Directional Smoothness and Gradient Methods: Convergence and Adaptivity

Aaron Mishkin[*]
Stanford University
amishkin@cs.stanford.edu

Ahmed Khaled[*]
Princeton University
ahmed.khaled@princeton.edu

Yuanhao Wang
Princeton University
yuanhaoa@princeton.edu

Aaron Defazio
FAIR, Meta AI
adefazio@meta.com

Robert M. Gower
CCM, Flatiron Institute
gowerrobert@gmail.com

Abstract

We develop new sub-optimality bounds for gradient descent (GD) that depend on the conditioning of the objective along the path of optimization rather than on global, worst-case constants. Key to our proofs is directional smoothness, a measure of gradient variation that we use to develop upper-bounds on the objective. Minimizing these upper-bounds requires solving implicit equations to obtain a sequence of strongly adapted step-sizes; we show that these equations are straightforward to solve for convex quadratics and lead to new guarantees for two classical step-sizes. For general functions, we prove that the Polyak step-size and normalized GD obtain fast, path-dependent rates despite using no knowledge of the directional smoothness. Experiments on logistic regression show our convergence guarantees are tighter than the classical theory based on L-smoothness.

1 Introduction

Gradient methods for differentiable functions are typically analyzed under the assumption that f is L-smooth, meaning ∇f is L-Lipschitz continuous. This condition implies f is upper-bounded by a quadratic and guarantees that gradient descent (GD) with step-size $\eta < 2/L$ decreases the optimality gap at each iteration (Bertsekas, 1997). However, experience shows that GD can still decrease the objective when f is not L-smooth, particularly for deep neural networks (Bengio, 2012; Z. Li et al., 2020; J. Cohen et al., 2021). Even for functions verifying smoothness, convergence rates are often pessimistic and fail to predict optimization speed in practice (Paquette et al., 2023).

One alternative to global smoothness is local Lipschitz continuity of the gradient ("local smoothness"). Local smoothness assumes different Lipschitz constants hold for different neighbourhoods, which avoids global assumptions and improves rates. However, such analyses typically rely on boundedness of the iterates and then use local smoothness to obtain L-smoothness over a compact set (Malitsky and Mishchenko, 2020). Boundedness is guaranteed in several ways: Junyu Zhang and Hong (2020) break optimization into stages, Patel and Berahas (2022) use stopping-times, and Lu and S. Mei (2023) employ a line-search. Unfortunately, these approaches modify the underlying optimization algorithm, require local smoothness oracles (Park et al., 2021), or rely on highly complex arguments.

In contrast, we prove simple rates for GD without global smoothness by deriving bounds of the form,

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{M(x_{k+1}, x_k)}{2} \|x_{k+1} - x_k\|_2^2, \tag{1}$$

[*]Equal contribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-0473

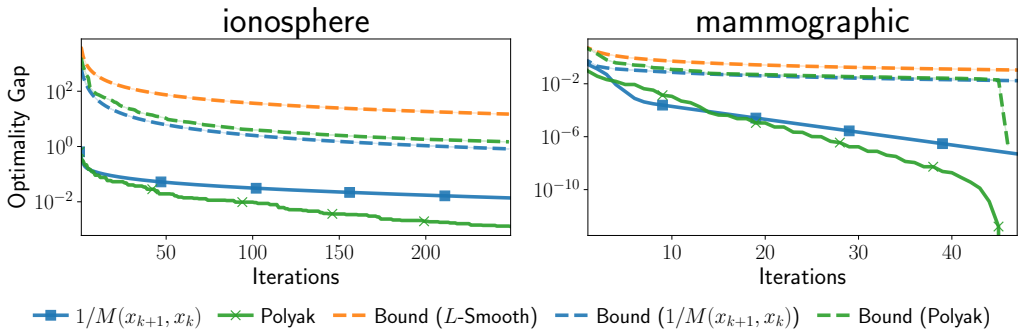

Figure 1: Comparison of actual (solid lines) and theoretical (dashed lines) convergence rates for GD with (i) step-sizes strongly adapted to the directional smoothness ($\eta_k = 1/M(x_{k+1}, x_k)$) and (ii) the Polyak step-size. Both problems are logistic regressions on UCI repository datasets (Asuncion and Newman, 2007). Our bounds using directional smoothness are tighter than those based on global L-smoothness of f and adapt to the optimization path. For example, on `mammographic` our theoretical rate for the Polyak step-size concentrates rapidly exactly when the optimizer shows fast convergence.

where the *directional smoothness function* $M(x_{k+1}, x_k)$ depends only on properties of f along the chord between x_k and x_{k+1}. Our sub-optimality bounds provide a path-dependent perspective on GD and are tighter than conventional analyses when the step-size sequence is adapted to the directional smoothness, meaning $\eta_k < 2/M(x_{k+1}, x_k)$. See Figure 1 for two real-data examples highlighting our improvement over classical rates. We summarize all our contributions as follows.

Directional Smoothness. We introduce three constructive directional smoothness functions $M(x, y)$. The first, point-wise smoothness, depends only on the end-points x, y and is easily computed, while the second, path-wise smoothness, yields a tighter bound, but depends on the chord $\mathcal{C} = \{\alpha x + (1-\alpha)y : \alpha \in [0, 1]\}$. The last function, which we call the optimal point-wise smoothness, is both easy-to-evaluate and provides the tightest possible quadratic upper bound.

Sub-optimality bounds. We leverage directional smoothness functions to prove new sub-optimality bounds for GD on convex functions. Our bounds are localized to the GD trajectory, hold for any step-size sequence, and are tighter than the classic analysis using L-smoothness. They are also more general since we do not need to assume that f is globally L-smooth to show progress; all we require is a sequence of step-sizes adapted to the directional smoothness function. Furthermore, our approach extends naturally to acceleration, allowing us to prove optimal rates for (strongly)-convex functions.

Adaptive Step-Sizes in the Quadratic Case. In the general setting, computing step-sizes which are adapted to the directional smoothness requires solving a challenging non-linear root-finding problem. For quadratic problems, we show that the ideal step-size that satisfies $\eta_k = 1/M(x_{k+1}, x_k)$ is the Rayleigh quotient and is connected to the hedging algorithm (Altschuler and Parrilo, 2023).

Exponential Search. Moving beyond quadratics, we prove that the equation $\eta_k = 1/M(x_{k+1}, x_k)$ admits a solution under mild conditions, meaning ideal step-sizes can be computed using Newton's method. Since computing these step-sizes is typically impractical, we adapt exponential search (Carmon and Hinder, 2022) to obtain similar path-dependent complexities up to a log-log penalty.

Polyak and Normalized GD. More importantly, we show that the Polyak step-size (Polyak, 1987) and normalized GD achieve fast, path-dependent rates *without* knowledge of the directional smoothness. Our analysis reveals that the Polyak step-size adapts to *any* directional smoothness to obtain the tightest possible convergence rate. This property is not shared by constant step-size GD and may explain the superiority of the Polyak step-size in many practical settings.

1.1 Additional Related Work

Directional smoothness is a relaxation of non-uniform smoothness (J. Mei et al., 2021), which restricts the smoothness function M to depend only on x, the origin point. J. Mei et al. (2021) leverage non-uniform smoothness and a non-uniform Łojasiewicz inequality to break lower-bounds for first-order optimization. Similarly, Berahas et al. (2023) show that a weak local smoothness oracle can break lower bounds for gradient methods. A major advantage of our work over such oracle-based approaches is that we construct explicit directional smoothness functions that are easy to evaluate.

Similar to non-uniform smoothness, Grimmer (2019) and Orabona (2023) consider Hölder-type growth conditions with constants that depend on a neighbourhood of x. Since directional smoothness is stronger than and implies these Hölder error bounds, our M functions can be leveraged to make their results fully explicit (the Hölder bounds are non-constructive). Finally, while they also analyze normalized GD, our rates are anytime and do not use online-to-batch reductions like Orabona (2023).

Directional smoothness is also related to (L_0, L_1)-smoothness (Jingzhao Zhang et al., 2020; B. Zhang et al., 2020), which can be interpreted as a directional smoothness function with exponential dependence on the distance between x and y. The extension of (L_0, L_1)-smoothness to (r, l)-smoothness by H. Li et al. (2023) shows how to bound sequences of such directional smoothness functions, even for accelerated methods. These approaches are complementary to ours and showcase a setting where directional smoothness leads to concrete convergence rates.

Our work is most closely connected to that by Malitsky and Mishchenko (2020), who use a smoothed version of $M(x, y)$ to set the step-size. Vladarean et al. (2021) apply a similar smoothed step-size scheme to primal-dual hybrid gradient methods, while Zhao and Huang (2024) relate directional smoothness to Barzilai-Borwein updates (Barzilai and Borwein, 1988) and Vainsencher et al. (2015) use local smoothness over neighbourhoods of the global minimizer to set the step-size for SVRG.

Finally, we note that adaptivity to directional smoothness is different from adaptivity to the sequence of observed gradients obtained by methods such as Adagrad (Duchi et al., 2010; Streeter and McMahan, 2010). Adagrad and its variants are most useful when the gradients are bounded, such as in Lipschitz optimization, although they can also be used to obtain rates for smooth functions (Levy, 2017). We do not address adaptivity to gradients in this work.

2 Directional Smoothness

We say that a convex function f is L-smooth if for all $x, y \in \mathbb{R}^d$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2. \tag{2}$$

Minimizing this quadratic upper bound in y gives the classical GD update with step-size $\eta_k = 1/L$. However, this viewpoint leads to rates which depend on the global, worst-case growth of f. This is both counter-intuitive and undesirable because the iterates of GD, $x_{k+1} = x_k - \eta_k \nabla f(x_k)$, depend only on local properties of f. Ideally, the analysis should also depend only on the local conditioning along the path $\{x_1, x_2, \ldots\}$. Towards this end, we generalize the smoothness upper-bound as follows.

Definition 2.1. We call $M : \mathbb{R}^{d,d} \to \mathbb{R}_+$ a *directional smoothness function* for f if for all $x, y \in \mathbb{R}^d$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M(x, y)}{2}\|y - x\|^2. \tag{3}$$

If a function is L-smooth, then $M(x, y) = L$ is a trivial choice of directional smoothness function. In the rest of this section, we construct different M functions that provide tighter bounds on f while still being possible to evaluate. The first is the *point-wise directional smoothness*,

$$D(x, y) := \frac{2\|\nabla f(y) - \nabla f(x)\|_2}{\|y - x\|_2}. \tag{4}$$

Point-wise smoothness is a directional estimate of L and satisfies $D(x, y) \leq 2L$. Indeed, L can be equivalently defined as the supremum of $D(x, y)/2$ over the domain of f (Beck, 2017). If f is convex and differentiable, then $D(x, y)$ is a directional smoothness function according to Definition 2.1.

Lemma 2.2. *If f is convex and differentiable, then the point-wise directional smoothness satisfies,*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{D(x, y)}{2}\|y - x\|_2^2. \tag{5}$$

See Appendix A (we defer all proofs to the relevant appendices). In the worst-case, the point-wise directional smoothness D is weaker than the standard upper-bound $M(x, y) = L$ by a factor of two. This is *not* an artifact of the analysis and is generally unavoidable, as the next proposition shows.

Proposition 2.3. *There exists a convex, differentiable f and $x, y \in \mathbb{R}^d$ such that if $t < 2$, then*

$$f(y) > f(x) + \langle \nabla f(x), y - x \rangle + \frac{t\|\nabla f(x) - \nabla f(y)\|}{2\|y - x\|_2}\|y - x\|_2^2. \tag{6}$$

https://doi.org/10.52202/079017-0473

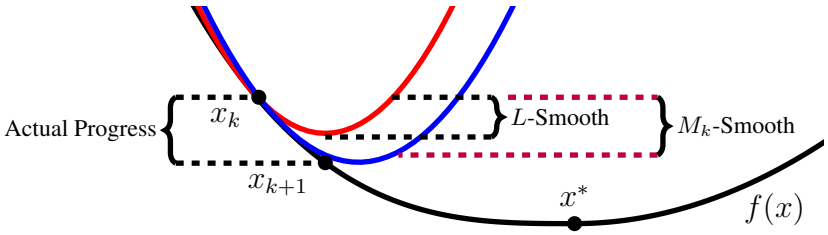

Figure 2: Illustration of GD with $\eta_k = 1/L$. Even though this step-size exactly minimizes the upper-bound from L-smoothness, M_k directional smoothness better predicts the progress of the gradient step because $M_k \ll L$. Our rates improve on L-smoothness because of this tighter bound.

While the point-wise smoothness is easy to compute, this additional factor of two can make Equation (5) looser than L-smoothness — on isotropic quadratics, for example. As an alternative, we define the *path-wise directional smoothness*,

$$A(x,y) := \sup_{t\in[0,1]} \frac{\langle \nabla f(x+t(y-x)) - \nabla f(x), y-x\rangle}{t\|y-x\|^2}, \tag{7}$$

and show it verifies the quadratic upper-bound and satisfies Definition 2.1 even without convexity.

Lemma 2.4. *For any differentiable function f, the path-wise smoothness* (7) *satisfies*

$$f(y) \le f(x) + \langle \nabla f(x), y-x\rangle + \frac{A(x,y)}{2}\|y-x\|_2^2. \tag{8}$$

Path smoothness is tighter than point-wise smoothness since $A(x,y) \le D(x,y)$, but hard to compute because it depends on the chord between x and y. That is, it depends on the properties of f on the line $\{tx + (1-t)y : t \in [0,1]\}$ rather than solely on the points x and y like the point-wise smoothness.

Point-wise and path-wise smoothness are constructive, but they may not yield the tightest bounds in all situations. The tightest directional smoothness function, which we call the *optimal point-wise smoothness*, is the smallest number for which the quadratic upper bound holds,

$$H(x,y) = \frac{|f(y) - f(x) - \langle \nabla f(x), y-x\rangle|}{\frac{1}{2}\|y-x\|^2} \tag{9}$$

By definition, H is the tightest possible directional smoothness function; it lower bounds any constant C that satisfies the quadratic bound (2). Thus, $H(x,y) \le M(x,y)$ for any smoothness function M.

The directional smoothness functions introduced in this section represent different trade-offs between computability and tightness. The optimal point-wise smoothness $H(x,y)$ requires access to both the function and gradient values, whereas the point-wise directional-smoothness $D(x,y)$ requires only access to the gradients and convexity. In contrast, the path-wise direction smoothness $A(x,y)$ satisfies Lemma 2.4 with or without convexity, but may be hard to evaluate.

3 Path-Dependent Sub-Optimality Bounds

Using directional smoothness, we obtain a descent lemma which depends only on local geometry,

$$f(x_{k+1}) \le f(x_k) - \left(\eta_k - \frac{\eta_k^2 M(x_k, x_{k+1})}{2}\right)\|\nabla f(x_k)\|_2^2. \tag{10}$$

See Lemma A.1. If $\eta_k < 2/M(x_k, x_{k+1})$, then GD is guaranteed to decrease the function value and we call η_k *adapted* to $M(x_k, x_{k+1})$. However, computing adapted step-sizes is not always straightforward. For instance, finding $\eta_k = 1/M(x_k, x_{k+1}(\eta_k))$ requires solving a non-linear equation.

The rest of this section leverages directional smoothness to derive new guarantees for GD with arbitrary step-sizes. We emphasize that these results are *sub-optimality bounds*, rather than convergence rates; a sequence of adapted step-sizes is required to convert our propositions into a convergence theory. As a trade-off, our bounds reflect the locality of GD, rather than treating it as a global method.

We start with the case when f has lower curvature. Instead of using strong convexity or the PL-condition (Karimi et al., 2016), we propose the directional strong convexity constant,

$$\mu(x,y) = \inf_{t\in[0,1]} \frac{\langle \nabla f(x+t(y-x)) - \nabla f(x), y-x\rangle}{t\|y-x\|_2^2}. \tag{11}$$

If f is convex, then $\mu(x,y) \geq 0$ and it verifies the standard lower-bound from strong convexity,

$$f(y) \geq f(x) + \langle \nabla f(x), y-x\rangle + \frac{\mu(x,y)}{2}\|y-x\|_2^2. \tag{12}$$

Moreover, we have $\mu(x,y) \geq \mu$ when f is μ–strongly convex. We prove two bounds for convex functions using directional strong convexity. For brevity, we denote $M_i := M(x_i, x_{i+1})$, $\mu_i := \mu_i(x_i, x^*)$, $\delta_i = f(x_i) - f(x^*)$, and $\Delta_i = \|x_i - x^*\|_2^2$, where x^* is a minimizer of f.

Proposition 3.1. *If f is convex and differentiable, then GD with step-size sequence $\{\eta_k\}$ satisfies,*

$$\delta_k \leq \left[\prod_{i\in\mathcal{G}} (1+\eta_i\lambda_i\mu_i)\right]\delta_0 + \sum_{i\in\mathcal{B}}\left[\prod_{j>i, j\in\mathcal{G}} (1+\eta_j\lambda_j\mu_j)\right]\frac{\eta_i\lambda_i}{2}\|\nabla f(x_i)\|_2^2, \tag{13}$$

where $\lambda_i = \eta_i M_i - 2$, $\mathcal{G} = \{i : \eta_i < 2/M_i\}$, and $\mathcal{B} = [k]\setminus\mathcal{G}$.

The analysis splits iterations into good steps $\mathcal{G}$, where η_k is adapted to the directional smoothness, and bad steps $\mathcal{B}$, where the step-size is too large and GD may increase the optimality gap. When f is L-smooth and μ-strongly convex, using the step-size sequence $\eta_k = 1/L$ gives

$$f(x_{k+1}) - f(x^*) \leq \left[\prod_{i=0}^{k}\left(1 - \frac{\mu_i\,(2 - M_i/L)}{L}\right)\right](f(x_0) - f(x^*)) \tag{14}$$

where $\mu_i\,(2 - M_i/L) \geq \mu$. Thus, Equation (13) gives at least as tight a rate as standard assumptions by localizing to the convergence path using *any* directional smoothness M. When $M_i < L$, the gap in constants yields a strictly improved rate (see Figure 2). We also prove a more elegant bound.

Proposition 3.2. *If f is convex and differentiable, then GD with step-size sequence $\{\eta_k\}$ satisfies,*

$$\Delta_k \leq \left[\prod_{i=0}^{k}\frac{|1-\mu_i\eta_i|}{1+\mu_{i+1}\eta_i}\right]\Delta_0 + \sum_{i=0}^{k}\left[\prod_{j>i}\frac{|1-\mu_j\eta_j|}{1+\mu_{j+1}\eta_j}\right]\frac{\left(M_i\eta_i^3 - \eta_i^2\right)}{1+\mu_{i+1}\eta_i}\|\nabla f(x_i)\|_2^2. \tag{15}$$

Unlike Proposition 3.1, this analysis shows linear progress at each iteration and does not divide k into good steps and bad steps. In exchange, the second term in Equation (15) reflects how much convergence is degraded when η_k is not adapted to the directional smoothness function M. We conclude this section with a bound for when there is no lower curvature, meaning $\mu_i = 0$.

Proposition 3.3. *Let $\overline{x}_k = \sum_{i=0}^k \eta_i x_{i+1} / \sum_{i=0}^k \eta_i$. If f is convex and differentiable, then GD satisfies,*

$$f(\overline{x}_k) - f(x^*) \leq \frac{\|x_0 - x^*\|_2^2}{2\sum_{i=0}^k \eta_i} + \frac{\sum_{i=0}^k \eta_i^2(\eta_i M_i - 1)\|\nabla f(x_i)\|_2^2}{2\sum_{i=0}^k \eta_i}. \tag{16}$$

Eq. (16) is faster than standard analyses whenever $M_i < L$; it will be a key tool in the next sections.

3.1 Path-Dependent Acceleration

Now we show that directional smoothness can also be used to derive path-dependent sub-optimality bounds for accelerated algorithms — that is, methods obtaining optimal rates for smooth, convex optimization. In particular, we study Nesterov's accelerated gradient descent (AGD) (Nesterov, 1983) and prove that directional smoothness leads to tighter rates given adapted step-sizes. Throughout this section we assume that f is μ-strongly convex with $\mu = 0$ when f is merely convex.

Although our analysis uses estimating sequences (Nesterov et al., 2018), we state AGD in the following "momentum" formulation, where y_k is the momentum and α_k the momentum parameter,

$$\begin{aligned} x_{k+1} &= y_k - \eta_k \nabla f(y_k) \\ \alpha_{k+1}^2 &= (1-\alpha_{k+1})\alpha_k^2\frac{\eta_{k+1}}{\eta_k} + \eta_{k+1}\alpha_{k+1}\mu \\ y_{k+1} &= x_{k+1} + \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2 + \alpha_{k+1}}\left(x_{k+1} - x_k\right). \end{aligned} \tag{17}$$

 https://doi.org/10.52202/079017-0473

If $\eta_k \leq 1/M(x_k, x_{k+1})$, then Equation (10) combined with $1 - \eta_k M(x_k, x_{k+1})/2 \geq 1/2$ implies,

$$f(x_{k+1}) \leq f(y_k) - \frac{\eta_k}{2}\|\nabla f(y_k)\|_2^2. \tag{18}$$

Our analysis leverages the fact that this descent condition for x_{k+1} is the only connection between the smoothness of f and the convergence rate of AGD. Since Equation (18) depends only on the step-size η_k, we can replace L within the analysis of AGD with a sequence of adapted step-sizes. The following theorem controls the effect of these step-sizes to obtain path-dependent bounds.

Theorem 3.4. *Suppose f is differentiable, μ–strongly convex and AGD is run with adapted step-sizes $\eta_k \leq 1/M_k$. If $\mu > 0$ and $\alpha_0 = \sqrt{\eta_0 \mu}$, then AGD obtains the following accelerated rate:*

$$f(x_{k+1}) - f(x^*) \leq \prod_{i=0}^{k} (1 - \sqrt{\mu \eta_i}) \left[f(x_0) - f(x^*) + \frac{\mu}{2}\|x_0 - x^*\|_2^2 \right]. \tag{19}$$

Let $\eta_{min} = \min_{i \in [k]} \eta_i$. If $\mu \geq 0$ and $\alpha_0 \in (\sqrt{\mu \eta_0}, c)$, where c is the maximum value of α_0 for which $\gamma_0 = \frac{\alpha_0^2 - \eta_0 \alpha_0 \mu}{\eta_0 (1 - \alpha_0)}$ satisfies $\gamma_0 < 3/\eta_{min} + \mu$, then AGD obtains the following rate:

$$f(x_{k+1}) - f(x^*) \leq \frac{4}{\eta_{min}(\gamma_0 - \mu)(k+1)^2} \left[f(x_0) - f(x^*) + \frac{\gamma_0}{2}\|x_0 - x^*\|_2^2 \right]. \tag{20}$$

If $\eta_k = 1/M_k > 1/L$, then these rates are strictly faster than those obtained under L-smoothness and Theorem 3.4 shows that AGD provably benefits from taking the largest possible steps given the local geometry of f. However, obtaining accelerated rates when $\mu = 0$ requires prior knowledge of the minimum step-size; while this is straightforward for L-smooth functions, it is not clear how to extend such result to non-strongly convex acceleration with locally Lipschitz gradients. For example, while H. Li et al. (2023) show that the (r, l)-smoothness (a valid directional smoothness function) is bounded over the iterate trajectory, their rate does not adapt to the optimization path.

4 Adaptive Learning Rates

Converting our sub-optimality bounds into convergence rates requires adapted step-sizes satisfying $\eta_k < 2/M(x_k, x_{k+1})$. Given an adapted step-size, the directional descent lemma (Equation (10)) implies GD decreases f and we can obtain fast rates if the step-sizes are bounded below. However, x_{k+1} is itself a function of η_k, meaning adapted step-sizes are not straightforward to compute.

For L-smooth f, the different directional smoothness functions M introduced in Section 2 satisfy $M(x_k, x_{k+1}) \leq 2L$. This implies $\eta_k < \frac{1}{L}$ is trivially adapted. As such step-sizes don't capture local properties of f, we introduce the notion of *strongly adapted step-sizes*, which satisfy

$$\eta_k = 1/M(x_{k+1}(\eta_k), x_k). \tag{21}$$

Equation (10) implies GD with a strongly adapted step-size makes guaranteed progress as,

$$f(x_{k+1}) \leq f(x_k) - [2M(x_{k+1}, x_k)]^{-1} \|\nabla f(x_k)\|_2^2. \tag{22}$$

This progress is greater than that guaranteed by L-smoothness when $M(x_k, x_{k+1}) < L$ and holds even when f is not L-smooth. However, it is not clear a priori if (i) strongly adapted step-sizes exist or if (ii) any iterative method achieves the progress in Eq. (21). Surprisingly, we provide a positive answer to both questions. Strongly adapted η_k are computable and we also prove GD with the Polyak step-size adapts to any choice of directional smoothness, including the optimal point-wise smoothness. Before presenting this strong result, we consider the illustrative case of quadratic minimization.

4.1 Adaptivity in Quadratics

Now we show that step-sizes adapted to both the point-wise smoothness M and the path-wise smoothness A exist when f is quadratic. Let $f(x) = x^\top B x/2 - c^\top x$, where B is positive semi-definite. Assuming $\{\eta_k\}$ is strongly adapted to the directional smoothness, Equation (16) implies

$$f(\overline{x}_k) - f(x^*) \leq \frac{\|x_0 - x^*\|_2^2}{2\sum_{i=0}^k \eta_i} = \frac{\|x_0 - x^*\|_2^2}{2\sum_{i=0}^k \frac{1}{M(x_i, x_{i+1})}} \leq \frac{\|x_0 - x^*\|_2^2}{2(k+1)} \frac{\sum_{i=0}^k M(x_i, x_{i+1})}{k+1}, \tag{23}$$

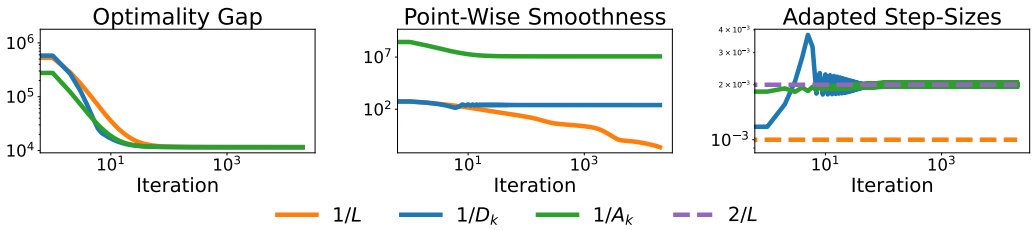

Figure 3: Performance of GD with different step-size rules for a synthetic quadratic problem. We run GD for 20,000 steps on 20 random quadratic problems with $L = 1000$ and Hessian skew. Left-to-right, the first plot shows the optimality gap $f(x_k) - f(x^*)$, the second shows the point-wise directional smoothness $D(x_k, x_{k+1})$, and the third shows step-sizes used by the different methods.

where we used $\eta_i M_i = 1$ as well as Jensen's inequality. This guarantee depends solely on the average directional smoothness along the optimization trajectory $\{x_0, x_1, \ldots\}$. When f is quadratic, we can exactly compute these smoothness constants. In particular, the point-wise directional smoothness is,

$$D(x_i, x_{i+1}) = 2\|B\nabla f(x_i)\|_2 / \|\nabla f(x_i)\|_2.$$

Notably, $D(x_i, x_{i+1})$ has no dependence on x_{i+1} and the corresponding strongly adapted step-size is given by $\eta_i = \|\nabla f(x_i)\|_2 / (2\|B\nabla f(x_i)\|_2)$ — see Lemma C.1. Remarkably, this expression recovers the step-size proposed by Dai and Yang (2006), who show it approximates the Cauchy step-size and converges to the "edge-of-stability" (J. Cohen et al., 2021) at $2/L$ as $k \to \infty$. Combining this simple expression with Equation (23) gives a fast, non-asymptotic convergence rate for GD and new theoretical justification for their work.

We can also compute the path-wise directional smoothness in closed form. As Lemma C.2 shows,

$$A(x_i, x_{i+1}) = \nabla f(x_i)^\top B \nabla f(x_i) / \nabla f(x)^\top \nabla f(x),$$

and $\eta_i = \nabla f(x_i)^\top \nabla f(x_i) / [\nabla f(x_i)^\top B \nabla f(x_i)]$ is the well-known Cauchy step-size. Path-wise directional smoothness thus provides another interpretation (and convergence guarantee) for the Cauchy step-size, which is traditionally derived by minimizing $f(x - \eta \nabla f(x))$ in η.

4.2 Adaptivity for Convex Functions

In the last subsection, we proved that strongly adapted step-sizes for the point-wise and path-wise directional smoothness functions have closed-form expressions when f is quadratic. Moreover, these step-sizes recover two classical schemes from the optimization literature, giving them new justification and fast convergence rates. Now we consider the existence of strongly adapted step-sizes for general convex functions. Our first result gives simple conditions for Equation (21) to have at least one solution when M is the point-wise directional smoothness.

Proposition 4.1. *If f is convex and continuously differentiable, then either (i) f is minimized along the ray $x(\eta) = x - \eta \nabla f(x)$ or (ii) there exists $\eta > 0$ satisfying $\eta = 1/D(x, x - \eta \nabla f(x))$.*

The next proposition uses a similar argument with slightly stronger conditions to show existence of strongly adapted step-sizes for the path-wise smoothness.

Proposition 4.2. *If f is convex and twice continuously differentiable, then either (i) f is minimized along the ray $x(\eta) = x - \eta \nabla f(x)$ or (ii) there exists $\eta > 0$ satisfying $\eta = 1/A(x, x - \eta \nabla f(x))$.*

Propositions 4.1 and 4.2 do not assume the global smoothness; although neither proof is constructive, it is possible to compute strongly adapted step-sizes for the point-wise directional smoothness using root-finding methods. We show in Section 5 that if f is twice differentiable, then strongly adapted step-sizes can be found via Newton's method using only Hessian-vector products, $\nabla^2 f(x) \nabla f(x)$.

4.2.1 Exponential Search

Now we show that the exponential search algorithm developed by Carmon and Hinder (2022) can be used to find step-sizes that adapt *on average* to the directional smoothness. Consider a fixed

https://doi.org/10.52202/079017-0473

optimization horizon k and denote by $x_i(\eta)$ the sequence of iterates obtained by running GD from x_0 using a fixed step-size η. Define the criterion function,

$$\psi(\eta) = \frac{\sum_{i=0}^{k} \|\nabla f(x_i(\eta))\|_2^2}{\sum_{i=0}^{k} M(x_i(\eta), x_{i+1}(\eta))\|\nabla f(x_i(\eta))\|_2^2}, \tag{24}$$

and suppose that we have a step-size η that satisfies $\psi(\eta)/2 \leq \eta \leq \psi(\eta)$. Using these bounds in Proposition 3.3 yields the following convergence rate,

$$f(\overline{x}_k) - f_* \leq \left[\frac{k\sum_{i=0}^{k} M(x_i, x_{i+1})\|\nabla f(x_i)\|_2^2}{\sum_{i=0}^{k}\|\nabla f(x_i)\|_2^2}\right] \|x_0 - x^*\|_2^2. \tag{25}$$

While η does not adapt to each directional smoothness $M(x_i, x_{i+1})$ along the path, it adapts to a *weighted average* of the directional smoothness constants, where the weights are the observed squared gradient norms. This is always smaller than the maximum directional smoothness along the trajectory and can be much smaller than the global smoothness. Furthermore, we have reduced our problem to finding $\eta \in [\psi(\eta)/2, \psi(\eta)]$, which is similar to the problem Carmon and Hinder (2022) solve with exponential search. We adopt their approach as Algorithm 1 and give a convergence guarantee.

Theorem 4.3. *Assume f is convex and L-smooth. Then Algorithm 1 with $\eta_0 > 0$ requires at most $2K(\log\log(2\eta_0/L) \vee 1)$ iterations of GD and in the last run it outputs a step-size η and point $\overline{x}_K = \frac{1}{K}\sum_{i=0}^{K-1} x_i(\eta)$ such that exactly one of the following holds:*

$$\textit{Case 1:} \quad \eta = \eta_0 \quad \textit{and} \quad f(\overline{x}_K) - f(x^*) \leq \frac{\|x_0 - x^*\|_2^2}{2K\eta_0}$$

$$\textit{Case 2:} \quad \eta \neq \eta_0 \quad \textit{and} \quad f(\overline{x}_K) - f^* \leq \frac{\|x_0 - x^*\|_2^2}{2K}\left[\frac{\sum_{i=0}^{k} M_i\|\nabla f(x_i')\|_2^2}{\sum_{i=0}^{k}\|\nabla f(x_i')\|_2^2}\right],$$

where $M_i \overset{def}{=} M(x_i', x_{i+1}')$ and x_i' are the iterates generated by GD with step-size $\eta' \in [\eta, 2\eta]$.

Theorem 4.3 requires f to be L-smooth, but has only a $\log\log$ dependence on the global smoothness constant. Moreover, the rate scales with the weighted average of smoothness constants along a very close trajectory $\{x_1', x_2', \ldots\}$. In the next section, we give convergence bounds that depend on the unweighted average of the directional smoothness constants along the *actual* optimization trajectory.

4.2.2 Polyak's Step-Size Rule

Our theory so-far suggests using strongly adapted step-sizes, but neither root-finding nor exponential search are practical methods for large-scale optimization. Thus, we now consider other step-size selection rules which may leverage directional smoothness. In particular, the Polyak step-size sets,

$$\eta_k = \gamma\left(f(x_k) - f(x^*)\right)/\|\nabla f(x_k)\|_2^2, \tag{26}$$

for some $\gamma > 0$, which is optimal for smooth and non-smooth optimization (Hazan and Kakade, 2019) given knowledge of $f(w^*)$. Surprisingly, we show that GD with the Polyak step-size also achieves the same guarantee as strongly adapted step-sizes without knowledge of the directional smoothness.

Theorem 4.4. *Suppose that f is convex and differentiable and let M be any directional smoothness function for f. Let $\Delta_0 := \|x_0 - x^*\|_2^2$. Then GD with the Polyak step-size and $\gamma \in (1, 2)$ satisfies*

$$f(\overline{x}_k) - f(x^*) \leq \frac{c(\gamma)\Delta_0}{2\sum_{i=0}^{k-1} M(x_i, x_{i+1})^{-1}}, \tag{27}$$

where $c(\gamma) = \gamma/(2-\gamma)(\gamma-1)$ and $\overline{x}_k = \sum_{i=0}^{k-1}\left[M(x_i, x_{i+1})^{-1}x_i\right] / \left(\sum_{i=0}^{k-1} M(x_i, x_{i+1})^{-1}\right)$.

Theorem 4.4 measures sub-optimality at an average iterate obtained using the directional smoothness. However, it also holds for the best iterate, $\hat{x}_k = \arg\min_{i\in[k]} f(x_i)$, meaning no knowledge of the directional smoothness is required to obtain the guarantee. We prove an alternative guarantee for the Polyak step-size in Theorem D.2, where the progress depends on the sum of step-sizes rather than on the average directional smoothness. This shows that the step-size in Equation (26) can itself be viewed as a measure of local smoothness, albeit without formal justification.

Compared with the standard guarantee for the Polyak step-size under L-smoothness, $f(\overline{x}_k) - f(x^*) \leq 2L\Delta_0/k$ (Hazan and Kakade, 2019), our analysis in Theorem 4.4 with the choice $\gamma = 1.5$ yields

$$f(\overline{x}_k) - f(x^*) \leq \frac{3\Delta_0}{\sum_{i=0}^{k-1} M(x_i, x_{i+1})^{-1}} \leq \frac{3\Delta_0}{k} \frac{\sum_{k=0}^{k-1} M(x_i, x_{i+1})}{k},$$

where the second bound follows from Jensen's inequality and shows that the convergence depends on the average directional smoothness along the trajectory, rather than on L. If f is L-smooth, then $M(x_k, x_{k+1}) \leq L$ immediately recovers the classic rate for Polyak's method up to a $3/2$ constant factor. If f is not L-smooth, but $M(x_k, x_{k+1})$ is bounded, then Equation (27) generalizes the $O(1/k)$ rate proved concurrently by Takezawa et al. (2024), but for any choice of directional smoothness (of which (L_0, L_1)-smoothness (Jingzhao Zhang et al., 2020) is but one).

Comparison with strongly adapted step-sizes. As we saw for quadratics, strongly adapted step-sizes for any directional smoothness function allow us to obtain the following convergence rate,

$$f(\overline{x}_k) - f(x^*) \leq \frac{\|x_0 - x^*\|_2^2}{2\sum_{i=0}^{k-1} M(x_i, x_{i+1})^{-1}}.$$

This is matches the guarantee given by Equation (27) up to constant factors. As a result, we give a positive answer to the question posed earlier in this section: GD with the Polyak step-size achieves the same convergence for any smoothness function M as GD with step-sizes strongly adapted to M.

Application to the optimal directional smoothness. Theorem 4.4 holds for *every* directional smoothness function M. Therefore we can specialize Equation (27) with the optimal point-wise directional smoothness H (as defined in Equation (4)) and $\gamma = 1.5$ to get the guarantee,

$$\min_{i \in [k-1]} [f(x_i) - f(x^*)] \leq \frac{3\|x_0 - x^*\|_2^2}{\sum_{i=0}^{k-1} H(x_i, x_{i+1})^{-1}}. \tag{28}$$

This rate requires computing the iterate with the minimum function value, but that is easy to track during optimization. Unlike our previous results, Equation (28) requires no access to the optimal point-wise smoothness, yet obtains a dependence on the tightest constant possible.

4.3 Normalized Gradient Descent

Now we change directions slightly and study normalized GD, whose convergence also depends on the directional smoothness. Normalized GD uses step-sizes which are divided by the gradient magnitude,

$$x_{k+1} = x_k - \frac{\eta_k}{\|\nabla f(x_k)\|_2} \nabla f(x_k). \tag{29}$$

Our next theorem shows that normalized GD obtains a guarantee which depends solely on the average of the point-wise directional smoothness $D_k := D(x_k, x_{k+1})$ despite no explicit knowledge of D_k.

Theorem 4.5. *Suppose that f is convex and differentiable. Let D be the point-wise directional smoothness defined by Equation* (4) *and $\Delta_0 := \|x_0 - x^*\|_2^2$. Then normalized GD with a sequence of non-increasing step-sizes η_k satisfies*

$$f(\hat{x}_k) - f(x^*) \leq \frac{\Delta_0 + \sum_{i=0}^{k-1} \eta_i^2}{2k^2} \left(\frac{f(x_0)}{\eta_0^2} - \frac{f(x^*)}{\eta_{k-1}^2} \right) + \frac{\Delta_0 + \sum_{i=0}^{k-1} \eta_i^2}{2k} \sum_{i=0}^{k-1} \frac{M(x_i, x_{i+1})}{k}, \tag{30}$$

where $\hat{x}_k = \arg\min_{i \in [k-1]} f(x_i)$. If $\max_{i \in [k-1]} M(x_i, x_{i+1})$ is bounded for all k (i.e. f is L-smooth), then for $\eta_i = 1/\sqrt{i}$ we have $f(\hat{x}_k) - f(x^*) \in \mathcal{O}(1/k)$ and for $\eta_i = 1/\sqrt{i}$ we get the anytime result $f(\hat{x}_k) - f(x^*) \in \mathcal{O}(\log(k)/k)$.

Theorem 4.5 gives a rate for normalized GD which is valid for any convex f without any dependence on global smoothness. However, does not adapt to any smoothness function like the Polyak step-size.

5 Experiments

We evaluate the practical improvement of our convergence rates over those using L-smoothness on two logistic regression problems taken from the UCI repository (Asuncion and Newman, 2007).

https://doi.org/10.52202/079017-0473

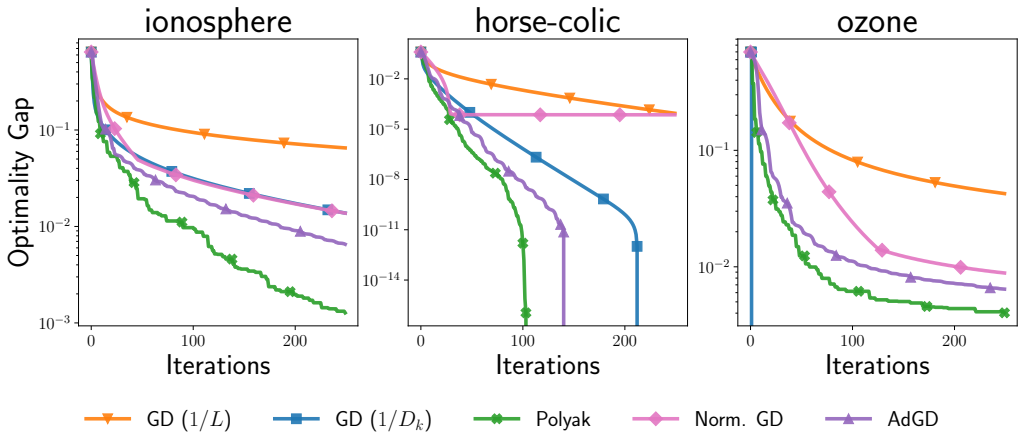

Figure 4: Comparison of GD with $\eta_k = 1/L$, step-sizes strongly adapted to the point-wise smoothness ($\eta_k = 1/D(x_k, x_{k+1})$), and the Polyak step-size against normalized GD (Norm. GD) and the AdGD method on three logistic regression problems. AdGD uses a smoothed version of the point-wise directional smoothness from the previous iteration to set η_k. We find that GD methods with adaptive step-sizes consistently outperform GD with $\eta_k = 1/L$ and even obtain a linear rate on `horse-colic`.

Figure 1 compares GD with strongly adapted step-sizes $\eta = 1/M_k$, where M_k is the point-wise smoothness, against GD with the Polyak step-size. We also plot the exact convergence rates for each method, Equation (16) and Equation (27), respectively, and compare against the classical guarantee for both methods. Our convergence rates are an order of magnitude tighter on the `ionosphere` dataset and display a remarkable ability to adapt to the path of optimization on `mammographic`.

Figure 3 compares the performance of GD with strongly adapted step-sizes and with the fixed step-size $\eta_k = 1/L$ for a synthetic quadratic with Hessian skew (R. Pan et al., 2022). Results are averaged over twenty random problems. We find that strongly adapted step-sizes lead to significantly faster convergence. Since $A_k, D_k \ll L$, the adapted step-sizes are larger than $2/L$, especially at the start of training; they eventually converge to $2/L$, indicating these methods operate at the edge-of-stability (J. Cohen et al., 2021; J. M. Cohen et al., 2022). This is consistent with Ahn et al. (2022) and Y. Pan and Y. Li (2023), who show local smoothness is correlated with edge-of-stability behavior.

We conclude with a comparison of empirical convergence rates on three additional logistic regression problems from the UCI repository. We compare GD with $\eta_k = 1/L$, GD with step-sizes strongly adapted to the point-wise smoothness ($\eta_k = 1/D_k$), GD with the Polyak step-size (Polyak), and normalized GD (Norm. GD) against the AdGD method (Malitsky and Mishchenko, 2020). The Polyak step-size performs best on every dataset but `ozone`, where GD with $\eta_k = 1/D_k$ solves the problem to high accuracy in just a few iterations. Thus, although Polyak step-sizes have the optimal dependence on directional smoothness, computing strongly adapted step-sizes can still be advantageous.

6 Conclusion

We present new sub-optimality bounds for GD under novel measures of local gradient variation which we call directional smoothness functions. Our results hold for any step-sizes, improve over standard analyses when η_k is adapted to the choice of directional smoothness, and depend only on properties of f local to the optimization path. For convex quadratics, we show that computing step-sizes strongly adapted to directional smoothness functions is straightforward and recovers two well-known step-size schemes, including the Cauchy step-size. In the general case, we prove that an algorithm based on exponential search gives a weighted-version of the path-dependent convergence rate with no need for adapted step-sizes. We also show that GD with the Polyak step-size and normalized GD both obtain fast rates with no dependence on the global smoothness parameter. Crucially, the Polyak step-size adapts to any choice of directional smoothness, including the tightest possible parameter.

Acknowledgements

Aaron Mishkin was supported by NSF Grant DGE-1656518, by NSERC Grant PGSD3-547242-2020, and by an internship at the Center for Computational Mathematics, Flatiron Institute. We thank Si Yi Meng for insightful discussions during the preparation of this work and Fabian Schaipp for use of the step-back code. We also thank the anonymous reviewers for comments leading to improvements in Proposition 3.2 and the addition of Theorem D.2.

References

Ahn, Kwangjun, Jingzhao Zhang, and Suvrit Sra (2022). "Understanding the unstable convergence of gradient descent". In: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Vol. 162. Proceedings of Machine Learning Research, pp. 247–257.

Altschuler, Jason M. and Pablo A. Parrilo (2023). "Acceleration by Stepsize Hedging I: Multi-Step Descent and the Silver Stepsize Schedule". In: *CoRR* abs/2309.07879.

Asuncion, Arthur and David Newman (2007). *UCI machine learning repository*.

Barzilai, Jonathan and Jonathan M Borwein (1988). "Two-point step size gradient methods". In: *IMA journal of numerical analysis* 8.1, pp. 141–148.

Beck, Amir (2017). *First-order methods in optimization*. MOS-SIAM series on optimization. Philadelphia : Philadelphia: Society for Industrial and Applied Mathematics ; Mathematical Optimization Society. ISBN: 978-161-197-4-9-9-7.

Bengio, Yoshua (2012). "Practical Recommendations for Gradient-Based Training of Deep Architectures". In: *Neural Networks: Tricks of the Trade - Second Edition*. Vol. 7700. Lecture Notes in Computer Science, pp. 437–478.

Berahas, Albert S., Lindon Roberts, and Fred Roosta (2023). "Non-Uniform Smoothness for Gradient Descent". In: *arXiv preprint arXiv:2311.08615* abs/2311.08615.

Bertsekas, Dimitri P (1997). "Nonlinear programming". In: *Journal of the Operational Research Society* 48.3, pp. 334–334.

Bubeck, Sébastien et al. (2015). "Convex optimization: Algorithms and complexity". In: *Foundations and Trends® in Machine Learning* 8.3-4, pp. 231–357.

Carmon, Yair and Oliver Hinder (2022). "Making SGD Parameter-Free". In: *Conference on Learning Theory, 2-5 July 2022, London, UK*. Vol. 178. Proceedings of Machine Learning Research, pp. 2360–2389.

Cohen, Jeremy et al. (2021). "Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability". In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Cohen, Jeremy M. et al. (2022). "Adaptive Gradient Methods At the Edge of Stability". In: *arXiv preprint arXiv:2207.14484* abs/2207.14484.

Dai, Y. H. and X. Q. Yang (2006). "A New Gradient Method with an Optimal Stepsize Property". In: *Computational Optimization and Applications* 33.1, pp. 73–88.

Duchi, John C., Elad Hazan, and Yoram Singer (2010). "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization". In: *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pp. 257–269.

Fernández-Delgado, Manuel et al. (2014). "Do we need hundreds of classifiers to solve real world classification problems?" In: *The journal of machine learning research* 15.1, pp. 3133–3181.

Grimmer, Benjamin (2019). "Convergence Rates for Deterministic and Stochastic Subgradient Methods without Lipschitz Continuity". In: *SIAM J. Optim.* 29.2, pp. 1350–1365.

Hazan, Elad and Sham Kakade (2019). "Revisiting the Polyak step size". In: *arXiv preprint arXiv:1905.00313*.

He, Kaiming et al. (2015). "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034.

Hogan, William W (1973). "Point-to-set maps in mathematical programming". In: *SIAM review* 15.3, pp. 591–603.

Karimi, Hamed, Julie Nutini, and Mark Schmidt (2016). "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*. Springer, pp. 795–811.

 https://doi.org/10.52202/079017-0473

Levy, Kfir Y. (2017). "Online to Offline Conversions, Universality and Adaptive Minibatch Sizes". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 1613–1622.

Li, Haochuan et al. (2023). "Convex and Non-convex Optimization Under Generalized Smoothness". In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Li, Zhiyuan, Kaifeng Lyu, and Sanjeev Arora (2020). "Reconciling Modern Deep Learning with Traditional Optimization Analyses: The Intrinsic Learning Rate". In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Liu, Dong C and Jorge Nocedal (1989). "On the limited memory BFGS method for large scale optimization". In: *Mathematical programming* 45.1-3, pp. 503–528.

Lu, Zhaosong and Sanyou Mei (2023). "Accelerated first-order methods for convex optimization with locally Lipschitz continuous gradient". In: *SIAM Journal on Optimization* 33.3, pp. 2275–2310.

Malitsky, Yura and Konstantin Mishchenko (2020). "Adaptive Gradient Descent without Descent". In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*. Vol. 119. Proceedings of Machine Learning Research, pp. 6702–6712.

Mei, Jincheng et al. (2021). "Leveraging Non-uniformity in First-order Non-convex Optimization". In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. Vol. 139. Proceedings of Machine Learning Research, pp. 7555–7564.

Mishkin, Aaron, Mert Pilanci, and Mark Schmidt (2024). "Faster Convergence of Stochastic Accelerated Gradient Descent under Interpolation". In: *arXiv preprint arXiv:2404.02378*.

Nesterov, Yurii (1983). "A method for solving the convex programming problem with convergence rate O (1/k2)". In: *Dokl akad nauk Sssr*. Vol. 269, p. 543.

Nesterov, Yurii et al. (2018). *Lectures on convex optimization*. Vol. 137. Springer.

Orabona, Francesco (2023). "Normalized Gradients for All". In: *arXiv preprint arXiv:2308.05621* abs/2308.05621.

Pan, Rui, Haishan Ye, and Tong Zhang (2022). "Eigencurve: Optimal Learning Rate Schedule for SGD on Quadratic Objectives with Skewed Hessian Spectrums". In: *ICLR*.

Pan, Yan and Yuanzhi Li (2023). "Toward understanding why adam converges faster than sgd for transformers". In: *arXiv preprint arXiv:2306.00204*.

Paquette, Courtney et al. (2023). "Halting time is predictable for large models: A universality property and average-case analysis". In: *Foundations of Computational Mathematics* 23.2, pp. 597–673.

Park, Jea-Hyun, Abner J Salgado, and Steven M Wise (2021). "Preconditioned accelerated gradient descent methods for locally Lipschitz smooth objectives with applications to the solution of nonlinear PDEs". In: *Journal of Scientific Computing* 89.1, p. 17.

Paszke, Adam et al. (2019). "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32.

Patel, Vivak and Albert S. Berahas (2022). "Gradient descent in the absence of global Lipschitz continuity of the gradients: Convergence, divergence and limitations of its continuous approximation". In: *arXiv preprint arXiv:2210.02418*.

Polyak, Boris T (1987). "Introduction to optimization". In.

Streeter, Matthew and H. Brendan McMahan (2010). "Less Regret Via Online Conditioning". In: *arXiv preprint arXiv:1002.4862*.

Takezawa, Yuki et al. (2024). "Polyak Meets Parameter-free Clipped Gradient Descent". In: *CoRR* abs/2405.15010. DOI: 10.48550/ARXIV.2405.15010. arXiv: 2405.15010. URL: https://doi.org/10.48550/arXiv.2405.15010.

Vainsencher, Daniel, Han Liu, and Tong Zhang (2015). "Local Smoothness in Variance Reduced Optimization". In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 2179–2187.

Virtanen, Pauli et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272.

Vladarean, Maria-Luiza, Yura Malitsky, and Volkan Cevher (2021). "A first-order primal-dual method with adaptivity to local smoothness". In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 6171–6182.

Zhang, Bohang et al. (2020). “Improved Analysis of Clipping Algorithms for Non-convex Optimization”. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Zhang, Jingzhao et al. (2020). “Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity”. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Zhang, Junyu and Mingyi Hong (2020). “First-order algorithms without Lipschitz gradient: A sequential local optimization approach”. In: *arXiv preprint arXiv:2010.03194*.

Zhao, Weijing and He Huang (2024). “Adaptive stepsize estimation based accelerated gradient descent algorithm for fully complex-valued neural networks”. In: *Expert Systems with Applications* 236, p. 121166.

 https://doi.org/10.52202/079017-0473

A Proofs for Section 2

Lemma 2.2. *If f is convex and differentiable, then the point-wise directional smoothness satisfies,*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{D(x,y)}{2}\|y - x\|_2^2. \tag{5}$$

Proof. By the convexity of f we have

$$f(x) + \langle \nabla f(x), y - x \rangle \leq f(y).$$

Rearranging and then using Cauchy-Schwarz we get

$$\begin{aligned} f(x) &\leq f(y) + \langle \nabla f(x), x - y \rangle \\ &= f(y) + \langle \nabla f(y), x - y \rangle + \langle \nabla f(x) - \nabla f(y), x - y \rangle \\ &\leq f(y) + \langle \nabla f(y), x - y \rangle + \|\nabla f(x) - \nabla f(y)\|\|x - y\| \\ &= f(y) + \langle \nabla f(y), x - y \rangle + \frac{D(x,y)}{2}\|x - y\|^2. \end{aligned}$$

□

Lemma 2.4. *For any differentiable function f, the path-wise smoothness* (7) *satisfies*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{A(x,y)}{2}\|y - x\|_2^2. \tag{8}$$

Proof. Starting from the fundamental theorem of calculus,

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x \rangle &= \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle \, dt \\ &\leq \int_0^1 A(x,y)t\|x - y\|_2^2 dt \\ &= \frac{A(x,y)}{2}\|y - x\|_2^2. \end{aligned}$$

which completes the proof. □

Proposition 2.3. *There exists a convex, differentiable f and $x, y \in \mathbb{R}^d$ such that if $t < 2$, then*

$$f(y) > f(x) + \langle \nabla f(x), y - x \rangle + \frac{t\|\nabla f(x) - \nabla f(y)\|}{2\|y - x\|_2}\|y - x\|_2^2. \tag{6}$$

Proof. Let H_f denote the optimal pointwise directional smoothness associated with some convex and differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ (as defined in Equation (4)), and D_f denote the pointwise directional smoothness associated with f (as defined in Equation (4)). For any t, the statement of (6) is equivalent to saying $H_f > t\frac{\|\nabla f(x) - \nabla f(y)\|}{\|x - y\|}$ for all $x, y \in \mathbb{R}^d$ and convex, differentiable f. Observe that Lemma 2.2 already shows that for all convex and differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$

$$H_f(x,y) \leq D_f(x,y) = 2\frac{\|\nabla f(x) - \nabla f(y)\|}{\|x - y\|}$$

for all $x, y \in \mathbb{R}^d$. In order to show that this is tight, we suppose by the way of contradiction that there exists some $2 > t \geq 0$ such that for all convex and differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$

$$H_f(x,y) \leq t \cdot \frac{\|\nabla f(x) - \nabla f(y)\|}{\|x - y\|} \tag{31}$$

for all $x, y \in \mathbb{R}$. We shall show that no such t exists by showing for each such t there exists a function f_t such that Equation (31) does not hold.

Consider $f_\epsilon(x) = \sqrt{x^2 + \epsilon^2}$ for $\epsilon \leq 1$. The function f is differentiable. Moreover

$$f'_\epsilon(x) = \frac{x}{\sqrt{x^2 + \epsilon^2}}, \qquad f''_\epsilon(x) = \frac{\epsilon^2}{(\epsilon^2 + x^2)^{\frac{3}{2}}} \geq 0.$$

Therefore f is convex. Let $g(x) = |x|$. Fix $x = 1$ and $y = 0$, we have

$$\begin{aligned}
|g(x) - g(y) - \text{sign}(y) \cdot (x - y)| &\leq |f_\epsilon(x) - f_\epsilon(y) - f'_\epsilon(y) \cdot (x - y)| + |g(x) - f_\epsilon(x)| \\
&\quad + |g(y) - f_\epsilon(y)| + |(f'_\epsilon(y) - \text{sign}(y)) \cdot (x - y)| \\
&= |f_\epsilon(x) - f_\epsilon(y) - f'_\epsilon(y) \cdot (x - y)| + \left|1 - \sqrt{1 + \epsilon^2}\right| \\
&\quad + \left|0 - \sqrt{\epsilon^2}\right| + |(0 - 0) \cdot (1 - 0)| \\
&\leq |f_\epsilon(x) - f_\epsilon(y) - f'_\epsilon(y) \cdot (x - y)| + 2\epsilon.
\end{aligned}$$

Now observe that

$$g(x) - g(y) - \text{sign}(y) \cdot (x - y) = |1| - |0| - 0 \cdot (1 - 0) = 1.$$

Therefore

$$|f_\epsilon(x) - f_\epsilon(y) - f'_\epsilon(y) \cdot (x - y)| \geq 1 - 2\epsilon. \tag{32}$$

By definition we have $\frac{1}{2}\|x - y\|^2 = \frac{1}{2}$, therefore

$$H_f(x, y) \;=\; \frac{|f_\epsilon(x) - f_\epsilon(y) - f'_\epsilon(y) \cdot (x - y)|}{\frac{1}{2}\|x - y\|^2} \geq 2 - 4\epsilon. \tag{33}$$

But by our starting assumption we have that there exists some $t < 2$ such that $H_f(x, y) \leq t\frac{\|f'(x) - f'(y)\|}{|x - y|}$ for all differentiable and convex functions f. Applying this to $f = f_\epsilon$ we get

$$H_{f_\epsilon}(1, 0) \leq t\frac{|f'_\epsilon(1) - f'_\epsilon(0)|}{|1|} = t \cdot \frac{1}{\sqrt{1 + \epsilon^2}} \leq t. \tag{34}$$

Combining Equations (33) and (34) we have

$$2 - 4\epsilon \leq H_{f_\epsilon}(1, 0) \leq t$$

Rearranging we get

$$2 - t \leq 4\epsilon$$

Choosing $\epsilon = \frac{2-t}{8} > 0$ we get a contradiction. It follows that the minimal t such that $H(x, y) \leq t\frac{|f'(x) - f'(y)|}{|x - y|}$ for all convex and differentiable f is $t = 2$. □

Lemma A.1. *One step of gradient descent with step-size $\eta_k > 0$ makes progress as*

$$f(x_{k+1}) \leq f(x_k) - \eta_k \left(1 - \frac{\eta_k M(x_k, x_{k+1})}{2}\right) \|\nabla f(x_k)\|_2^2.$$

Proof. Starting from Equation (4), we have

$$\begin{aligned}
f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{M(x_k, x_{k+1})}{2}\|x_{k+1} - x_k\|_2^2 \\
&= f(x_k) - \eta_k \|\nabla f(x_k)\|_2^2 + \frac{\eta_k^2 M(x_k, x_{k+1})}{2}\|\nabla f(x_k)\|_2^2 \\
&= f(x_k) - \eta_k \left(\frac{1 - \eta_k M(x_k, x_{k+1})}{2}\right) \|\nabla f(x_k)\|_2^2.
\end{aligned}$$

□

B Proofs for Section 3

Lemma B.1. *If f is convex, then for any $x, y \in \mathbb{R}^d$,*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu(x, y)}{2}\|y - x\|_2^2. \tag{35}$$

If f is μ strongly convex, then $\mu(x, y) \geq \mu$.

https://doi.org/10.52202/079017-0473

Proof. The fundamental theorem of calculus implies

$$\begin{aligned} f(x) - \langle \nabla f(x), y - x \rangle &= \int_0^1 \langle \nabla f(x + t(y-x)) - \nabla f(x), y - x \rangle \, dt \\ &\geq \int_0^1 \mu(x,y) t \|x - y\|_2^2 dt \\ &= \frac{\mu(x,y)}{2} \|y - x\|_2^2. \end{aligned}$$

Note that we have implicitly used convexity to verify the inequality in the second line in the case where $\mu(x, y) = 0$. Now assume that f is μ strongly convex. As a standard consequence of strong-convexity, we obtain:

$$\begin{aligned} \frac{\langle \nabla f(x + t(y-x)) - \nabla f(x), y - x \rangle}{t\|x - y\|_2^2} &= \frac{\langle \nabla f(x + t(y-x)) - \nabla f(x), x + t(y-x) - x \rangle}{t^2\|x - y\|_2^2} \\ &\geq \mu \frac{\|x - t(y-x) - x\|_2^2}{t^2\|y - x\|_2^2} \\ &= \mu. \end{aligned}$$

□

Proposition 3.1. *If f is convex and differentiable, then GD with step-size sequence $\{\eta_k\}$ satisfies,*

$$\delta_k \leq \left[\prod_{i \in \mathcal{G}} (1 + \eta_i \lambda_i \mu_i) \right] \delta_0 + \sum_{i \in \mathcal{B}} \left[\prod_{j > i, j \in \mathcal{G}} (1 + \eta_j \lambda_j \mu_j) \right] \frac{\eta_i \lambda_i}{2} \|\nabla f(x_i)\|_2^2, \tag{13}$$

where $\lambda_i = \eta_i M_i - 2$, $\mathcal{G} = \{i : \eta_i < 2/M_i\}$, and $\mathcal{B} = [k] \backslash \mathcal{G}$.

Proof. First note that $\lambda_i < 0$ for $i \in \mathcal{G}$ and $\lambda_i \geq 0$ for $i \in \mathcal{B}$. We start from Equation (10),

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \eta_k \left(\frac{\eta_k M(x_k, x_{k+1})}{2} - 1 \right) \|\nabla f(x_k)\|_2^2 \\ &= f(x_k) + \mathbb{1}_{k \in \mathcal{G}} \cdot \left[\frac{\eta_k \lambda_k}{2} \|\nabla f(x_k)\|_2^2 \right] + \mathbb{1}_{k \in \mathcal{B}} \cdot \left[\frac{\eta_k \lambda_k}{2} \|\nabla f(x_k)\|_2^2 \right] \\ &\leq f(x_k) + \mathbb{1}_{k \in \mathcal{G}} \cdot \left[\eta_k \lambda_k \mu_k \left(f(x_k) - f(x^*) \right) \right] + \mathbb{1}_{k \in \mathcal{B}} \cdot \left[\frac{\eta_k \lambda_k}{2} \|\nabla f(x_k)\|_2^2 \right], \end{aligned}$$

where we used that directional strong convexity gives

$$\|\nabla f(x_k)\|_2^2 \geq 2\mu_k \left(f(x_k) - f(x^*) \right).$$

Subtracting $f(x^*)$ from both sides and then recursively applying the inequality gives the result. □

Proposition 3.2. *If f is convex and differentiable, then GD with step-size sequence $\{\eta_k\}$ satisfies,*

$$\Delta_k \leq \left[\prod_{i=0}^{k} \frac{|1 - \mu_i \eta_i|}{1 + \mu_{i+1} \eta_i} \right] \Delta_0 + \sum_{i=0}^{k} \left[\prod_{j > i} \frac{|1 - \mu_j \eta_j|}{1 + \mu_{j+1} \eta_j} \right] \frac{(M_i \eta_i^3 - \eta_i^2)}{1 + \mu_{i+1} \eta_i} \|\nabla f(x_i)\|_2^2. \tag{15}$$

Proof. Let $\Delta_k = \|x_k - x^*\|_2^2$ and observe

$$\Delta_k = \|x_k - x_{k+1} + x_{k+1} - x^*\|_2^2 = \Delta_{k+1} + \|x_k - x_{k+1}\|_2^2 + 2 \langle x_k - x_{k+1}, x_{k+1} - x^* \rangle.$$

Using this expansion in $\Delta_{k+1} - \Delta_k$, we obtain

$$\begin{aligned} \Delta_{k+1} - \Delta_k &= -\|x_k - x_{k+1}\|_2^2 - 2 \langle x_k - x_{k+1}, x_{k+1} - x^* \rangle \\ &= -\eta_k^2 \|\nabla f(x_k)\|_2^2 - 2\eta_k \langle \nabla f(x_k), x_{k+1} - x^* \rangle \\ &= -\eta_k^2 \|\nabla f(x_k)\|_2^2 - 2\eta_k \langle \nabla f(x_k), x_{k+1} - x_k \rangle - 2\eta_k \langle \nabla f(x_k), x_k - x^* \rangle. \end{aligned}$$

Now we control the inner-products with directional strong convexity and directional smoothness.

$$\begin{aligned}
&\leq -\eta_k^2\|\nabla f(x_k)\|_2^2 - 2\eta_k\left\langle \nabla f(x_k), x_{k+1}-x_k\right\rangle + 2\eta_k\left[f(x^*) - f(x_k) - \frac{\mu_k}{2}\Delta_k\right]\\
&\leq -\eta_k^2\|\nabla f(x_k)\|_2^2 + 2\eta_k\left[f(x_k) - f(x_{k+1}) + \frac{M(x_k,x_{k+1})\eta_k^2}{2}\|\nabla f(x_k)\|_2^2\right]\\
&\quad + 2\eta_k\left[f(x^*) - f(x_k) - \frac{\mu_k}{2}\Delta_k\right]\\
&= \eta_k^2\left(M(x_k,x_{k+1})\eta_k - 1\right)\|\nabla f(x_k)\|_2^2 + 2\eta_k\left[f(x^*) - f(x_{k+1})\right] - \mu_k\eta_k\Delta_k\\
&\leq \eta_k^2\left(M(x_k,x_{k+1})\eta_k - 1\right)\|\nabla f(x_k)\|_2^2 - \eta_k\mu_{k+1}\Delta_{k+1} - \mu_k\eta_k\Delta_k,
\end{aligned}$$

where the last inequality follows from μ_{k+1} strong convexity between x_{k+1} and x^*. Re-arranging this expression allows us to deduce a rate with error terms depending on the local smoothness,

$$\begin{aligned}
\implies (1+\mu_{k+1}\eta_k)\Delta_{k+1} &\leq (1-\mu_k\eta_k)\Delta_k + \eta_k^2\left(M(x_k,x_{k+1})\eta - 1\right)\|\nabla f(x_k)\|_2^2\\
&\leq |1-\mu_k\eta_k|\,\Delta_k + \eta_k^2\left(M(x_k,x_{k+1})\eta - 1\right)\|\nabla f(x_k)\|_2^2\\
\implies \Delta_{k+1} &\leq \frac{|1-\mu_k\eta_k|}{1+\mu_{k+1}\eta_k}\Delta_k + \frac{\eta_k^2\left(M(x_k,x_{k+1})\eta - 1\right)}{1+\mu_{k+1}\eta_k}\|\nabla f(x_k)\|_2^2\\
&\leq \left[\prod_{i=0}^{k}\frac{|1-\mu_i\eta_i|}{1+\mu_{i+1}\eta_i}\right]\Delta_0\\
&\quad + \sum_{i=0}^{k}\left[\prod_{j=i+1}^{k}\frac{|1-\mu_j\eta_j|}{1+\mu_{j+1}\eta_j}\right]\frac{\eta_i^2\left(M(x_i,x_{i+1})\eta_i - 1\right)}{1+\mu_{i+1}\eta_i}\|\nabla f(x_i)\|_2^2.
\end{aligned}$$

□

Proposition 3.3. *Let $\overline{x}_k = \sum_{i=0}^k \eta_i x_{i+1}/\sum_{i=0}^k \eta_i$. If f is convex and differentiable, then GD satisfies,*

$$f(\overline{x}_k) - f(x^*) \leq \frac{\|x_0 - x^*\|_2^2}{2\sum_{i=0}^k \eta_i} + \frac{\sum_{i=0}^k \eta_i^2(\eta_i M_i - 1)\|\nabla f(x_i)\|_2^2}{2\sum_{i=0}^k \eta_i}. \tag{16}$$

Proof. Let $\Delta_k = \|x_k - x^*\|_2^2$ and observe

$$\Delta_k = \|x_k - x_{k+1} + x_{k+1} - x^*\|_2^2 = \Delta_{k+1} + \|x_k - x_{k+1}\|_2^2 + 2\left\langle x_k - x_{k+1}, x_{k+1} - x^*\right\rangle.$$

Using this expansion in $\Delta_{k+1} - \Delta_k$, we obtain

$$\begin{aligned}
\Delta_{k+1} - \Delta_k &= -\|x_k - x_{k+1}\|_2^2 - 2\left\langle x_k - x_{k+1}, x_{k+1} - x^*\right\rangle\\
&= -\eta_k^2\|\nabla f(x_k)\|_2^2 - 2\eta_k\left\langle \nabla f(x_k), x_{k+1} - x^*\right\rangle\\
&= -\eta_k^2\|\nabla f(x_k)\|_2^2 - 2\eta_k\left\langle \nabla f(x_k), x_{k+1} - x_k\right\rangle - 2\eta_k\left\langle \nabla f(x_k), x_k - x^*\right\rangle.
\end{aligned}$$

Now we use convexity and directional smoothness to control the two inner-products as follows:

$$\begin{aligned}
\Delta_{k+1} - \Delta_k &\leq -\eta_k^2\|\nabla f(x_k)\|_2^2 - 2\eta_k\left(f(x_k) - f(x^*)\right) - 2\eta_k\left\langle \nabla f(x_k), x_{k+1} - x_k\right\rangle\\
&\leq -\eta_k^2\|\nabla f(x_k)\|_2^2 - 2\eta_k\left(f(x_k) - f(x^*)\right) + 2\eta_k(f(x_k) - f(x_{k+1}))\\
&\qquad + \eta_k^3 M(x_k,x_{k+1})\|\nabla f(x_k)\|_2^2\\
&= \eta_k^2(\eta_k M(x_k,x_{k+1}) - 1)\|\nabla f(x_k)\|_2^2 - 2\eta_k\left(f(x_{k+1}) - f(x^*)\right).
\end{aligned}$$

Re-arranging this equation and summing over iterations implies the following sub-optimality bound:

$$\sum_{i=0}^{k}\frac{\eta_i}{\sum_{i=0}^k \eta_i}\left(f(x_{i+1}) - f(x^*)\right) \leq \frac{\Delta_0 + \sum_{i=0}^k \eta_i^2(\eta_i M(x_i,x_{i+1}) - 1)\|\nabla f(x_i)\|_2^2}{2\sum_{i=0}^k \eta_i}.$$

Convexity of f and Jensen's inequality now imply the final result,

$$\implies f(\overline{x}_k) - f(x^*) \leq \frac{\Delta_0 + \sum_{i=0}^k \eta_i^2(\eta_i M(x_i,x_{i+1}) - 1)\|\nabla f(x_i)\|_2^2}{2\sum_{i=0}^k \eta_i}.$$

□

https://doi.org/10.52202/079017-0473

B.1 Path-Dependent Acceleration: Proofs

This section proves Theorem 3.4 using estimating sequences. Throughout this section, we assume $\mu \geq 0$ is the global strong convexity parameter, where $\mu = 0$ covers the non-strongly convex case. We start from the estimating sequences version of Equation (17), which is given as follows:

$$\begin{aligned} \alpha_k^2 &= \eta_k(1-\alpha_k)\gamma_k + \eta_k\alpha_k\mu \\ \gamma_{k+1} &= (1-\alpha_k)\gamma_k + \alpha_k\mu \\ y_k &= \frac{1}{\gamma_k + \alpha_k\mu}\left[\alpha_k\gamma_k v_k + \gamma_{k+1}x_k\right] \\ x_{k+1} &= y_k - \eta_k\nabla f(y_k) \\ v_{k+1} &= \frac{1}{\gamma_{k+1}}\left[(1-\alpha_k)\gamma_k v_k + \alpha_k\mu y_k - \alpha_k\nabla f(y_k)\right]. \end{aligned} \tag{36}$$

The algorithm is initialized with $x_0 = v_0$ and some $\gamma_0 > 0$. Note that $y_0 = x_0 = v_0$ since y_k is a convex combination of x_k and v_k. First we prove that this scheme is equivalent to the one given in Equation (17).

Lemma B.2. *Equation* (36) *and Equation* (17) *lead to equivalent updates for the y_k, x_k, and α_k sequences. Moreover, given initialization $\gamma_0 > 0$, the corresponding initialization for α_0 is,*

$$\alpha_0 = \frac{\eta_0}{2}\left((\mu - \gamma_0) + \sqrt{(\gamma_0 - \mu)^2 + 4\gamma_0/\eta_0}\right). \tag{37}$$

Proof. The proof follows Nesterov et al. (2018, Theorem 2.2.3). Expanding the definition of v_{k+1}, we obtain

$$\begin{aligned} v_{k+1} &= \frac{1}{\gamma_{k+1}}\left[\frac{(1-\alpha_k)}{\alpha_k}\left[(\gamma_k + \alpha_k\mu)y_k - \gamma_{k+1}x_k\right] + \alpha_k\mu y_k - \alpha_k\nabla f(y_k)\right] \\ &= \frac{1}{\gamma_{k+1}}\left[\frac{(1-\alpha_k)\gamma_k}{\alpha_k}y_k + \mu y_k\right] - \frac{(1-\alpha_k)}{\alpha_k}x_k - \frac{\alpha_k}{\gamma_{k+1}}\nabla f(y_k) \\ &= x_k - \frac{\eta_k}{\alpha_k}\nabla f(y_k) + \frac{1}{\alpha_k}(y_k - x_k) \\ &= x_k + \frac{1}{\alpha_k}\left(x_{k+1} - x_k\right). \end{aligned}$$

Plugging this back into the expression for y_{k+1},

$$\begin{aligned} y_{k+1} &= \frac{1}{\gamma_{k+1} + \alpha_{k+1}\mu}\left[\alpha_{k+1}\gamma_{k+1}v_{k+1} + \gamma_{k+2}x_{k+1}\right] \\ &= \frac{1}{\gamma_{k+1} + \alpha_{k+1}\mu}\left[\alpha_{k+1}\gamma_{k+1}(x_k + \frac{1}{\alpha_k}\left(x_{k+1} - x_k\right)) + \gamma_{k+2}x_{k+1}\right] \\ &= \frac{\alpha_{k+1}\gamma_{k_1} + \alpha_k(1-\alpha_{k+1})\gamma_{k+1} + \alpha_k\alpha_{k+1}\mu}{\alpha_k(\gamma_{k+1} + \alpha_{k+1}\mu)}x_{k+1} - \frac{\alpha_{k+1}\gamma_{k+1}(1-\alpha_k)}{\alpha_k(\gamma_{k+1} + \alpha_{k+1}\mu)}x_k \\ &= x_{k+1} + \frac{\alpha_{k+1}\gamma_{k+1}(1-\alpha_k)}{\alpha_k(\gamma_{k+1} + \alpha_{k+1}\mu)}(x_{k+1} - x_k) \\ &= x_{k+1} + \frac{\alpha_{k+1}\gamma_{k+1}(1-\alpha_k)}{\alpha_k\left(\gamma_{k+1} + \alpha_{k+1}^2/\eta_k - (1-\alpha_{k+1})\gamma_{k+1}\right)}(x_{k+1} - x_k) \\ &= x_{k+1} + \frac{\alpha_k(1-\alpha_k)}{(\alpha_{k+1} + \alpha_k^2)}(x_{k+1} - x_k). \end{aligned}$$

Note that this update is consistent with Equation (17). Since $\gamma_k = \alpha_k^2/\eta_k$, we can write,

$$\begin{aligned} \alpha_{k+1}^2 &= \eta_{k+1}(1-\alpha_{k+1})\gamma_k + \eta_{k+1}\alpha_{k+1}\mu \\ &= \frac{\eta_{k+1}}{\eta_k}(1-\alpha_{k+1})\alpha_k^2 + \eta_{k+1}\alpha_{k+1}\mu, \end{aligned}$$

which is also consistent with Equation (17). Finally, the initialization for α_0 is determined by γ_0 in Equation (36) as,

$$\alpha_0^2 = \eta_0(1-\alpha_0)\gamma_0 + \eta_0\alpha_0\mu.$$

The quadratic formula now implies,

$$\alpha_0 = \frac{\eta_0}{2}\left((\mu-\gamma_0) + \sqrt{(\gamma_0-\mu)^2 + 4\gamma_0/\eta_0}\right).$$

This completes the proof. □

As mentioned, our proof uses the concept of estimating sequences.

Definition B.3. Two sequences λ_k, ϕ_k are estimating sequences for f if $\lambda_l \geq 0$ for all $k \in \mathbb{N}$, $\lim_{k\to\infty}\lambda_k = 0$, and,

$$\phi_k(x) \leq (1-\lambda_k)f(x) + \lambda_k\phi_0(x), \tag{38}$$

for all $x \in \mathbb{R}^d$.

We use the same estimating sequences as developed by Nesterov et al. (2018). Let $\lambda_0 = 1$, $\phi_0(x) = f(x_0) + \frac{\gamma_0}{2}\|x - x_0\|_2^2$, and define the updates,

$$\begin{aligned}\lambda_{k+1} &= (1-\alpha_k)\lambda_k \\ \phi_{k+1}(x) &= (1-\alpha_k)\phi_k(x) + \alpha_k\left(f(y_k) + \langle\nabla f(y_k), x - y_k\rangle + \frac{\mu}{2}\|x-y_k\|_2^2\right),\end{aligned} \tag{39}$$

where $\mu \geq 0$ is the strong convexity parameter, with $\mu = 0$ when f is merely convex. It is straightforward to differentiate ϕ_{k+1} to see that v_{k+1} of Equation (36) is the minimizer. Indeed, Nesterov et al. (2018, Lemma 2.2.3) shows that this choice for the estimating sequences obeys the following canonical form:

$$\phi_{k+1}(x) = \min_z \phi_{k+1}(z) + \frac{\gamma_{k+1}}{2}\|x - v_{k+1}\|_2^2, \tag{40}$$

where γ_{k+1} and v_{k+1} are given by Equation (36) and the minimum value is,

$$\begin{aligned}\min_z \phi_{k+1}(z) = (1-\alpha_k)\min_z \phi_k(z) + \alpha_k f(y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(y_k)\|_2^2 \\ + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|y_k - v_k\|_2^2 + \langle\nabla f(y_k), v_k - y_k\rangle\right).\end{aligned} \tag{41}$$

Before we can prove our main theorem, we must show that these choices for λ_k and ϕ_k yield a valid estimating sequence. The following proofs build on (Nesterov et al., 2018) and (Mishkin et al., 2024).

Lemma B.4. *Assume $\alpha_k \in (0,1]$ for all $k \in \mathbb{N}$. If $\mu > 0$ and $\gamma_0 = \mu$, then*

$$\lambda_k = \prod_{i=0}^{k-1}(1 - \sqrt{\eta_k\mu}). \tag{42}$$

If $\mu \geq 0$ and $\gamma_0 \in (\mu, \mu + 3/\eta_{min})$, then,

$$\lambda_k \leq \frac{4}{\eta_{min}(\gamma_0 - \mu)(k+1)^2}. \tag{43}$$

Proof. Assume $\gamma_0 = \mu > 0$. Then $\gamma_k = \mu$ for all k and,

$$\begin{aligned}\alpha_k^2 &= (1-\alpha_k)\eta_k\mu + \alpha_k\eta_k\mu \\ &= \eta_k\mu.\end{aligned}$$

As a consequence,

$$\lambda_k = \prod_{i=0}^{k-1}(1 - \sqrt{\eta_k\mu}),$$

as claimed.

 https://doi.org/10.52202/079017-0473

Now assume $\gamma_0 \in (\mu, 3L + \mu)$. Modifying the proof by Nesterov et al. (2018, Lemma 2.2.4), we compute as follows:

$$\gamma_{k+1} - \mu = (1 - \alpha_k)\gamma_k + (\alpha_k - 1)\mu = (1 - \alpha_k)(\gamma_k - \mu)$$

Recursing on this equality implies

$$\gamma_{k+1} = (\gamma_0 - \mu) \prod_{i=0}^{k} (1 - \alpha_k) = \lambda_{k+1}(\gamma_0 - \mu).$$

If $\alpha_k = 1$ or $\lambda_k = 0$, then using $\lambda_{k+1} = (1 - \alpha_k)\lambda_k$ implies $\lambda_{k+1} = 0$ and the result trivially holds. Otherwise, recall $\alpha_k^2/\gamma_{k+1} = \eta_k$ to obtain,

$$\begin{aligned} 1 - \frac{\lambda_{k+1}}{\lambda_k} &= \alpha_k = (\gamma_{k+1}\eta_k)^{1/2} \\ &= (\eta_k\mu + \eta_k\lambda_{k+1}(\gamma_0 - \mu))^{1/2} \\ \implies \frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k} &= \frac{1}{\lambda_{k+1}^{1/2}} \left[\frac{\eta_k\mu}{\lambda_{k+1}} + \eta_k(\gamma_0 - \mu)\right]^{1/2} \\ &\geq \frac{1}{\lambda_{k+1}^{1/2}} \left[\frac{\eta_{\min}\mu}{\lambda_{k+1}} + \eta_{\min}(\gamma_0 - \mu)\right]^{1/2}. \end{aligned}$$

Finally, this implies

$$\begin{aligned} \frac{2}{\lambda_{k+1}^{1/2}} \left(\frac{1}{\lambda_{k+1}^{1/2}} - \frac{1}{\lambda_k^{1/2}}\right) &\geq \left(\frac{1}{\lambda_{k+1}^{1/2}} - \frac{1}{\lambda_k^{1/2}}\right) \left(\frac{1}{\lambda_{k+1}^{1/2}} + \frac{1}{\lambda_k^{1/2}}\right) \\ &\geq \frac{1}{\lambda_{k+1}^{1/2}} \left[\frac{\eta_{\min}\mu}{\lambda_{k+1}} + \eta_{\min}(\gamma_0 - \mu)\right]^{1/2}. \end{aligned}$$

Moreover, this bound holds uniformly for all $k \in \mathbb{N}$. We have now exactly reached Eq. 2.2.11 of Nesterov et al. (2018, Lemma 2.2.4) with L replaced by $1/\eta_{\min}$. Applying that Lemma with this modification, we obtain

$$\lambda_k \leq \frac{4}{\eta_{\min}(\gamma_0 - \mu)(k+1)^2},$$

which completes the proof. □

Lemma B.5. *If f is strongly convex with parameter $\mu \geq 0$ and $\eta_k \leq 1/\mu$ for all $k \in \mathbb{N}$, then λ_k and ϕ_k are estimating sequences.*

Proof. Using the quadratic formula, we find

$$\alpha_k = \frac{\mu - \gamma_k \pm \sqrt{(\mu - \gamma_k)^2 + 4\hat{\gamma}_k/\eta_k}}{2/\eta_k}.$$

Thus,

$$(\mu - \gamma_k) + \left((\mu - \gamma_k)^2 + 4\hat{\gamma}_k/\eta_k\right)^{1/2} > 0.$$

is sufficient for $\alpha_k > 0$. This holds if $\mu \geq \gamma_k$. Otherwise, we require,

$$(\mu - \gamma_k)^2 + 4\hat{\gamma}_k/\eta_k > (\mu - \gamma_k)^2,$$

which holds if and only if $\eta_k, \gamma_k > 0$. On the other hand, we also need $\alpha_k \leq 1$, which is satisfied when,

$$4 + 4\eta_k(\gamma_k - \mu) + \eta_k^2(\mu - \gamma_k)^2 \leq \eta_k^2(\mu - \gamma_k)^2 + 4\eta_k\gamma_k \iff \eta_k \leq \frac{1}{\mu},$$

as claimed.

Recall $\lambda_0 = 1$ and $\lambda_{k+1} = (1 - \alpha_k)\lambda_k$. Since $\alpha_k \in (0, 1]$, $\lambda_k \geq 0$ holds by induction. It remains to show that λ_k tends to zero, which holds by Lemma B.4 since we have shown $\alpha_k \in (0, 1]$ for all k.

Now we establish the last piece,

$$\phi_k(x) \leq (1 - \lambda_k)f(x) + \lambda_k\phi_0(x).$$

But this follows immediately by Nesterov et al. (2018, Lemma 2.2.2). □

Now we can prove the last major lemma before our convergence result.

Lemma B.6. *Suppose f is strongly convex with parameter $\mu \geq 0$ and η_k is a sequence of adapted step-sizes, meaning $\eta_k \leq 1/M(x_k, x_{k+1})$. Then for every $k \in \mathbb{N}$,*

$$\min_z \phi_k(z) \geq f(x_k).$$

Proof. We use an inductive proof again. The inductive assumption is

$$\min_z \phi_k(z) \geq f(x_k),$$

It is easy to see this holds at $k = 0$ since,

$$\phi_0(x) = f(x_0) + \frac{\gamma_0}{2}\|x - v_0\|_2^2,$$

implies $\min_z \phi_0(z) = f(x_0)$. Using Equation (41), we obtain

$$\begin{aligned}
\min_z \phi_{k+1}(z) &= (1-\alpha_k)\min_z \phi_k(z) + \alpha_k f(y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(y_k)\|^2 \\
&\quad + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|y_k - v_k\|^2 + \langle \nabla f(y_k, z_k), v_k - y_k\rangle\right) \\
&\geq (1-\alpha_k) f(x_k) + \alpha_k f(y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(y_k)\|^2 \\
&\quad + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|y_k - v_k\|^2 + \langle \nabla f(y_k), v_k - y_k\rangle\right),
\end{aligned}$$

where the inequality holds by the inductive assumption. Using convexity of f and recalling $\frac{\alpha_k^2}{\gamma_{k+1}} = \eta_k$ from the definition of the update (Equation (36)),

$$\begin{aligned}
\min_z \phi_{k+1}(z) &\geq (1-\alpha_k)\left(f(y_k) + \langle \nabla f(y_k), x_k - y_k\rangle\right) + \alpha_k f(y_k) - \frac{\eta_k}{2}\|\nabla f(y_k)\|^2 \\
&\quad + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|y_k - v_k\|^2 + \langle \nabla f(y_k), v_k - y_k\rangle\right) \\
&= f(y_k) + (1-\alpha_k)\langle \nabla f(y_k), x_k - y_k\rangle - \frac{\eta_k}{2}\|\nabla f(y_k)\|^2 \\
&\quad + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|y_k - v_k\|^2 + \langle \nabla f(y_k), v_k - y_k\rangle\right)
\end{aligned}$$

Using the fact that the step-sizes are adapted and invoking the directional descent lemma (i.e. Equation (18)) now implies

$$\begin{aligned}
\min_z \phi_{k+1}(z) \geq f(x_{k+1}) + (1-\alpha_k)\Bigg(&\langle \nabla f(y_k), x_k - y_k\rangle \\
&+ \frac{\alpha_k\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|y_k - v_k\|^2 + \langle \nabla f(y_k), v_k - y_k\rangle\right)\Bigg).
\end{aligned}$$

The remainder of the proof is largely unchanged from the analysis in Nesterov et al. (2018). The definition of y_k gives $x_k - y_k = \frac{\alpha_k\gamma_k}{\gamma_{k+1}}(y_k - v_k)$, which we use to obtain

$$\begin{aligned}
\min_z \phi_{k+1}(z) &\geq f(x_{k+1}) + (1-\alpha_k)\Bigg(\frac{\alpha_k\gamma_k}{\gamma_{k+1}}\langle \nabla f(y_k), y_k - v_k\rangle \\
&\quad + \frac{\alpha_k\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|y_k - v_k\|^2 + \langle \nabla f(y_k), v_k - y_k\rangle\right)\Bigg) \\
&= f(x_{k+1}) + \frac{\mu\alpha_k(1-\alpha_k)\gamma_k}{2\gamma_{k+1}}\|y_k - v_k\|^2 \\
&\geq f(x_{k+1}),
\end{aligned}$$

since $\frac{\mu\alpha_k(1-\alpha_k)\gamma_k}{2\gamma_{k+1}} \geq 0$. We conclude the desired result by induction. □

 https://doi.org/10.52202/079017-0473

The main accelerated result now follows almost immediately.

Theorem 3.4. *Suppose f is differentiable, μ–strongly convex and AGD is run with adapted step-sizes $\eta_k \leq 1/M_k$. If $\mu > 0$ and $\alpha_0 = \sqrt{\eta_0 \mu}$, then AGD obtains the following accelerated rate:*

$$f(x_{k+1}) - f(x^*) \leq \prod_{i=0}^{k} (1 - \sqrt{\mu \eta_i}) \left[f(x_0) - f(x^*) + \frac{\mu}{2} \|x_0 - x^*\|_2^2 \right]. \tag{19}$$

Let $\eta_{min} = \min_{i \in [k]} \eta_i$. If $\mu \geq 0$ and $\alpha_0 \in (\sqrt{\mu \eta_0}, c)$, where c is the maximum value of α_0 for which $\gamma_0 = \frac{\alpha_0^2 - \eta_0 \alpha_0 \mu}{\eta_0 (1 - \alpha_0)}$ satisfies $\gamma_0 < 3/\eta_{min} + \mu$, then AGD obtains the following rate:

$$f(x_{k+1}) - f(x^*) \leq \frac{4}{\eta_{min}(\gamma_0 - \mu)(k+1)^2} \left[f(x_0) - f(x^*) + \frac{\gamma_0}{2} \|x_0 - x^*\|_2^2 \right]. \tag{20}$$

Proof. We analyze the equivalent formulation given in Equation (36). See Lemma B.2 for a formal proof that these two schemes produce the same x_k, y_k, and α_k iterates. Note that our proof follows Nesterov et al. (2018) and Mishkin et al. (2024) closely; while their results are very similar, we are not aware of pre-existing works which adapt them to our specific setting.

First, observe that $M(x_k, x_{k+1}) \geq \mu$ for all $k \in \mathbb{N}$. Since the step-sizes η_k are assumed to satisfy $\eta_k \leq 1/M(x_k, x_{k+1})$, we also have that $\eta_k \leq 1/\mu$ for every k.

Thus, Lemma B.4 and Lemma B.5 apply. Using the definition of an estimating sequence and Lemma B.6, we obtain,

$$\begin{aligned} f(x_k) &\leq \min_z \phi_k(z) \\ &\leq \min_z (1 - \lambda_k) f(z) + \lambda_k \phi_0(z) \\ &\leq (1 - \lambda_k) f(x^*) + \lambda_k \phi_0(x^*). \end{aligned}$$

Re-arranging this equation and expanding the definition ϕ_0 (Equation (39)), we deduce the following:

$$\begin{aligned} f(x_k) - f(x^*) &\leq \lambda_k \left(\phi_0(x^*) - f(x^*) \right) \\ &= \lambda_k \left(f(x_0) - f(x^*) + \frac{\gamma_0}{2} \|x_0 - x^*\|_2^2 \right). \end{aligned}$$

We see that the rate of convergence of AGD is entirely controlled by the convergence of the sequence λ_k. If $\mu > 0$ and $\gamma_0 = \mu$, then Lemma B.4 implies

$$f(x_k) - f(x^*) \leq \prod_{i=0}^{k-1} (1 - \sqrt{\mu \eta_i}) \left[f(x_0) - f(x^*) + \frac{\mu}{2} \|x_0 - x^*\|_2^2 \right].$$

By Lemma B.2, this initialization is equivalent to choosing $\alpha_0 = \sqrt{\eta_0 \mu}$, which is the setting claimed in the theorem.

Alternatively, if $\mu \geq 0$ and $\gamma_0 \in (\mu, \mu + 3/\eta_{\min})$, then,

$$f(x_k) - f(x^*) \leq \frac{4}{\eta_{\min}(\gamma_0 - \mu) k^2} \left[f(x_0) - f(x^*) + \frac{\gamma_0}{2} \|x_0 - x^*\|_2^2 \right],$$

where the equality,

$$\gamma_0 = \frac{\alpha_0^2 - \eta_0 \alpha_0 \mu}{\eta_0 (1 - \alpha_0)},$$

holds by Lemma B.2. Since $\alpha_0 \leq 1$ for $\eta_0 \leq 1/\mu$, η_0 is an increasing function of γ_0. Thus, an upper-bound c on α_0 can be deduced from that on γ_0 using the quadratic formula:

$$\begin{aligned} c &= -\frac{3\eta_0}{2\eta_{\min}} + \frac{\eta_0}{2} \left(\frac{9}{(\eta_{\min})^2} + 4 \frac{3\eta_{\min} + \mu}{\eta_0} \right)^{1/2} \\ &= \frac{3\eta_0}{2\eta_{\min}} \left[\left(1 + 4(\eta_{\min})^2 \frac{3\eta_{\min} + \mu}{9\eta_0} \right)^{1/2} - 1 \right]. \end{aligned}$$

□

C Proofs for Section 4.1

Lemma C.1. *Let B be a positive semi-definite matrix and suppose that*

$$f(x) = \frac{1}{2}x^\top Bx - c^\top x.$$

Let $x_{i+1} = x_i - \eta\nabla f(x_i)$. Then for any $\eta > 0$, the pointwise directional smoothness between the gradient descent iterates x_i, x_{i+1} is given by

$$\frac{1}{2}D(x_i, x_{i+1}) = \frac{\|B\nabla f(x_i)\|_2}{\|\nabla f(x_i)\|_2}.$$

Proof. We have by straightforward algebra,

$$\begin{aligned}
\frac{1}{2}D(x_i, x_{i+1}) &= \frac{\|\nabla f(x_{i+1}) - \nabla f(x_i)\|_2}{\|x_{i+1} - x_i\|_2} \\
&= \frac{\|\left[Bx_{i+1} - c\right] - \left[Bx_i - c\right]\|_2}{\|x_{i+1} - x_i\|_2} \\
&= \frac{\|B\left[x_{i+1} - x_i\right]\|_2}{\|x_{i+1} - x_i\|_2} \\
&= \frac{\|B\left[-\eta\nabla f(x_i)\right]\|_2}{\|-\eta\nabla f(x_i)\|_2} \\
&= \frac{\|B\nabla f(x_i)\|_2}{\|\nabla f(x_i)\|_2}.
\end{aligned}$$

□

Lemma C.2. *Let B be a positive semi-definite matrix and suppose that*

$$f(x) = \frac{1}{2}x^\top Bx - c^\top x.$$

Let $x_{i+1} = x_i - \eta\nabla f(x_i)$. Then for any $\eta > 0$, the path-wise directional smoothness between the gradient descent iterates x_i, x_{i+1} is given by by

$$A(x_i, x_{i+1}) = \frac{\nabla f(x_i)^\top B\nabla f(x_i)}{\nabla f(x_i)^\top \nabla f(x_i)}.$$

Proof. Let $A_t(x, y) = \frac{\langle \nabla f(x+t(y-x)) - \nabla f(x), y-x\rangle}{t\|x-y\|_2^2}$. We have

$$\begin{aligned}
A_t(x, y) &= \frac{\langle \nabla f(x + t(y-x)) - \nabla f(x), y - x\rangle}{t\|x-y\|_2^2} \\
&= \frac{\langle (B(x + t(y-x))) - c - \left[Bx - c\right], y - x\rangle}{t\|x-y\|_2^2} \\
&= \frac{\langle t \cdot B(y-x), y - x\rangle}{t\|x-y\|_2^2} \\
&= \frac{(y-x)^\top B(y-x)}{\|x-y\|_2^2}.
\end{aligned}$$

The path-wise directional smoothness A is therefore

$$\begin{aligned}
A(x, y) &= \sup_{t\in[0,1]} A_t(x, y) \\
&= \sup_{t\in[0,1]} \frac{(y-x)^\top B(y-x)}{\|x-y\|_2^2} \\
&= \frac{(y-x)^\top B(y-x)}{\|x-y\|_2^2}.
\end{aligned}$$

https://doi.org/10.52202/079017-0473

Plugging in $y = x - \eta\nabla f(x) = x - \eta[Bx - c]$ in the above gives

$$\begin{aligned} A(x, x - \eta\nabla f(x)) &= \frac{(-\eta\,[Bx - c])\, B(-\eta)\,[Bx - c]}{\|\eta[Bx - c]\|_2^2} \\ &= \frac{(Bx - c)^\top B(Bx - c)}{\|Bx - c\|_2^2} \\ &= \frac{(Bx - c)^\top B(Bx - c)}{\|Bx - c\|_2^2} \\ &= \frac{\nabla f(x)^\top B\nabla f(x)}{\nabla f(x)^\top \nabla f(x)}. \end{aligned}$$

□

D Proofs for Section 4.2

Proposition 4.1. *If f is convex and continuously differentiable, then either (i) f is minimized along the ray $x(\eta) = x - \eta\nabla f(x)$ or (ii) there exists $\eta > 0$ satisfying $\eta = 1/D(x, x - \eta\nabla f(x))$.*

Proof. Let $\mathcal{I} = \{\eta : \nabla f(x - \eta\nabla f(x)) = \nabla f(x)\}$. For every $\eta \in \mathcal{I}$, it holds that

$$-\langle \nabla f(x - \eta\nabla f(x)), \nabla f(x)\rangle = -\|\nabla f(x)\|_2^2.$$

However, since f is convex, the directional derivative

$$-\langle \nabla f(x - \eta'\nabla f(x)), \nabla f(x)\rangle\,,$$

is monotone non-decreasing in η'. We deduce that $\mathcal{I}$ must be an interval of form $[0, \overline{\eta}]$. If $\overline{\eta}$ is not bounded, then f is linear along $-\nabla f(x)$ and is minimized by taking $\eta \to \infty$. Therefore, we may assume $\overline{\eta}$ is finite.

Let $\eta > \overline{\eta}$. Then we have the following:

$$\begin{aligned} x - \eta\nabla f(x) &= x - \frac{2\|x - \eta\nabla f(x) - x\|_2}{\|\nabla f(x - \eta\nabla f(x)) - \nabla f(x)\|_2}\nabla f(x) \\ \iff \nabla f(x) &= \frac{2\|\nabla f(x)\|_2}{\|\nabla f(x - \eta\nabla f(x)) - \nabla f(x)\|_2}\nabla f(x), \end{aligned}$$

from which we deduce

$$\|\nabla f(x - \eta\nabla f(x)) - \nabla f(x)\|_2 = 2\|\nabla f(x)\|_2,$$

is sufficient for the implicit equation to hold. Squaring both sides and multiplying by $1/2$, we obtain the following alternative root-finding problem:

$$h(\eta) := \frac{1}{2}\|\nabla f(x - \eta\nabla f(x))\|_2^2 - \langle \nabla f(x - \eta\nabla f(x)), \nabla f(x)\rangle - \frac{1}{2}\|\nabla f(x)\|_2^2 = 0. \tag{44}$$

Because f is C^1, h is a continuous function and it suffices to show that there exists an interval in which h crosses 0. From the display above, we see

$$h(\overline{\eta}) = -\|\nabla f(x)\|_2^2 < 0.$$

Continuity now implies $\exists\eta' > \overline{\eta}$ such that $h(\eta') < 0$. Now, suppose $h(\eta) \le 0$ for all $\eta \ge \eta'$. Working backwards, we see that this can only occur when

$$\eta \le \frac{2\|x - \eta\nabla f(x) - x\|_2}{\|\nabla f(x - \eta\nabla f(x)) - \nabla f(x)\|_2} = \frac{1}{D(x(\eta), x - \eta\nabla f(x))}$$

for all $\eta \ge \eta'$. The directional descent lemma (Equation (10)) now implies

$$f(x - \eta\nabla f(x)) \le f(x) - \eta\left(1 - \frac{\eta D(x, x - \eta\nabla f(x))}{2}\right)\|\nabla f(x)\|_2^2 \le f(x) - \frac{\eta}{2}\|\nabla f(x)\|_2^2,$$

Taking limits on both sides as $\eta \to \infty$ implies $f(x - \eta\nabla f(x))$ is minimized along the ray $x(\eta) = x - \eta\nabla f(x)$. Thus, we deduce that either there exists $\eta'' > \eta'$ such that $h(\eta'') > 0$ exists, or f is minimized along the gradient direction as claimed. □

Proposition 4.2. *If f is convex and twice continuously differentiable, then either (i) f is minimized along the ray $x(\eta) = x - \eta\nabla f(x)$ or (ii) there exists $\eta > 0$ satisfying $\eta = 1/A(x, x - \eta\nabla f(x))$.*

Proof. Let

$$\mathcal{J} = \left\{\eta : \langle \nabla f(x - \eta\nabla f(x)), \nabla f(x)\rangle = \|\nabla f(x)\|_2^2\right\}.$$

Since f is convex, the directional derivative

$$-\langle \nabla f(x - \eta'\nabla f(x)), \nabla f(x)\rangle,$$

is monotone non-decreasing in η'. We deduce that $\mathcal{J}$ must be an interval of form $[0, \overline{\eta}]$. If $\overline{\eta}$ is not bounded, then convexity implies

$$\begin{aligned}\lim_{\eta\to\infty} f(x - \eta\nabla f(x)) &\leq \lim_{\eta\to\infty} f(x) - \eta\langle \nabla f(x - \eta\nabla f(x)), \nabla f(x)\rangle \\ &= -\infty,\end{aligned}$$

meaning f is minimized along $-\nabla f(x)$. Therefore, we may assume $\overline{\eta}$ is finite.

We have

$$\begin{aligned}x - \eta\nabla f(x) = x - \frac{1}{A(x, x - \eta\nabla f(x))}\nabla f(x) \\ \iff \eta = \inf_{t\in[0,1]} \frac{t\eta\|\nabla f(x)\|_2^2}{\langle \nabla f(x) - \nabla f(x - t\eta\nabla f(x)), \nabla f(x)\rangle}.\end{aligned}$$

Thus, for $\eta > \overline{\eta}$, the equation we must solve reduces to

$$h(\eta) := \eta - \inf_{t\in[0,1]} \frac{t\eta\|\nabla f(x)\|_2^2}{\langle \nabla f(x) - \nabla f(x - t\eta\nabla f(x)), \nabla f(x)\rangle} = 0.$$

Since f is C^2, h is continuous (see, e.g. Hogan (1973, Theorem 7)) and it suffices to show that there exists an interval over which h crosses 0.

Using Taylor's theorem, we can re-write this expression as

$$h(\eta) = \eta - \inf_{t\in[0,1]} \frac{\|\nabla f(x)\|_2^2}{\langle \nabla f(x), \nabla^2 f(x - \alpha(t\eta)\nabla f(x))\nabla f(x)\rangle},$$

where for some $\alpha(t\eta) \in [0, t\eta]$. Examining the denominator, we find that,

$$\int_0^t \nabla f(x)^\top \nabla^2 f(x - t\overline{\eta}\nabla f(x))\nabla f(x)dt = \langle \nabla f(x - \overline{\eta}\nabla f(x)) - \nabla f(x), \nabla f(x)\rangle = 0,$$

which, since f is convex, implies

$$\nabla f(x)^\top \nabla^2 f(x - \alpha\nabla f(x))\nabla f(x) = 0,$$

for every $\alpha \in [0, \overline{\eta}]$. By continuity of the Hessian, for every $\epsilon > 0$, there exists $\delta > 0$ such that $\eta' \in [\overline{\eta}, \overline{\eta} + \delta]$ guarantees,

$$\nabla f(x)^\top \nabla^2 f(x - \eta'\nabla f(x))\nabla f(x) < \epsilon.$$

Substituting this into our expression for h,

$$\begin{aligned}h(\eta') &= \eta' - \inf_{t\in[0,1]} \frac{\|\nabla f(x)\|_2^2}{\langle \nabla f(x), \nabla^2 f(x - \alpha(t\eta')\nabla f(x))\nabla f(x)\rangle} \\ &< \overline{\eta} + \delta - \frac{\|\nabla f(x)\|_2^2}{\epsilon} \\ &< 0,\end{aligned}$$

for ϵ, δ sufficiently small. Thus, there exists $\eta' > \overline{\eta}$ for which $h(\eta') < 0$.

Now let us show that $h(\eta'') > 0$ for some η''. For convenience, define

$$g(\eta) = \inf_{t\in[0,1]} \frac{t\|\nabla f(x)\|_2^2}{\langle \nabla f(x) - \nabla f(x - t\eta\nabla f(x)), \nabla f(x)\rangle},$$

 https://doi.org/10.52202/079017-0473

Algorithm 1 Gradient Descent with Exponential Search

Procedure ExponentialSearch(x, η_0)
for $k = 1, 2, 3, \ldots$ **do**
  $\eta_{\text{out}} \leftarrow \text{RootFindingBisection}\left(x, 2^{-2^k}\eta_0, \eta_0\right)$.
  if $\eta_{\text{out}} < \infty$ **then**
    Return η_{out}
  end if
end for
End Procedure
Procedure RootFindingBisection($x, \eta_{\text{lo}}, \eta_{\text{hi}}$)
Define $\phi(\eta) = \eta - \psi(\eta)$ where $\psi(\eta)$ is given in (24) \\ One access to ϕ requires T descent steps.
if $\phi(\eta_{\text{hi}}) \leq 0$ **then**
  Return η_{hi}
end if
if $\phi(\eta_{\text{lo}}) > 0$ **then**
  Return ∞
end if
while $\eta_{\text{hi}} > 2\eta_{\text{lo}}$ **do**
  $\eta_{\text{mid}} = \sqrt{\eta_{\text{lo}}\eta_{\text{hi}}}$
  if $\phi(\eta_{\text{mid}}) > 0$ **then**
    $\eta_{\text{hi}} = \eta_{\text{mid}}$
  else
    $\eta_{\text{lo}} = \eta_{\text{mid}}$
  end if
  \\ Invariant: $\phi(\eta_{\text{hi}}) > 0$, and $\phi(\eta_{\text{lo}}) \leq 0$.
end while
Return η_{lo}
End Procedure

which is a continuous and monotone non-increasing function. Take $\eta \to \infty$ and let

$$\lim_{\eta\to\infty} g(\eta) = c,$$

where the limit exists, but may be $-\infty$. Indeed, it must hold that $c < \infty$ since,

$$\lim_{\eta\to\infty} g(\eta) < g(\eta') < \infty.$$

If $c < 0$, then taking η'' large enough that $g(\eta'') \leq 0$ suffices. Alternatively, if $c \geq 0$, then there exists $\tilde{\eta}$ such that $g(\eta) \leq c + \epsilon$ for every $\eta \geq \tilde{\eta}$. Choosing $\eta'' > \max\{\tilde{\eta}, c\} + \epsilon$ yields

$$h(\eta'') = \eta'' - g(\eta'') > c + \epsilon - c - \epsilon = 0.$$

This completes the proof. □

Theorem 4.3. *Assume f is convex and L-smooth. Then Algorithm 1 with $\eta_0 > 0$ requires at most $2K(\log\log(2\eta_0/L) \vee 1)$ iterations of GD and in the last run it outputs a step-size η and point $\overline{x}_K = \frac{1}{K}\sum_{i=0}^{K-1} x_i(\eta)$ such that exactly one of the following holds:*

$$\textit{Case 1:} \quad \eta = \eta_0 \quad \textit{and} \quad f(\overline{x}_K) - f(x^*) \leq \frac{\|x_0 - x^*\|_2^2}{2K\eta_0}$$

$$\textit{Case 2:} \quad \eta \neq \eta_0 \quad \textit{and} \quad f(\overline{x}_K) - f^* \leq \frac{\|x_0 - x^*\|_2^2}{2K}\left[\frac{\sum_{i=0}^k M_i\|\nabla f(x_i')\|_2^2}{\sum_{i=0}^k \|\nabla f(x_i')\|_2^2}\right],$$

where $M_i \overset{def}{=} M(x_i', x_{i+1}')$ and x_i' are the iterates generated by GD with step-size $\eta' \in [\eta, 2\eta]$.

Proof of Theorem 4.3. This analysis follows (Carmon and Hinder, 2022). First, instantiate Equation (16) from Proposition 3.3 with $\eta_i = \eta$ for all i to obtain

$$f(\overline{x}_k) - f^* \leq \frac{\|x_0 - x^*\|^2}{2\eta k} + \frac{\eta\left[\eta\sum_{i=0}^k M(x_i, x_{i+1})\|\nabla f(x_i)\|^2 - \sum_{i=0}^k \|\nabla f(x_i)\|^2\right]}{2k}. \quad (45)$$

Now, observe that if we get a "Lucky strike" and $\phi(\eta_{\text{hi}}) = \phi(\eta_0) \leq 0$, then specializing Equation (45) for $\eta = \eta_0$ we get

$$\begin{aligned} f(\overline{x}_k) - f(x^*) &\leq \frac{\|x_0 - x^*\|_2^2}{2\eta_0 k} + \frac{\eta_0}{2k}\left[\eta_0 \sum_{i=0}^k M(x_i, x_{i+1})\|\nabla f(x_i)\|_2^2 - \sum_{i=0}^k \|\nabla f(x_i)\|_2^2\right] \\ &= \frac{\|x_0 - x^*\|_2^2}{2\eta_0 k} + \frac{\eta_0\sum_{i=0}^k M(x_i, x_{i+1})\|\nabla f(x_i)\|_2^2}{2k} \cdot \phi(\eta_0) \\ &\leq \frac{\|x_0 - x^*\|_2^2}{2\eta_0 k}. \end{aligned}$$

This covers the first case of Theorem 4.3.

With the first case out of the way, we may assume that $\phi(\eta_{\text{hi}}) > 0$. This implies that $\eta_{\text{hi}} > \frac{1}{L}$, since if $\eta \leq \frac{1}{L}$ we have $\phi(\eta) \leq 0$. Now observe that when $\eta_{\text{lo}} = 2^{2^{-k}}\eta_0 \leq \frac{1}{L}$, we have that $\phi(\eta_{\text{lo}}) \leq 0$, therefore it takes at most $k = \lceil \log\log \frac{\eta_0}{L^{-1}} \rceil$ to find such an η_{lo}. From here on, we suppose that $\phi(\eta_{\text{hi}}) > 0$ and $\phi(\eta_{\text{lo}}) \leq 0$. Now observe that the algorithm's main loop always maintains the invariant $\phi(\eta_{\text{hi}}) > 0$ and $\phi(\eta_{\text{lo}}) \leq 0$, and every iteration of the loop halves $\log \frac{\eta_{\text{hi}}}{\eta_{\text{lo}}}$, therefore we make at most $\lceil \log\log \eta_0 L \rceil$ loop iterations. The output step-size η_{lo} satisfies $\frac{\eta_{\text{hi}}}{2} \leq \eta_{\text{lo}} \leq \eta_{hi}$ and $\phi(\eta_{\text{lo}}) \leq 0$. Specializing Equation (45) for $\eta = \eta_0$ and using that $\phi(\eta_{\text{lo}}) \leq 0$ we get

$$\begin{aligned} f(\overline{x}_k) - f(x^*) &\leq \frac{\|x_0 - x^*\|_2^2}{2\eta_{\text{lo}} k} + \frac{\eta_{\text{lo}}\sum_{i=0}^k M(x_i(\eta_{\text{lo}}), x_{i+1}(\eta_{\text{lo}}))\|\nabla f(x_i(\eta_{\text{lo}}))\|_2^2}{2k} \cdot \phi(\eta_{\text{lo}}) \\ &\leq \frac{\|x_0 - x^*\|_2^2}{2\eta_{\text{lo}} k}. \end{aligned} \quad (46)$$

By the loop invariant $\phi(\eta_{\text{hi}}) > 0$ we have

$$\phi(\eta_{\text{hi}}) > 0 \Leftrightarrow \eta_{\text{hi}} > \frac{\sum_{i=0}^K \|\nabla f(x_i(\eta_{\text{hi}}))\|_2^2}{\sum_{i=0}^K \|\nabla f(x_i(\eta_{\text{hi}}))\|_2^2 M(x_i(\eta_{\text{hi}}), x_{i+1}(\eta_{\text{hi}}))}$$

By the loop termination condition we have $\eta_{\text{lo}} \geq \frac{\eta_{\text{hi}}}{2}$, combining this with the last equation we get

$$\eta_{\text{lo}} \geq \frac{\eta_{\text{hi}}}{2} \geq \frac{1}{2}\frac{\sum_{i=0}^K \|\nabla f(x_i(\eta_{\text{hi}}))\|_2^2}{\sum_{i=0}^K \|\nabla f(x_i(\eta_{\text{hi}}))\|_2^2 M(x_i(\eta_{\text{hi}}), x_{i+1}(\eta_{\text{hi}}))}.$$

Plugging this into Equation (46) we obtain

$$f(\overline{x}_k) - f(x^*) \leq \frac{\|x_0 - x^*\|_2^2}{k} \cdot \frac{\sum_{i=0}^K \|\nabla f(x_i(\eta_{\text{hi}}))\|_2^2 M(x_i(\eta_{\text{hi}}), x_{i+1}(\eta_{\text{hi}}))}{\sum_{i=0}^K \|\nabla f(x_i(\eta_{\text{hi}}))\|_2^2}$$

It remains to notice that $\eta_{\text{hi}} \in [\eta_{\text{lo}}, 2\eta_{\text{lo}}]$. □

Theorem 4.4. *Suppose that f is convex and differentiable and let M be any directional smoothness function for f. Let $\Delta_0 := \|x_0 - x^*\|_2^2$. Then GD with the Polyak step-size and $\gamma \in (1, 2)$ satisfies*

$$f(\overline{x}_k) - f(x^*) \leq \frac{c(\gamma)\Delta_0}{2\sum_{i=0}^{k-1} M(x_i, x_{i+1})^{-1}}, \quad (27)$$

where $c(\gamma) = \gamma/(2-\gamma)(\gamma-1)$ and $\overline{x}_k = \sum_{i=0}^{k-1}\left[M(x_i, x_{i+1})^{-1}x_i\right] / \left(\sum_{i=0}^{k-1} M(x_i, x_{i+1})^{-1}\right)$.

For the proof of this theorem, we will need the following proposition:

https://doi.org/10.52202/079017-0473

Proposition D.1. *Let $x \in \mathbb{R}^d$. Define $\eta_x = \gamma \frac{f(x)-f(x^*)}{\|\nabla f(x)\|^2}$ for some $\gamma \in (1,2)$ and let $\tilde{x} = x - \eta_x \nabla f(x)$. Then,*

$$f(x) - f(x^*) \geq \frac{\gamma - 1}{\gamma^2} \frac{2}{M(x,\tilde{x})} \|\nabla f(x)\|_2^2.$$

Proof. Observe

$$\begin{aligned} f(x) - f(x^*) &= f(x) - f(\tilde{x}) + f(\tilde{x}) - f(x^*) \\ &\geq f(x) - f(\tilde{x}). \end{aligned} \tag{47}$$

By smoothness we have

$$\begin{aligned} f(\tilde{x}) &\leq f(x) + \langle \nabla f(x), \tilde{x} - x \rangle + \frac{M(x,\tilde{x})}{2} \|\tilde{x} - x\|^2 \\ &= f(x) - \eta_x \|\nabla f(x)\|_2^2 + \frac{\eta_x^2 M(x,\tilde{x})}{2} \|\nabla f(x)\|_2^2. \end{aligned}$$

Plugging back into Equation (47) we get

$$f(x) - f(x^*) \geq \eta_x \|\nabla f(x)\|_2^2 - \frac{\eta_x^2 M(x,\tilde{x})}{2} \|\nabla f(x)\|_2^2.$$

Let us now use the definition of $\eta_x = \gamma \frac{f(x)-f(x^*)}{\|\nabla f(x)\|_2^2}$ to get

$$f(x) - f(x^*) \geq \gamma (f(x) - f(x^*)) - \frac{\gamma \eta_x M(x,\tilde{x})}{2} (f(x) - f(x^*)).$$

Assuming that $f(x) \neq f(x^*)$ then we get by cancellation

$$1 \geq \gamma - \frac{\gamma \eta_x M(x,\tilde{x})}{2}.$$

Using the definition of η_x again

$$1 - \gamma \geq -\gamma^2 \frac{M(x,\tilde{x})}{2} \frac{f(x) - f(x^*)}{\|\nabla f(x)\|_2^2}$$

Rearranging we get

$$f(x) - f(x^*) \geq \frac{\gamma - 1}{\gamma^2} \frac{2}{M(x,\tilde{x})} \|\nabla f(x)\|_2^2.$$

If $f(x) = f(x^*)$ then $\|\nabla f(x)\|_2^2 = 0$, both sides are identically zero and the statement holds trivially. □

Now we can prove our theorem on the convergence of GD with Polyak step-sizes:

Proof of Theorem 4.4. We start by considering the distance to the optimum and expanding the square

$$\begin{aligned} \|x_{k+1} - x^*\|_2^2 &= \|x_k - x^*\|_2^2 + 2 \langle x_{k+1} - x_k, x_k - x^* \rangle + \|x_{k+1} - x_k\|_2^2 \\ &= \|x_k - x^*\|_2^2 - 2\eta_k \langle \nabla f(x_k), x_k - x^* \rangle + \eta_k^2 \|\nabla f(x_k)\|_2^2. \end{aligned} \tag{48}$$

Let $\delta_k = f(x_k) - f(x^*)$. By convexity we have $f(x^*) \geq f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle$. Therefore we can upper bound Equation (48) as

$$\begin{aligned} \|x_{k+1} - x^*\|_2^2 &\leq \|x_k - x^*\|_2^2 - 2\eta_k \delta_k + \eta_k^2 \|\nabla f(x_k)\|_2^2 \\ &= \|x_k - x^*\|_2^2 - 2\eta_k \delta_k + \eta_k \left(\gamma \frac{\delta_k}{\|\nabla f(x_k)\|_2^2} \right) \|\nabla f(x_k)\|_2^2 \\ &= \|x_k - x^*\|_2^2 - (2 - \gamma) \eta_k \delta_k, \end{aligned} \tag{49}$$

where in the second line we used the definition of η_k. By Proposition D.1 we have

$$\delta_k \geq \frac{\gamma-1}{\gamma}\frac{2}{M(x_k,x_{k+1})}\|\nabla f(x_k)\|_2^2. \tag{50}$$

Using this in Equation (49) gives

$$\begin{aligned}
\|x_{k+1}-x^*\|_2^2 &\leq \|x_k-x^*\|_2^2 - (2-\gamma)\eta_k\frac{\gamma-1}{\gamma^2}\frac{2}{M(x_k,x_{k+1})}\|\nabla f(x_k)\|_2^2 \\
&= \|x_k-x^*\|_2^2 - (2-\gamma)\frac{\gamma-1}{\gamma^2}\frac{2}{M(x_k,x_{k+1})}\left(\gamma\frac{\delta_k}{\|\nabla f(x_k)\|_2^2}\right)\|\nabla f(x_k)\|_2^2 \\
&= \|x_k-x^*\|_2^2 - \frac{2(2-\gamma)(\gamma-1)}{\gamma M(x_k,x_{k+1})}\delta_k.
\end{aligned}$$

Rearranging we get

$$\frac{2(2-\gamma)(\gamma-1)}{\gamma M(x_k,x_{k+1})}\delta_k \leq \|x_k-x^*\|_2^2 - \|x_{k+1}-x^*\|_2^2.$$

Summing up and telescoping we get

$$\sum_{i=0}^{k-1}\frac{2(2-\gamma)(\gamma-1)}{\gamma M(x_i,x_{i+1})}\delta_i \leq \|x_0-x^*\|_2^2.$$

Let $\overline{x}_k = \frac{1}{\sum_{i=0}^{k-1} M(x_i,x_{i+1})^{-1}}\sum_{i=0}^{k-1} M(x_i,x_{i+1})^{-1}x_i$, then by the convexity of f and Jensen's inequality we have

$$\begin{aligned}
f(\overline{x}_k)-f(x^*) &\leq \frac{1}{\sum_{i=0}^{k-1} M(x_i,x_{i+1})^{-1}}\sum_{i=0}^{k-1} M(x_i,x_{i+1})^{-1}\delta_i \\
&\leq \frac{\gamma}{2(2-\gamma)(\gamma-1)}\frac{1}{\sum_{i=0}^{k-1} M(x_i,x_{i+1})^{-1}}\|x_0-x^*\|_2^2.
\end{aligned}$$

□

Theorem D.2. *If f is convex and differentiable, then GD with the Polyak step-size and $\gamma < 2$ satisfies,*

$$f(\overline{x}_k)-f(x^*) \leq \frac{1}{(2-\gamma)\sum_{i=0}^{k}\eta_i}\|x_0-x^*\|_2^2, \tag{51}$$

where $\overline{x}_k = \sum_{i=0}^{k-1}\eta_i x_i / \left(\sum_{i=0}^{k-1}\eta_i\right)$.

Proof. The proof begins in the same manner as that for Theorem 4.4,

$$\begin{aligned}
\|x_{k+1}-x^*\|_2^2 &= \|x_k-x^*\|_2^2 + 2\langle x_{k+1}-x_k, x_k-x^*\rangle + \|x_{k+1}-x_k\|_2^2 \\
&= \|x_k-x^*\|_2^2 - 2\eta_k\langle\nabla f(x_k), x_k-x^*\rangle + \eta_k^2\|\nabla f(x_k)\|_2^2 \\
&\leq \|x_k-x^*\|_2^2 - 2\eta_k\delta_k + \eta_k^2\|\nabla f(x_k)\|_2^2 \\
&= \|x_k-x^*\|_2^2 - 2\eta_k\delta_k + \eta_k\left(\gamma\frac{\delta_k}{\|\nabla f(x_k)\|_2^2}\right)\|\nabla f(x_k)\|_2^2 \\
&= \|x_k-x^*\|_2^2 - (2-\gamma)\eta_k\delta_k.
\end{aligned}$$

Re-arranging, summing from $i=0$ to $k-1$, and dividing by $\sum_{i=0}^{k-1}\eta_i$,

$$\begin{aligned}
\implies \sum_{i=0}^{k-1}\frac{\eta_i}{\sum_{i=0}^{k}\eta_i}(f(x_i)-f(w^*)) &\leq \frac{1}{(2-\gamma)\sum_{i=0}^{k}\eta_i}\left[\|x_0-x^*\|_2^2 - \|x_k-x^*\|_2^2\right] \\
\implies f(\overline{x}_k)-f(x^*) &\leq \frac{1}{(2-\gamma)\sum_{i=0}^{k}\eta_i}\|x_0-x^*\|_2^2,
\end{aligned}$$

which completes the proof. □

 https://doi.org/10.52202/079017-0473

Lemma D.3. *Normalized GD with step-sizes η_k satisfies*

$$-\frac{\eta_k}{\|\nabla f(x_k)\|_2}\langle \nabla f(x_k), \nabla f(x_{k+1})\rangle \leq \eta_k^2 M(x_k, x_{k+1}) - \eta_k \|\nabla f(x_k)\|_2. \tag{52}$$

Proof. By convexity we have

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle x_{k+1} - x_k, \nabla f(x_{k+1})\rangle \\ &= f(x_k) - \frac{\eta_k}{\|\nabla f(x_k)\|_2}\langle \nabla f(x_k), \nabla f(x_{k+1})\rangle \end{aligned} \tag{53}$$

Now note that

$$\begin{aligned} -\frac{\eta_k}{\|\nabla f(x_k)\|_2}\langle \nabla f(x_k), \nabla f(x_{k+1})\rangle &= \frac{\eta_k}{\|\nabla f(x_k)\|_2}\langle \nabla f(x_k), \nabla f(x_k) - \nabla f(x_{k+1})\rangle \\ &\quad - \eta_k \|\nabla f(x_k)\|_2 \end{aligned} \tag{54}$$

$$\leq \eta_k \|\nabla f(x_k) - \nabla f(x_{k+1})\| - \eta_k \|\nabla f(x_k)\|_2, \tag{55}$$

where we used Cauchy-Schwarz. Recalling the definition of directional smoothness

$$M(x_k, x_{k+1}) \stackrel{\text{def}}{=} \frac{\|\nabla f(x_k) - \nabla f(x_{k+1})\|}{\|x_k - x_{k+1}\|} = \frac{\|\nabla f(x_k) - \nabla f(x_{k+1})\|}{\eta_k}$$

in Equation (55) gives

$$-\frac{\eta_k}{\|\nabla f(x_k)\|_2}\langle \nabla f(x_k), \nabla f(x_{k+1})\rangle \leq \eta_k^2 M(x_k, x_{k+1}) - \eta_k \|\nabla f(x_k)\|_2.$$

□

Theorem 4.5. *Suppose that f is convex and differentiable. Let D be the point-wise directional smoothness defined by Equation* (4) *and $\Delta_0 := \|x_0 - x^*\|_2^2$. Then normalized GD with a sequence of non-increasing step-sizes η_k satisfies*

$$f(\hat{x}_k) - f(x^*) \leq \frac{\Delta_0 + \sum_{i=0}^{k-1}\eta_i^2}{2k^2}\left(\frac{f(x_0)}{\eta_0^2} - \frac{f(x^*)}{\eta_{k-1}^2}\right) + \frac{\Delta_0 + \sum_{i=0}^{k-1}\eta_i^2}{2k}\sum_{i=0}^{k-1}\frac{M(x_i, x_{i+1})}{k}, \tag{30}$$

where $\hat{x}_k = \arg\min_{i\in[k-1]} f(x_i)$. If $\max_{i\in[k-1]} M(x_i, x_{i+1})$ is bounded for all k (i.e. f is L-smooth), then for $\eta_i = 1/\sqrt{i}$ we have $f(\hat{x}_k) - f(x^*) \in \mathcal{O}(1/k)$ and for $\eta_i = 1/\sqrt{i}$ we get the anytime result $f(\hat{x}_k) - f(x^*) \in \mathcal{O}(\log(k)/k)$.

Proof. Here we will first establish that for any non-increasing sequence of step-sizes $\eta_k > 0$ we have that

$$\min_{i\in[k-1]} f(x_i) - f(x^*) \leq \frac{1}{2}\frac{\Delta_0 + \sum_{i=0}^{k-1}\eta_i^2}{k}\left(\frac{f(x_0)}{k\eta_0^2} - \frac{f(x^*)}{k\eta_{k-1}^2} + \sum_{i=0}^{k-1}\frac{M(x_i, x_{i+1})}{k}\right). \tag{56}$$

The specialized results follow by assuming that $\sum_{i=0}^{k-1}\frac{M(x_i,x_{i+1})}{k}$ is bounded, which it is the case of L–Lipschitz gradients. In particular the $\min_{i\in[k-1]} f(x_i) - f(x^*) \in \mathcal{O}(1/T)$ result follows by plugging in $\eta_i = 1/\sqrt{k}$ and using that

$$\begin{aligned} \sum_{i=0}^{k-1}\eta_i^2 &= \sum_{i=0}^{k-1}\frac{1}{k} = 1 \\ \frac{f(x_0)}{k\eta_0^2} - \frac{f(x^*)}{k\eta_{k-1}^2} &= f(x_0) - f(x^*). \end{aligned}$$

Alternatively we get $\min_{i\in[k-1]} f(x_i) - f(x^*) \in \mathcal{O}(\log(T)/T)$ by plugging in $\eta_i = 1/\sqrt{i+1}$ and using that

$$\begin{aligned} \sum_{i=0}^{k-1}\eta_i^2 &= \sum_{i=0}^{k-1}\frac{1}{i+1} \leq \log(k) \\ \frac{f(x_0)}{k\eta_0^2} - \frac{f(x^*)}{k\eta_{k-1}^2} &= \frac{f(x_0)}{k} - f(x^*). \end{aligned}$$

With this in mind, let us prove Equation (56).

By convexity,

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \frac{\eta_k}{\|\nabla f(x_k)\|_2} \nabla f(x_k)^\top \nabla f(x_{k+1}) \\ &\leq f(x_k) - \eta_k \|\nabla f(x_k)\|_2 + \eta_k^2 M(x_k, x_{k+1}). \qquad \text{(Using (52))} \end{aligned}$$

Re-arranging, dividing through by η_k^2, and then summing over $i = 0, \cdots, k-1$ gives

$$\begin{aligned} \sum_{i=0}^{k-1} \frac{\|\nabla f(x_i)\|_2}{\eta_i} &\leq \frac{f(x_0)}{\eta_0^2} + \sum_{i=1}^{k-2} f(x_i) \left(\frac{1}{\eta_i^2} - \frac{1}{\eta_{i-1}^2} \right) - \frac{f(x^*)}{\eta_{k-1}^2} + \sum_{i=0}^{k-1} M(x_i, x_{i+1}) \\ &\leq \frac{f(x_0)}{\eta_0^2} - \frac{f(x^*)}{\eta_{k-1}^2} + \sum_{i=0}^{k-1} M(x_i, x_{i+1}), \end{aligned} \tag{57}$$

where we used that $\eta_{i-1} \leq \eta_i \implies \frac{1}{\eta_i^2} - \frac{1}{\eta_{i-1}^2} \leq 0$. Using Jensen's inequality over the map $a \mapsto 1/a$, which is convex for a positive, gives

$$\sum_{i=0}^{k-1} \frac{\eta_i}{\|\nabla f(x_k)\|_2} \geq \frac{k^2}{\sum_{i=0}^{k-1} \|\nabla f(x_k)\|_2 / \eta_i} \overset{(57)}{\geq} \frac{k^2}{\frac{f(x_0)}{\eta_0^2} - \frac{f(x^*)}{\eta_{k-1}^2} + \sum_{i=0}^{k-1} M(x_i, x_{i+1})}. \tag{58}$$

Meanwhile, recall our notation $\Delta_i = \|x_i - x^*\|_2^2$. Expanding the squares and using that $f(x)$ is convex, we have that

$$\begin{aligned} \Delta_{i+1} &= \Delta_i - 2 \frac{\eta_i}{\|\nabla f(x_i)\|_2} \nabla f(x_i)^\top (x_i - x^*) + \eta_i^2 \\ &\leq \Delta_i - 2\eta_i \frac{f(x_i) - f(x^*)}{\|\nabla f(x_k)\|_2} + \eta_i^2. \end{aligned}$$

As before, we use $\delta_i := f(x_i) - f(x^*)$. Re-arranging, summing both sides of the above over $i = 0, \ldots, k-1$ and using telescopic cancellation gives

$$\sum_{i=0}^{k-1} \eta_i \frac{\delta_i}{\|\nabla f(x_i)\|} \leq \frac{\Delta_0 + \sum_{i=0}^{t-1} \eta_i^2}{2}.$$

Using the above along with (58) gives,

$$\begin{aligned} \min_{i \in [k-1]} \delta_i &\leq \frac{1}{\sum_{i=0}^{k-1} \frac{\eta_i}{\|\nabla f(x_i)\|_2}} \sum_{i=0}^{k-1} \eta_i \frac{\delta_i}{\|\nabla f(x_i)\|_2} \\ &\leq \frac{1}{2} \frac{\Delta_0 + \sum_{i=0}^{k-1} \eta_i^2}{\sum_{i=0}^{k-1} \frac{\eta_i}{\|\nabla f(x_i)\|}} \\ &\leq \frac{1}{2} \frac{\Delta_0 + \sum_{i=0}^{k-1} \eta_i^2}{k} \left(\frac{f(x_0)}{k\eta_0^2} - \frac{f(x^*)}{k\eta_{k-1}^2} + \sum_{i=0}^{k-1} \frac{M(x_i, x_{i+1})}{k} \right) \end{aligned}$$

□

E Experimental Details

In this section we provide additional details necessary to reproduce our experiments. We run our logistic regression experiments using PyTorch (Paszke et al., 2019). For the UCI datasets, we use the pre-processed version of the data provided by Fernández-Delgado et al. (2014), although we do not use their evaluation procedure as it is known have test-set leakage. Instead, we randomly perform an 80–20 train-test split and use the test set for validation. Unless otherwise stated, all methods are initialized using the Kaiming initialization (He et al., 2015), which is standard in PyTorch.

https://doi.org/10.52202/079017-0473

In order to compute the strongly adapted step-sizes, we run the SciPy (Virtanen et al., 2020) implementation of Newton method on Equation (44). In general, we find this procedure is surprisingly robust, although it can be slow.

Figure 1: We pick two datasets from the UCI repository to showcase different behaviors of the upper-bounds. We compute a tight-upper bound on L as follows. Recall that for logistic regression problems the Hessian is given by

$$\nabla^2 f(x) = A^\top \text{Diag}\left(\frac{1}{\sigma(-y \cdot Ax) + 2 + \sigma(y \cdot Ax)}\right) A,$$

where A is the data matrix and $\sigma(z) = \frac{1}{1+\exp(z)}$ is the sigmoid function. A short calculation shows that the diagonal matrix

$$\text{Diag}\left(\frac{1}{\sigma(-y \cdot Ax) + 2 + \sigma(y \cdot Ax)}\right) \preceq \frac{1}{4}\mathbf{I},$$

which is tight when $x = 0$. As a result, $L = \lambda_{\max}(A^\top A)/4$. We compute this manually. We also compute the optimal value for the logistic regression problem using the SciPy implementation of BFGS (Liu and Nocedal, 1989). We use this value for $f(x^*)$ to compute the Polyak step-size and when plotting sub-optimality. It turns out that the upper-bound based on L-smoothness for both GD with the Polyak step-size (Hazan and Kakade, 2019) and standard GD (Bubeck et al., 2015) is

$$f(x_k) - f(x^*) \leq \frac{2L\|x_0 - x^*\|_2^2}{k}.$$

Figure 3: We run these experiments using vanilla NumPy. As mentioned in the text, we generate a quadratic optimization problem

$$\min_x \frac{1}{2} x^\top Ax - b^\top x,$$

where the eigenvalues of A were generated to follow power law distribution with parameter $\alpha = 3$. We scaled the eigenvalues to ensure $L = 1000$. The dimension of the problem we create is $d = 300$. We repeat the experiment for 20 random trials and plot the mean and standard deviations.

Figure 4: We pick three different datasets from the UCI repository to showcase the possible convergence behavior of the optimization methods. We compute L and $f(w^*)$ as described above for Figure 1. For normalized GD, we use the step-size schedule $\eta_k = \eta_0/\sqrt{k}$ as suggested by our theory. To pick η_0, we run a grid search on the grid generated by `np.logspace(-8, 1, 20)`. We implement AdGD from scratch and use a starting step-size of $\eta_0 = 10^{-3}$. We use the same procedure to compute the strongly adapted step-sizes as described above.

F Computational Details

The experiments in Figure 3 were run on a MacBook Pro (16 inch, 2019) with a 2.6 GHz 6-Core Intel i7 CPU and 16GB of memory. All other experiments were run on a Slurm cluster with several different node configurations. Our experiments on the cluster were run with nodes using (i) Nvidia A100 GPUs (80GB or 40GB memory) or Nvidia H100-80GB GPUs with Icelake CPUs, or (ii) Nvidia V100-32GB or V100-16GB GPUs with Skylake CPUs. All jobs were allocated a single GPU and 24GB of RAM.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims in the abstract and introduction are justified with rigorous proofs and supported by experiments.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of our theoretical results, including necessary assumptions, are addressed in the text.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

https://doi.org/10.52202/079017-0473

Answer: [Yes]

Justification: Our theoretical results are accompanied by their necessary assumptions and rigorous proofs are provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All experimental procedures and hyper-parameter settings are provided in Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will release the code to reproduce our experiments upon acceptance of the paper. All non-synthetic data used in this paper is open source and freely accessible online.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The details necessary to interpret our experimental results are provided in Section 5, while additional details necessary to reproduce the experiments are given in Appendix E.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: Our experiments consider only deterministic optimization methods.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

 https://doi.org/10.52202/079017-0473

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We detail the compute resources used in our paper in Appendix F.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All authors are familiar with the code of ethics and conform to its principles. Moreover, our research is primarily theoretical and is of minor ethical concern.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our research is primarily theoretical and has no societal impact beyond the impact of general advances in the fields of machine learning and optimization.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No new data or models are released as part of this paper.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All libraries, models, and data sources are appropriately referenced.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

https://doi.org/10.52202/079017-0473

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets are introduced.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We did not use any crowdsourcing for this work

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: IRB approval was not required for this paper.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-0473