1}(S_j) \tag{4.16}$$

$$= -S_j - m_j + S_j + \frac{3m_j}{2} + \frac{S_j}{2} = \frac{S_{j+1}}{2}. \tag{4.17}$$

To show the lower-bound property, recall that $f_j(x) \geq f_{j-1}(S_j) - 3m_j/32 = F(S_j) - 3m_j/32 = S_j/2 - 3m_j/32$. Since $S_j \to \infty$ and $m_j \to 0$ by construction, $F(x) \geq \inf_j S_j/2 - 3m_j/32 > -\infty$.

Finally, we verify that F is not globally Lipschitz continuous. For a contradiction, suppose that there exists an $L > 0$ such that, for any $x, x' \in \mathbb{R}$, $|\dot{F}(x) - \dot{F}(x')| \leq L|x - x'|$. Since $m_j \to 0$, there exists $j \in \mathbb{N}$ such that $Lm_j/2 < 1$. Then, $|\dot{F}(S_j + m_j/2) - \dot{F}(S_j)| \leq Lm_j/2 < 1$. However, by Proposition 4.2, $\dot{F}(S_j + m_j/2) = \dot{f}_j(S_j + m_j/2) = 0$ and $\dot{F}(S_j) = \dot{f}_j(S_j) = -1$, which implies that $|\dot{F}(S_j + m_j/2) - \dot{F}(S_j)| = 1$, which is a contradiction. ■

4.3. Properties of gradient descent on the objective function.

We are now ready to show that gradient descent with diminishing step sizes generates iterates such that the iterates diverge, the sequence of objective function values evaluated at the iterates diverges, and the sequence of gradient function values evaluated at the iterates remains bounded away from zero.

Proposition 4.4. *Let $\{m_k : k+1 \in \mathbb{N}\}$ be any positive sequence such that $\sum_k m_k$ diverges and $m_k \to 0$. Define $F : \mathbb{R} \to \mathbb{R}$ as in (4.1). Suppose that $x_0 = 0$, and let $\{x_k : k \in \mathbb{N}\}$ be generated according to (2.3) with $M_k = m_k I$ for all $k+1 \in \mathbb{N}$. Then, $\{M_k\}$ satisfies Properties 2.4, 2.5, and 2.6. Moreover,* (a) $\lim_k x_k = \infty$, (b) $\lim_k F(x_k) = \infty$, *and* (c) $\lim_k |\dot{F}(x_k)| = 1$.

Proof. To prove the result, we need only show that $x_k = S_k$, where we recall that $S_0 = 0$ and $S_k = \sum_{j=0}^{k-1} m_j$. For $k = 0$, $x_0 = 0 = S_0$. Suppose that this holds up to k. Then, by Proposition 4.3,

$$x_{k+1} = x_k - M_k \dot{F}(x_k) = S_k - m_k(-1) = S_{k+1}. \tag{4.18}$$

Now, since S_k diverges, the iterates diverge (part (a)). Moreover, by Proposition 4.3, since $F(x_k) = F(S_k) = S_k/2$, the objective function also diverges (part (b)). Finally, by Proposition 4.3, $\dot{F}(x_k) = \dot{F}(S_k) = -1$ (part (c)). ■

In summary, as the example from Proposition 4.4 and the example of $F(x) = \exp(-x^2)$ show, under our assumptions about the objective function and properties of gradient descent, we cannot conclude anything additional about the objective behavior of the function or the gradient in the regime where the iterates generated by gradient descent with diminishing step sizes diverge.

5. Conclusion.

In this paper, we have analyzed the global behavior of gradient descent with diminishing step sizes for differentiable nonconvex functions whose gradients are only locally Lipschitz continuous. To the best of our knowledge, we have provided the most general convergence analysis of gradient descent with diminishing step sizes. Specifically, we have shown that the iterates cannot produce erratic behavior in the objective function or gradient function when they persist in a region for sufficiently long, even if they eventually escape. We also construct specific examples to show the types of erratic behaviors that can occur when the iterates escape off to infinity. Our analysis has also raised a number of interesting questions with varying degrees of practical interest.

1. Is there a notion of continuity on the gradients that is appropriate for data science yet more restrictive than Assumption 2.2 for which Theorem 3.6 or Theorem 3.8 hold uniformly over the family of functions specified by this notion of continuity?

2. Is there a choice of step sizes that ensures the subsequential limit points of the iterates is a set that is a singleton?
3. Is there a function class that is necessary and sufficient to avoid the divergence regime and the corresponding erratic behaviors for gradient descent with diminishing step size?

Acknowledgments. The authors thank the detailed and insightful feedback and criticism from the associate editors and reviewers, which has substantially improved the quality of this work.

REFERENCES

[1] L. Armijo, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pac. J. Math., 16 (1966), pp. 1–3.

[2] H. H. Bauschke, J. Bolte, and M. Teboulle, *A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications*, Math. Oper. Res., 42 (2017), pp. 330–348.

[3] A. Beck, *First-Order Methods in Optimization*, SIAM, Philadelphia, 2017.

[4] M. Benaïm, *Dynamics of stochastic approximation algorithms*, in Seminaire de probabilites XXXIII, Lecture Notes in Math. 1709, Springer, Cham 1999, pp. 1–68.

[5] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Stoch. Model. Appl. Probab. 22, Springer, Cham, 2012.

[6] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Nashua, NH, 2016.

[7] S. Bittanti, P. Bolzern, and M. Campi, *Convergence and exponential convergence of identification algorithms with directional forgetting factor*, Automatica, 26 (1990), pp. 929–932.

[8] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd, *First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems*, SIAM J. Optim., 28 (2018), pp. 2131–2151.

[9] L. Bottou, *Large-scale machine learning with stochastic gradient descent*, in Proceedings of COMPSTAT'2010, Springer, Cham, 2010, pp. 177–186.

[10] L. Cao and H. M. Schwartz, *Exponential convergence of the Kalman filter based parameter estimation algorithm*, Int. J. Adapt. Control Signal Process., 17 (2003), pp. 763–783.

[11] C. Cartis, N. I. Gould, and P. L. Toint, *Adaptive cubic regularisation methods for unconstrained optimization. Part* I*: Motivation, convergence and numerical results*, Math. Program., 127 (2011), pp. 245–295.

[12] C. Cartis, N. I. Gould, and P. L. Toint, *Adaptive cubic regularisation methods for unconstrained optimization. Part* II*: Worst-case function-and derivative-evaluation complexity*, Math. Program., 130 (2011), pp. 295–319.

[13] C. Cartis, N. I. Gould, and P. L. Toint, *Evaluation Complexity of Algorithms for Nonconvex Optimization: Theory, Computation and Perspectives*, SIAM, Philadelphia, 2022.

[14] A. Cauchy, *Méthode générale pour la résolution des systemes d'équations simultanées*, C. R. Sci. Paris, 25 (1847), pp. 536–538.

[15] H. B. Curry, *The method of steepest descent for non-linear minimization problems*, Quart. Appl. Math., 2 (1944), pp. 258–261.

[16] F. E. Curtis and K. Scheinberg, *Adaptive stochastic optimization: A framework for analyzing stochastic optimization algorithms*, IEEE Signal Process. Mag., 37 (2020), pp. 32–42.

[17] F. E. Curtis, K. Scheinberg, and R. Shi, *A stochastic trust region algorithm based on careful step normalization*, INFORMS J. Optim., 1 (2019), pp. 200–220.

[18] A. Défossez, L. Bottou, F. Bach, and N. Usunier, *A Simple Convergence Proof of Adam and Adagrad*, preprint, arXiv:2003.02395, 2020.

[19] H. W. Dommel and W. F. Tinney, *Optimal power flow solutions*, IEEE Trans. Power Appar. Syst., (1968), pp. 1866–1876.

[20] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai, *Gradient descent finds global minima of deep neural networks*, in International Conference on Machine Learning, PMLR, 2019, pp. 1675–1685.

[21] S. S. Du, C. Jin, J. D. Lee, M. I. Jordan, A. Singh, and B. Poczos, *Gradient descent can take exponential time to escape saddle points*, Adv. Neural Inf. Process. Syst., 30 (2017), pp. 1–11.

[22] S. S. Du, X. Zhai, B. Poczos, and A. Singh, *Gradient Descent Provably Optimizes Over-Parameterized Neural Networks*, preprint, arXiv:1810.02054, 2018.

[23] J.-C. Fort and G. Pages, *Convergence of stochastic algorithms: From the Kushner–Clark theorem to the Lyapounov functional method*, Adv. Appl. Probab., 28 (1996), pp. 1072–1094.

[24] G. N. Grapiglia and G. F. Stella, *An adaptive trust-region method without function evaluations*, Comput. Optim. Appl., 82 (2022), pp. 31–60.

[25] S. Gratton, S. Jerad, and P. L. Toint, *Convergence Properties of an Objective-Function-Free Optimization Regularization Algorithm, Including an $o(\epsilon^{3/2})$ Complexity Bound*, preprint, arXiv:2203.09947, 2022.

[26] S. Gratton, S. Jerad, and P. L. Toint, *First-Order Objective-Function-Free Optimization Algorithms and Their Complexity*, preprint, arXiv:2203.01757, 2022.

[27] S. Gratton, S. Jerad, and P. L. Toint, *Complexity of a Class of First-Order Objective-Function-Free Optimization Algorithms*, preprint, arXiv:2203.01647, 2022.

[28] J. Hadamard, *Mémoire sur le problème d'analyse relatif à l'équilibre des plaques élastiques encastrées*, Memoires presentes par divers savants a l'Academie des sciences de l'Institut national de France 33, Imprimerie Nationale, Paris, 1908.

[29] G. Iyengar and A. K. C. Ma, *Fast gradient descent method for mean-CVaR optimization*, Ann. Oper. Res., 205 (2013), pp. 203–212.

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

[31] R. M. Johnstone, C. R. Johnson, Jr., R. R. Bitmead, and B. D. Anderson, *Exponential convergence of recursive least squares with exponential forgetting factor*, Systems Control Lett., 2 (1982), pp. 77–82.

[32] C. Josz, *Global convergence of the gradient method for functions definable in o-minimal structures*, Math. Program., (2023), pp. 1–29.

[33] H. Karimi, J. Nutini, and M. Schmidt, *Linear convergence of gradient and proximal-gradient methods under the Polyak–Lojasiewicz condition*, in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2016, pp. 795–811.

[34] X. Ke and J. Han, *A class of nonmonotone trust region algorithms for unconstrained optimization problems*, Sci. China Ser. A: Math., 41 (1998), pp. 927–932.

[35] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington, *Wide neural networks of any depth evolve as linear models under gradient descent*, Adv. Neural Inf. Process. Syst., 32 (2019), 124002.

[36] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, *Gradient descent only converges to minimizers*, in Conference on Learning Theory, PMLR, 2016, pp. 1246–1257.

[37] C. Lemaréchal, *Cauchy and the gradient method*, Doc. Math. Extra, 251 (2012), 10.

[38] H. Li, J. Qian, Y. Tian, A. Rakhlin, and A. Jadbabaie, *Convex and Non-Convex Optimization Under Generalized Smoothness*, preprint, arXiv:2306.01264, 2023.

[39] L. Ljung, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, 22 (1977), pp. 551–575.

[40] H. Lu, R. M. Freund, and Y. Nesterov, *Relatively smooth convex optimization by first-order methods, and applications*, SIAM J. Optim., 28 (2018), pp. 333–354.

[41] P. Mertikopoulos, N. Hallak, A. Kavis, and V. Cevher, *On the Almost Sure Convergence of Stochastic Gradient Descent in Non-Convex Problems*, preprint, arXiv:2006.11144, 2020.

[42] J. J. Moré, *Recent Developments in Algorithms and Software for Trust Region Methods*, Springer, Cham, 1983.

[43] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim., 19 (2009), pp. 1574–1609.

[44] Y. Nesterov, *Universal gradient methods for convex optimization problems*, Math. Program., 152 (2015), pp. 381–404.

[45] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, Cham, 2006.

[46] S. Oymak and M. Soltanolkotabi, *Overparameterized nonlinear learning: Gradient descent takes the shortest path?*, in International Conference on Machine Learning, PMLR, 2019, pp. 4951–4960.

[47] J. Parkum, N. K. Poulsen, and J. Holst, *Recursive forgetting algorithms*, Internat. J. Control, 55 (1992), pp. 109–128.

[48] V. Patel, *Stopping criteria for, and strong convergence of, stochastic gradient descent on Bottou–Curtis–Nocedal functions*, Math. Program., 195 (2021), pp. 693–734.

[49] V. Patel, B. Tian, and S. Zhang, *Global Convergence and Stability of Stochastic Gradient Descent*, preprint, arXiv:2110.01663, 2021.

[50] V. Patel and S. Zhang, *Stochastic Gradient Descent on Nonconvex Functions with General Noise Models*, preprint, arXiv:2104.00423, 2021.

[51] S. Reddi, S. Sra, B. Poczos, and A. J. Smola, *Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization*, Adv. Neural Inf. Process. Syst., 29 (2016), pp. 1145–1153.

[52] S. J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola, *Stochastic variance reduction for nonconvex optimization*, in International Conference on Machine Learning, PMLR, 2016, pp. 314–323.

[53] F. Stonyakin, A. Tyurin, A. Gasnikov, P. Dvurechensky, A. Agafonov, D. Dvinskikh, M. Alkousa, D. Pasechnyuk, S. Artamonov, and V. Piskunova, *Inexact model: A framework for optimization and variational inequalities*, Optim. Methods Softw., 36 (2021), pp. 1155–1201.

[54] C. Varner and V. Patel, *A Novel Gradient Methodology with Economical Objective Function Evaluations for Data Science Applications*, preprint, arXiv:2309.10894, 2023.

[55] S. A. Vavasis, *Black-box complexity of local minimization*, SIAM J. Optim., 3 (1993), pp. 60–80.

[56] R. Ward, X. Wu, and L. Bottou, *AdaGrad stepsizes: Sharp convergence over nonconvex landscapes*, J. Mach. Learn. Res., 21 (2020), pp. 9047–9076.

[57] X. Wu, R. Ward, and L. Bottou, *WNGrad: Learn the Learning Rate in Gradient Descent*, preprint, arXiv:1803.02865, 2018.

[58] J. Zhang, Y. Wang, and X. Zhang, *Superlinearly convergent trust-region method without the assumption of positive-definite Hessian*, J. Optim. Theory Appl., 129 (2006), pp. 201–218.

[59] J. Zhang, L. Wu, and X. Zhang, *A trust region method for optimization problem with singular solutions*, Appl. Math. Optim., 56 (2007), pp. 379–394.

[60] G. Zoutendijk, *Methods of Feasible Directions: A Study in Linear and Non-Linear Programming*, Elsevier, Amsterdam, 1960.

SUPPLEMENTARY MATERIALS: Gradient Descent in the Absence of Global Lipschitz Continuity of the Gradients[*]

Vivak Patel[†] and Albert S. Berahas[‡]

SM1. Equivalent Definitions for Local Lipschitz Continuity.

Lemma SM1.1. *A function $G : \mathbb{R}^p \to \mathbb{R}^p$ is locally Lipschitz continuous if and only if for every compact set $\mathcal{C} \subset \mathbb{R}^p$, there exists an $L \geq 0$ such that*

$$\frac{\|G(y) - G(z)\|_2}{\|y - z\|_2} \leq L, \ \forall y, z \in \mathcal{C}. \tag{SM1.1}$$

Proof. Suppose G is locally Lipschitz continuous. Suppose for a contradiction, there exists a compact set, $\mathcal{C}$, for which no such L exists. Then for every $\ell \in \mathbb{N}$, we can find a pair $y_\ell, z_\ell \in \mathcal{C}$ such that

$$\frac{\|G(y_\ell) - G(z_\ell)\|_2}{\|y_\ell - z_\ell\|_2} > \ell. \tag{SM1.2}$$

By compactness, there exists a subsequence $\{\ell_k : k \in \mathbb{N}\}$ and $y, z \in \mathcal{C}$ such that $y_{\ell_k} \to y$ and $z_{\ell_k} \to z$ as $k \to \infty$. If $\|y - z\|_2 > 0$, then, for $k \in \mathbb{N}$ sufficiently large,

$$\frac{\|G(y_{\ell_k}) - G(z_{\ell_k})\|_2}{\|y_{\ell_k} - z_{\ell_k}\|_2} \leq \frac{2\sup_{x \in \mathcal{C}} \|G(x)\|_2}{0.5 \|y - z\|_2} < \infty, \tag{SM1.3}$$

which is a contradiction. Hence, $\|y - z\|_2 = 0$; that is, $y = z$. This also provides a contradiction as G is locally Lipschitz continuous at $y = z$ and so for $k \in \mathbb{N}$ sufficiently large, y_{ℓ_k} and z_{ℓ_k} would be inside of $\mathcal{N}$ from Definition 2.3.

For the other direction of the result: for any point $x \in \mathbb{R}^p$ and any open ball containing x, we can take the closure of this open ball to generate a compact set $\mathcal{C}$. The result follows. ■

SM2. Continuous Hessians Implies Local Lipschitz Continuity.

Lemma SM2.1. *Suppose F is twice continuously differentiable for all $x \in \mathbb{R}^p$. Then $\dot{F}(x)$ is locally Lipschitz continuous.*

Proof. Let $\ddot{F}(x)$ denote the Hessian of F. Then, by assumption, $\|\ddot{F}(x)\|_2$ is a continuous function and it is bounded over any compact region. By Taylor's theorem, for any $x, y \in \mathbb{R}^p$, $\dot{F}(x) - \dot{F}(y) = \int_0^1 \ddot{F}(y + t(x - y))(x - y)dt$. Let $K \subset \mathbb{R}^p$ be compact. By continuity and

[*] Supplementary material for SIMODS MS#M152721.
https://doi.org/10.1137/22M1527210

[†] Department of Statistics, University of Wisconsin – Madison, Madison, WI 53706 USA (vivak.patel@wisc.edu, http://vivakpatel.org).

[‡] Department of Industrial and Operations Engineering, University of Michigan – Ann Arbor, Ann Arbor, MI 48109 USA (albertberahas@gmail.com, https://aberahas.engin.umich.edu).

compactness, there exists an L for K such that $\|\ddot{F}(x)\|_2 \leq L$ for all $x \in K$. Hence, by Hölder's inequality, for any $x, y \in K$, $\|\dot{F}(x) - \dot{F}(y)\| \leq L\|x - y\|_2$. As K is arbitrary, the result follows. ■

SM3. Some Properties of Subsequential Limits.

Lemma SM3.1. *Let $\{a_n : n \in \mathbb{N}\} \subset \mathbb{R}^p$. Let $\mathcal{C}$ be the set of its subsequential limits. Then $\mathcal{C}$ is closed.*

Proof. Let z be a limit point of $\mathcal{C}$. Then, we can construct a sequence $\{z_k : k \in \mathbb{N}\} \subset \mathcal{C}$ such that for every $K \in \mathbb{N}$ and for all $k \geq K$, $\|z_k - z\|_2 \leq 2^{-K-1}$. Moreover, since $z_k \in \mathcal{C}$, $\exists n_k \in \mathbb{N}$ such that $\|a_{n_k} - z_k\|_2 \leq 2^{-k-1}$. Let $\epsilon > 0$ and let $K \in \mathbb{N}$ such that $2^{-K} < \epsilon$. Then, $\forall k \geq K$, $\|a_{n_k} - z\|_2 \leq \|a_{n_k} - z_k\|_2 + \|z_k - z\|_2 \leq 2^{-K} < \epsilon$. Hence, $z = \lim_k a_{n_k} \in \mathcal{C}$. ■

Lemma SM3.2. *Let $\{a_n : n \in \mathbb{N}\} \subset \mathbb{R}$ such that $\liminf_n a_n$ and $\limsup_n a_n$ are finite. If $\lim_n a_{n+1} - a_n = 0$, then for any $z \in [\liminf_n a_n, \limsup_n a_n]$, there is a subsequence of $\{a_n : n \in \mathbb{N}\}$ that converges to z.*

Proof. We begin by showing that any closed interval strictly between the limit infimum and limit supremum contains a subsequential limit. Let $r_1 < r_2$ such that $\liminf_n a_n < r_1$ and $r_2 < \limsup_n a_n$. If there exists an infinite subsequence $\{a_{n_k} : k \in \mathbb{N}\} \subset [r_1, r_2]$, then sequential compactness implies that $\{a_{n_k} : k \in \mathbb{N}\}$ has a subsequence which converges in $[r_1, r_2]$. Suppose now, $\exists K \in \mathbb{N}$ such that $\forall n \geq K$, $a_n \not\in [r_1, r_2]$. Since the $\liminf_n a_n < r_1 < r_2 < \limsup_n a_n$, there exists a subsequence $\{a_{n_k} : k \in \mathbb{N}\}$ such that $a_{n_k} < r_1$ and $r_2 < a_{n_k+1}$. However, this is a contradiction since $a_{n_k+1} - a_{n_k} \to 0$ as $k \to \infty$. Hence, there is always a subsequence in any closed interval between $\liminf_n a_n$ and $\limsup_n a_n$.

We have that if z is either the limit infimum or limit supremum then there is a subsequence of $\{a_n : n \in \mathbb{N}\}$ that converges to this value. So take $\liminf_n a_n < z < \limsup_n a_n$. We now proceed by induction. Let $z_0 = \liminf_n a_n$. There is a subsequence that converges to a point in $[0.5(z + z_0), z]$. Let z_1 be this limit. If $z \neq z_1$, then $|z_1 - z| \leq 2^{-1}(z - z_0)$ and we define z_2 as the subsequential limit in $[0.5(z + z_1, z]$. If $z = z_1$ then we stop. Suppose we proceed by induction such that $\{z_j : j = 1, \ldots, k\}$ are subsequential limits such that $|z_j - z| \leq 2^{-j}(z - z_0)$. If $z \neq z_k$, then we can find z_{k+1} as the limit of a subsequence in $[0.5(z + z_k), z]$, which we denote z_{k+1}. Moreover, $|z_{k+1} - z| \leq 2^{-k-1}(z - z_0)$. If we never terminate at z for some $k \in \mathbb{N}$, then $\{z_k : k \in \mathbb{N}\}$ is a sequence of subsequential limits converging to z. By Lemma SM3.1, z is a subsequential limit. ■

SM4. Divergence Regime: Nonexistence of Objective Function Limit.

Here, we use as similar construction for Proposition 4.4 to construct an objective function F such that when gradient descent is applied to this objective function with a specific initialization, $\limsup_k F(x_k) = \infty$, $\liminf_k F(x_k) = 0$ and $|\dot{F}(x_k)| = 1$ for all k. We proceed in three general steps corresponding to each subsection below.

SM4.1. Objective Function Target Values.

Let $\{m_k : k+1 \in \mathbb{N}\}$ be a sequence of scalars such that $m_k > 0$, $\sum_k m_k = \infty$, and $m_k \to 0$ as $k \to \infty$. Define $S_0 = 0$ and $S_{k+1} = \sum_{j=0}^{k} m_k$ for all integers $k \geq 0$. We will now construct a sequence $\{O_k : k+1 \in \mathbb{N}\}$ which will serve as target values for each iterate of our objective function.

1. Let $O_0 = 0$. For convenience, let $u_0 = \ell_0 = 0$.

2. Let $\ell_1 = 1 + \min\{k \geq 0 : O_0 + \frac{1}{2}\sum_{j=0}^{k} m_j > 1\}$. From the divergence of $\sum_k m_k$, it is clear that such an ℓ_1 is finite. Define $O_k = O_0 + \frac{1}{2}\sum_{j=0}^{k-1} m_j$ for $k \in [1, \ell_1] \cap \mathbb{N}$.

3. Let $u_1 = 1 + \min\{k \geq \ell_1 : O_{\ell_1} - \sum_{j=\ell_1}^{k} m_j < 0\}$. Again, from the divergence of $\sum_k m_k$, u_1 is finite. Define $O_k = O_{\ell_1} - \sum_{j=\ell_1}^{k-1} m_j$ for $k \in [\ell_1 + 1, u_1] \cap \mathbb{N}$.

4. For $t \in \mathbb{N}$, let $\ell_{t+1} = 1 + \min\{k \geq u_t : O_{u_t} + \frac{1}{2}\sum_{j=u_t}^{k} m_j > t+1\}$. From the divergence of $\sum_k m_k$, ℓ_{t+1} is finite if u_t is finite. Define $O_k = O_{u_t} + \frac{1}{2}\sum_{j=u_t}^{k-1} m_j$ for $k \in [u_t + 1, \ell_{t+1}] \cap \mathbb{N}$.

5. For $t \in \mathbb{N}$, let $u_{t+1} = 1 + \min\{k \geq \ell_{t+1} : O_{\ell_{t+1}} - \sum_{j=\ell_{t+1}}^{k} m_j < 0\}$. From the divergence of $\sum_k m_k$, u_{t+1} is finite if ℓ_{t+1} is finite. Define $O_k = O_{\ell_{t+1}} - \sum_{j=\ell_{t+1}}^{k-1} m_j$ for $k \in [\ell_{t+1} + 1, \ldots, u_{t+1}] \cap \mathbb{N}$.

We point out several facts about the sequence $\{O_k : k+1 \in \mathbb{N}\}$. First, $\lim_t O_{\ell_t} = \infty$ by construction. Second, we verify, $\lim_t O_{u_t} = 0$. By construction, $O_{u_t-1} > 0$ and $0 > O_{u_t} = O_{u_t-1} - m_{u_t-1} \geq -m_{u_t-1}$. Since $m_{u_t-1} \to 0$ as $t \to \infty$, $\liminf_t O_{u_t} = 0$. In turn, the limit of the sequence exists and is zero. Third, we verify, $\limsup_k O_k = \infty$ and $\liminf_k O_k = 0$. For any $k \in \mathbb{N}_{>\ell_1}$, there exists a $t \in \mathbb{N}$ such that $k \in [\ell_t + 1, u_t]$ or $k \in [u_t + 1, \ell_{t+1}]$. If $k \in [\ell_t + 1, u_t]$, then $O_k \in [O_{u_t}, O_{\ell_t}]$. If $k \in [u_t + 1, \ell_{t+1}]$, then $O_k \in [O_{u_t}, O_{\ell_{t+1}}]$. Hence, the third fact holds because of the first two.

SM4.2. Construction of the Objective Function.

With these sequences established, we now state our objective function.

$$F(x) = \begin{cases} -x & x \leq 0, \\ \tilde{f}_0(x) & x \in (0, S_{\ell_1}], \\ \tilde{f}_t(x) & x \in (S_{\ell_t}, S_{\ell_{t+1}}],\ \forall t \in \mathbb{N}, \end{cases} \tag{SM4.1}$$

where

$$\tilde{f}_0(x) = \begin{cases} f_j(x) & x \in (S_j, S_{j+1}],\ j \in \{0, \ldots, \ell_1 - 1\}; \end{cases} \tag{SM4.2}$$

$$\tilde{f}_t(x) = \begin{cases} O_{\ell_t} - (x - S_{\ell_t}) & x \in (S_{\ell_t}, S_{u_t}] \\ f_j(x) & x \in (S_j, S_{j+1}],\ j \in \{u_t, \ldots, \ell_{t+1} - 1\}; \end{cases} \tag{SM4.3}$$

and

$$f_j(x) = \begin{cases} -x' + O_j & x' \in (0, \frac{m_j}{16}) \\ \frac{8}{m_j}(x' - \frac{m_j}{8})^2 - \frac{3m_j}{32} + O_j & x' \in [\frac{m_j}{16}, \frac{3m_j}{16}) \\ -\frac{5m_j}{16} \exp\left(\frac{5m_j/16}{x' - m_j/2} + 1\right) + \frac{m_j}{4} + O_j & x' \in [\frac{3m_j}{16}, \frac{m_j}{2}) \\ \frac{m_j}{4} + O_j & x' = \frac{m_j}{2} \\ \frac{5m_j}{16} \exp\left(\frac{-5m_j/16}{x' - m_j/2} + 1\right) + \frac{m_j}{4} + O_j & x' \in (\frac{m_j}{2}, \frac{13m_j}{16}) \\ \frac{-8}{m_j}(x' - \frac{7m_j}{8})^2 + \frac{19m_j}{32} + O_j & x' \in [\frac{13m_j}{16}, \frac{15m_j}{16}) \\ -x' + \frac{3m_j}{2} + O_j & x' \in [\frac{15m_j}{16}, m_j]. \end{cases} \tag{SM4.4}$$

with $x' = x - S_j$.

SM4.3. Properties of the Objective Function. Here, we verify, (SM4.1) satisfies Assumption 2.1 and 2.2. We need to verify certain properties of $\tilde{f}_t(x)$, which we do now.

Proposition SM4.1. *Let $t+1 \in \mathbb{N}$. The continuous extension of $\tilde{f}_t : (S_{\ell_t}, S_{\ell_{t+1}}] \to \mathbb{R}$, (SM4.3), to $[S_{\ell_t}, S_{\ell_{t+1}}]$ is*

1. *continuous on $[S_{\ell_t}, S_{\ell_{t+1}}]$ with values O_{ℓ_t}, O_{u_t} and $O_{\ell_{t+1}}$ at points S_{ℓ_t}, S_{u_t} and $S_{\ell_{t+1}}$, respectively;*
2. *bounded from below by $\min\{O_j - \frac{3m_j}{32} : j = u_t, \ldots, s_{\ell_{t+1}-1}\}$;*
3. *differentiable on $[S_{\ell_t}, S_{\ell_{t+1}}]$ with the one-sided derivatives being -1 at the end points of the interval;*
4. *locally Lipschitz continuous.*

Proof. We note that (SM4.3) has several components. The $f_j(x)$ are the same as those defined by (4.2) but shifted vertically by a constant. Hence, by Proposition 4.2, the continuous extension of $f_j(x)$ to $[S_j, S_{j+1}]$ is continuous; bounded from below by $O_j - \frac{3m_j}{32}$; differentiable with the one-sided derivatives being -1 on the end points of the interval; and locally Lipschitz continuous.

We use these facts to show the remaining properties of $\tilde{f}_t(x)$. First, to verify continuity, we need only verify that the components agree at the points $x \in \{S_{u_t}, S_{u_t+1}, \ldots, S_{\ell_{t+1}-1}\}$. When $x = S_{u_t}$,

$$\tilde{f}_t(S_{u_t}) = O_{\ell_t} + (S_{u_t} - S_{\ell_t}) = O_{\ell_t} + \left(\sum_{k=0}^{u_t-1} m_k - \sum_{k=0}^{\ell_t-1} m_k \right) = O_{\ell_t} + \sum_{k=\ell_t}^{u_t-1} = O_{u_t}. \tag{SM4.5}$$

Moreover,

$$\lim_{x\downarrow S_{u_t}} \tilde{f}_t(x) = \lim_{x\downarrow S_{u_t}} f_{u_t}(x) = \lim_{x\downarrow S_{u_t}} -(x - S_{u_t}) + O_{u_t} = O_{u_t}. \tag{SM4.6}$$

Hence, the evaluation of $\tilde{f}_t(x)$ at S_{u_t} agrees with its limit from the right. For the remaining points, let $j \in \{u_t + 1, \ldots, \ell_{t+1} - 1\}$. Then,

$$\tilde{f}_t(S_j) = f_{j-1}(S_j) = -(S_j - S_{j-1}) + \frac{3m_{j-1}}{2} + O_{j-1} = \frac{m_{j-1}}{2} + O_{u_t} + \frac{1}{2}\sum_{k=u_t}^{j-1} m_k = O_j. \tag{SM4.7}$$

Moreover,

$$\lim_{x\downarrow S_j} \tilde{f}_t(S_j) = \lim_{x\downarrow S_j} f_j(S_j) = \lim_{x\downarrow S_j} -(x - S_j) + O_j = O_j. \tag{SM4.8}$$

Hence, $\tilde{f}(x)$ is continuous. Moreover, we have also shown that the continuous extension of $\tilde{f}(x)$ has the stated values at $x \in \{S_{u_t}, S_{u_t+1}, \ldots, S_{\ell_{t+1}-1}\}$.

For the lower bound, we have that $\tilde{f}(x) \geq O_{u_t}$ for $x \in (S_{\ell_t}, S_{u_t}]$. By Proposition 4.2, each $f_j(x) \geq O_j - \frac{3m_j}{32}$. Hence, the lower bound follows.

We now verify differentiability. By the properties of a linear function and Proposition 4.2, each component of $\tilde{f}_t(x)$ is differentiable on its domain. We must check that these derivatives agree at $x \in \{S_{u_t}, S_{u_t+1}, \ldots, S_{\ell_{t+1}-1}\}$. For the linear function, the derivative is a constant of -1, and the continuous extension of $f_j(x)$ has derivative of -1 at each end of its intervals. Thus, the extension of $\tilde{f}_t(x)$ is differentiable and the one-sided derivatives are -1 at the end of the interval on which it is defined.

To check local Lipschitz continuity of $\tilde{f}_t(x)$, we note that each component of $\tilde{f}_t(x)$ is locally Lipschitz continuous in its domain either because it is a linear function or by Proposition 4.2. Hence, we need to only check that local Lipschitz continuity holds for each $x \in \{S_{u_t}, S_{u_t+1}, \ldots, S_{\ell_{t+1}-1}\}$. For $j \in \{u_t, \ldots, \ell_{t+1}-1\}$, the derivative of $\tilde{f}_t(x)$ is -1 in $(S_j - m_j/32, S_j + m_j/32)$. Hence, the derivative is locally Lipschitz continuous at the stated values of x. ■

Proposition SM4.2. *The function $F : \mathbb{R} \to \mathbb{R}$ as defined in* (SM4.1) *is continuous and differentiable on its domain; it is lower bounded; its derivative is locally Lipschitz continuous; $F(S_{\ell_t}) = O_{\ell_t}$, $\forall t \in \mathbb{N}$; $F(S_{u_t}) = O_{u_t}$ $\forall t \in \mathbb{N}$; and F's derivative is not globally Lipschitz continuous.*

Proof. The proof is similar to Proposition 4.3. Hence, we will only verify that F is lower bounded. By Proposition SM4.1, the component $\tilde{f}_t(X)$ of F for some $t+1 \in \mathbb{N}$ is bounded from below by some $O_j - \frac{3m_j}{32}$ for some choice of j. So it is enough for us to show, $\{O_j - \frac{3m_j}{32}\}$ is bounded from below. By construction, $\liminf_j O_j = 0$ and $\lim_j m_j = 0$. Hence, $\{O_j - \frac{3m_j}{32}\}$ is bounded from below. Thus, F is bounded from below. ■

SM4.4. Properties of Gradient Descent on the Objective Function.

We now show that when gradient descent is applied to the constructed problem, the objective function's limit supremum is infinite and limit infimum is zero, all while the gradient function remains bounded away from 0.

Proposition SM4.3. *Let $\{m_k : k+1 \in \mathbb{N}\}$ be any positive sequence such that $\sum_k m_k$ diverges and $m_k \to 0$. Define $F : \mathbb{R} \to \mathbb{R}$ as in* (SM4.1)*. Suppose $x_0 = 0$ and let $\{x_k : k \in \mathbb{N}\}$ be generated according to* (2.3) *with $M_k = m_k I$ for all $k+1 \in \mathbb{N}$. Then, $\{M_k\}$ satisfies Properties* 2.4 *and* 2.6*. Moreover, (a) $\lim_k x_k = \infty$; (b) $\limsup_k F(x_k) = \infty$; (c) $\liminf_k F(x_k) = 0$; and (d) $\lim_k |\dot{F}(x_k)| = -1$.*

Proof. We first show, $x_k = S_k$ for all $k \in \mathbb{N}$. $0 = x_0 = S_0$. Suppose the claim is true up to $k \in \mathbb{N}$. Then, $\exists t+1 \in \mathbb{N}$ such that $\dot{F}(x_k) = \dot{\tilde{f}}_t(x_k)$. Using Proposition SM4.1 or properties of a linear function, $\dot{F}(x_k) = \dot{F}(S_k) = \dot{\tilde{f}}_t(S_k) = -1$. Therefore,

$$x_{k+1} = x_k - M_k \dot{F}(x_k) = S_k - m_k \dot{\tilde{f}}_t(S_k) = S_k + m_k = S_{k+1}. \tag{SM4.9}$$

Thus, as $k \to \infty$, the iterates diverge and $\dot{F}(x_k) = -1$ for all $k+1 \in \mathbb{N}$. Now, $F(x_k) = F(S_k) = O_k$ for every $k+1 \in \mathbb{N}$. By properties of $\{O_k\}$, the limit supremum and limit infimum of this sequence is ∞ and 0, respectively. The result follows. ■

We stress that the choice of the limit supremum and limit infimum can be readily modified by choosing a different definition for $\{\ell_t\}$ and $\{u_t\}$. Hence, the limit supremum can be made to be finite and even agree with the limit infimum. Moreover, the limit infimum can be set larger than 0.

SM5. Divergence Regime: Objective Function Diverges, Gradient Function Converges to Zero.

Here, we construct an objective function that is bounded below and has locally Lipschitz continuous gradients. Importantly, when we apply gradient descent with diminishing step sizes to this objective function, the iterates of the procedure diverge, the objective function evaluated at the iterates will diverge, and the gradient function will converge to zero. This objective function will be constructed in a similar fashion to our other divergence regime examples.

SM5.1. Construction of the Objective Function.

Let $\{m_k : k+1\}$ be a positive sequence such that $\sum_k m_k$ diverges and $m_k \to 0$. Let $S_0 = 0$ and $S_{k+1} = \sum_{j=0}^k m_j$. We now show by contradiction, $\sum_k \frac{m_k}{S_{k+1}}$ diverges. Suppose $\sum_k \frac{m_k}{S_{k+1}}$ converges, which implies the Cauchy property. Then, using $1/2$, there exists a sufficiently large integer j such that

$$\frac{1}{2} > \sum_{k=j}^{j'} \frac{m_k}{S_{k+1}}, \tag{SM5.1}$$

for any $j' > j$. Given that $\{S_k\}$ is an increasing sequence,

$$\sum_{k=j}^{j'} \frac{m_k}{S_{k+1}} \geq \frac{1}{S_{j'+1}} \sum_{k=j}^{j'} m_k = 1 - \frac{S_j}{S_{j'+1}}. \tag{SM5.2}$$

The right hand side of this equality can be lower bounded by $3/4$ since $S_{j'+1}$ is diverging and S_j is fixed. Hence, we have a contradiction. Therefore, $\sum_k \frac{m_k}{S_{k+1}}$ diverges.

Let $K = \min\{k > 0 : S_k \geq 1\}$ and define

$$T_k = \begin{cases} S_k & k = 0, \ldots, K, \\ T_K + \sum_{j=K}^k \frac{m_j}{S_{j+1}} & k > K. \end{cases} \tag{SM5.3}$$

Moreover, define

$$d_k = \begin{cases} 1 & k = 0, \ldots, K, \\ \frac{1}{S_{k+1}} & k > K. \end{cases} \tag{SM5.4}$$

Finally, let

$$F(x) = \begin{cases} -x & x \leq 0, \\ f_j(x) & x \in (T_j, T_{j+1}],\ j+1 \in \mathbb{N}, \end{cases} \tag{SM5.5}$$

where $f_0(x), \dots, f_{K-1}(x)$ are identical to (4.3); and, letting $x' = x - T_j$,

$$f_j(x) = \begin{cases} -d_j x' + f_{j-1}(T_j) & x' \in \left[0, \frac{(2-d_j)m_j}{16S_{j+1}}\right) \\ \frac{8S_{j+1}}{m_j}\left(x' - \frac{m_j}{8S_{j+1}}\right)^2 - \frac{m_j}{S_{j+1}}\left(\frac{-d_j^2+4d_j}{32}\right) + f_{j-1}(T_j) & x' \in \left[\frac{(2-d_j)m_j}{16S_{j+1}}, \frac{3m_j}{16S_{j+1}}\right) \\ \frac{-5m_j}{16S_{j+1}} \exp\left(\frac{5/16}{S_{j+1}x'/m_j - 1/2} + 1\right) + \frac{m_j}{S_{j+1}}\left(\frac{11+d_j^2-4d_j}{32}\right) + f_{j-1}(T_j) & x' \in \left[\frac{3m_j}{16S_{j+1}}, \frac{m_j}{2S_{j+1}}\right) \\ \frac{m_j}{S_{j+1}}\left(\frac{11+d_j^2-4d_j}{32}\right) + f_{j-1}(T_j) & x' = \frac{m_j}{2S_{j+1}} \\ \frac{5m_j}{16S_{j+1}} \exp\left(\frac{-5/16}{S_{j+1}x'/m_j - 1/2} + 1\right) + \frac{m_j}{S_{j+1}}\left(\frac{11+d_j^2-4d_j}{32}\right) + f_{j-1}(T_j) & x' \in \left(\frac{m_j}{2S_{j+1}}, \frac{13m_j}{16S_{j+1}}\right) \\ \frac{-8S_{j+1}}{m_j}\left(x' - \frac{7m_j}{8S_{j+1}}\right)^2 + \frac{m_j}{S_{j+1}}\left(\frac{22+d_j^2-4d_j}{32}\right) + f_{j-1}(T_j) & x' \in \left[\frac{13m_j}{16S_{j+1}}, \frac{(d_{j+1}+14)m_j}{16S_{j+1}}\right) \\ -d_{j+1}x' + \frac{m_j}{S_{j+1}}\left(\frac{22+d_j^2+d_{j+1}^2-4d_j+28d_{j+1}}{32}\right) + f_{j-1}(T_j) & x' \in \left[\frac{(d_{j+1}+14)m_j}{16S_{j+1}}, \frac{m_j}{S_{j+1}}\right], \end{cases} \tag{SM5.6}$$

for $j \geq K$.

SM5.2. Properties of the Objective Function.

Here, we verify, (SM5.5) satisfies Assumption 2.1 and 2.2. We begin by studying the properties of $f_j(x)$ for $j \geq K$. Note, we already know the properties of $f_j(x)$ for $j < K$ by Proposition 4.2.

Proposition SM5.1. *Let $j > K$. The continuous extension of $f_j : (T_j, T_{j+1}] \to \mathbb{R}$, (SM5.6), to $[T_j, T_{j+1}]$ is continuous on its domain; bounded from below by $f_{j-1}(T_j) - m_j/(8S_{j+1})$; differentiable on its domain with $\dot{f}_j(T_j) = -d_j$ and $\dot{f}_j(T_{j+1}) = -d_{j+1}$; its derivative is locally Lipschitz continuous; and $f_j(T_{j+1}) \geq f_{j-1}(T_j) + 7m_j/(16S_{j+1})$.*

Proof. The proof of this result is similar to that of Proposition 4.2. Hence, we only produce the values of $f_j(x)$ and $\dot{f}_j(x)$ at key points.

1. At $x = T_j$, $f_j(T_j) = f_{j-1}(T_j)$. $\dot{f}_j(T_j) = -d_j$.
2. At $x = (2 - d_j)m_j/(16S_{j+1}) + T_j$,

$$f_j(x) = \frac{m_j}{S_{j+1}}\left(\frac{d_j^2 - 2d_j}{16}\right) + f_{j-1}(T_j), \tag{SM5.7}$$

and $\dot{f}_j(x) = -d_j$.

3. At $x = T_j + 3m_j/(16S_{j+1})$,

$$f_j(x) = \frac{m_j}{S_{j+1}}\left(\frac{1 + d_j^2 - 4d_j}{32}\right) + f_{j-1}(T_j), \tag{SM5.8}$$

and $\dot{f}_j(x) = 1$.

4. At $x = T_j + m_j/(2S_{j+1})$,

$$f_j(x) = \frac{m_j}{S_{j+1}}\left(\frac{11 + d_j^2 - 4d_j}{32}\right) + f_{j-1}(T_j), \tag{SM5.9}$$

and $\dot{f}_j(x) = 0$.

$$f_j(x) \tag{SM5.10}$$

5. At $x = T_j + 13m_j/(16S_{j+1})$,

$$f_j(x) = \frac{m_j}{S_{j+1}}\left(\frac{21 + d_j^2 - 4d_j}{32}\right) + f_{j-1}(T_j), \tag{SM5.11}$$

and $\dot{f}_j(x) = 1$.

6. At $x = T_j + (d_{j+1} + 14)m_j/(16S_j)$,

$$f_j(x) = \frac{m_j}{S_{j+1}}\left(\frac{22 + d_j^2 - d_{j+1}^2 - 4d_j}{32}\right) + f_{j-1}(T_j), \tag{SM5.12}$$

and $\dot{f}_j(x) = -d_{j+1}$.

7. At $x = T_j + m_j/S_{j+1}$,

$$f_j(x) = \frac{m_j}{S_{j+1}}\left(\frac{22 + d_j^2 + d_{j+1}^2 - 4d_j - 4d_{j+1}}{32}\right) + f_{j-1}(T_j), \tag{SM5.13}$$

and $\dot{f}_j(x) = -d_{j+1}$.

Note, $f_j(T_{j+1}) = f_j(T_j + m_j/S_{j+1}) \geq (22 - 8)m_j/(32S_{j+1}) + f_{j-1}(T_j)$. ■

Proposition SM5.2. *The function $F : \mathbb{R} \to \mathbb{R}$ as defined in (SM5.5) is continuous and differentiable on its domain; it is lower bounded; its derivative is locally Lipschitz continuous; $F(T_j) \geq 7T_j/16$ for $j + 1 \in \mathbb{N}$; $\dot{F}(T_j) = -d_j$ for all $j + 1 \in \mathbb{N}$; and F's derivative is not globally Lipschitz continuous.*

Proof. As the proof of this statement is similar to the other constructions, we only verify the values of the objective and the derivative at $\{T_j\}$. For $j = 0$, $F(T_0) = 0$. For $j = 1, \ldots, K$, $F(T_j) = F(S_j) = f_{j-1}(S_j) = S_j/2 \geq 7S_j/16 = 7T_j/16$ by Proposition 4.2. For $j > K$, $F(T_j) = f_{j-1}(T_j) \geq f_{j-1}(T_{j-1}) + \frac{7m_j}{16S_{j+1}} = F(T_{j-1}) + \frac{7m_j}{16S_{j+1}}$ by Proposition SM5.1. By induction, for $j + 1 \in \mathbb{N}$, $F(T_j) \geq 7T_j/16$. Similarly, either by Proposition 4.2 or Proposition SM5.1, $\dot{F}(T_j) = \dot{f}_j(T_j) = -d_j$. ■

SM5.3. Properties of Gradient Descent on the Objective.

We now show that when gradient descent is applied to the constructed problem, the objective function diverges, and the gradient function converges to zero.

Proposition SM5.3. *Let $\{m_k : k+1 \in \mathbb{N}\}$ be any positive sequence such that $\sum_k m_k$ diverges and $m_k \to 0$. Define $F : \mathbb{R} \to \mathbb{R}$ as in (SM5.5). Suppose $x_0 = 0$ and let $\{x_k : k \in \mathbb{N}\}$ be generated according to (2.3) with $M_k = m_k I$ for all $k+1 \in \mathbb{N}$. Then, $\{M_k\}$ satisfies Properties 2.4 and 2.6. Moreover, (a) $\lim_k x_k = \infty$; (b) $\lim_k F(x_k) = \infty$; and (c) $\lim_k |\dot{F}(x_k)| = 0$.*

Proof. We show that $x_k = T_k$ for all $k + 1 \in \mathbb{N}$. For the base case, $x_0 = 0 = T_0$. Suppose $x_k = T_k$ for some $k < K$. Then,

$$x_{k+1} = x_k - M_k\dot{F}(x_k) = T_k + m_k d_k = T_k + m_k = T_{k+1}. \tag{SM5.14}$$

This implies that $x_K = T_K$. Now, suppose $x_k = T_k$ for some $k > K$. Then,

$$x_{k+1} = x_k - M_K \dot{F}(x_k) = T_k + m_k d_k = T_k + \frac{m_k}{S_{k+1}} = T_{k+1}. \tag{SM5.15}$$

Hence, $F(x_k) = F(T_k) \geq 7T_k/16$, which diverges to infinity. Moreover, for $k > K$, $|\dot{F}(x_k)| = |\dot{F}(T_k)| = d_k = 1/S_{k+1}$ which tends to zero. ■