OPT2023: 15th Annual Workshop on Optimization for Machine Learning

A novel analysis of gradient descent under directional smoothness

Aaron Mishkin[*] AARONPMISHKIN@CS.STANFORD.EDU
Department of Computer Science, Stanford University

Ahmed Khaled[*] AHMED.KHALED@PRINCETON.EDU
Department of Electrical and Computer Engineering, Princeton University

Aaron Defazio AARON.DEFAZIO@GMAIL.COM
Fundamental AI Research team, Meta

Robert M. Gower GOWERROBERT@GMAIL.COM
CCM, Flatiron Institute

Abstract

We develop new sub-optimality bounds for gradient descent 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 an implicit equation to obtain an adapted step-size; we show that this equation is straightforward to solve for convex quadratics and leads to new guarantees for a classical step-size sequence. For general functions, we prove that exponential search can be used to obtain a path-dependent convergence guarantee with only a log-log dependency on the global smoothness constant. Experiments on quadratic functions showcase the utility of our theory and connections to the edge-of-stability phenominon.

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 gradient methods can still shrink the optimality gap when f is not L-smooth, particularly for deep neural networks (Bengio, 2012; J. Cohen et al., 2021; Li et al., 2020). Even for functions verifying smoothness, convergence rates are often pessimistic and fail to predict optimization speed in practice (Paquette et al., 2023).

In this paper, we prove new sub-optimality bounds for gradient descent without global smoothness assumptions by deriving upper-bounds of the form,

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

where the *directional smoothness* $M(x_{k+1}, x_k)$ depends only on properties of f along the chord between x_{k+1} and x_k. Our bounds provide a path-dependent perspective on gradient descent and

[*] Equal contribution

are tighter than conventional analyses when the step-size sequence is adapted to the directional smoothness, meaning $\eta_k < 2/M(x_{k+1}, x_k)$. As adapted step-sizes require solving an implicit equation, we show the exponential search from Carmon and Hinder (2022) can be used to obtain similar path-dependent complexities up to a log-log penalty. Our contributions are the following:

Directional smoothness constants. We introduce two related directional smoothness constants $M(y, x)$; one depends only on the end-points y, x and is easily computed, while the other yields a tighter bound but depends on the chord $\mathcal{C} = \{\alpha x + (1-\alpha)y : \alpha \in [0, 1]\}$.

Convergence rates. We leverage directional smoothness to prove new convergence rates for gradient descent under a directional strong convexity assumption as well as without a curvature condition. Our bounds are step-size independent and improve over the standard analyses.

Quadratic case. For quadratic objectives, we show solving the implicit equation $\eta_k = \frac{1}{M(x_{k+1}, x_k)}$ is feasible and results in an adaptive step-size that requires no knowledge of problem constants and performs as well as, or better than, gradient descent with the optimal fixed step-size $\frac{1}{L}$.

Exponential search. We give a simple restarting mechanism which comes within a double logarithm of the complexity obtained using adapted step-sizes that depend on the directional smoothness.

1.1. Related work

Local smoothness: Global smoothness of f can be avoided by using local Lipschitz continuity of the gradient ("local smoothness"). Such analyses often require the iterates to be bounded so that local smoothness gives a quadratic bound like Eq. (1). Zhang and Hong (2020) enforce boundedness by breaking optimization into stages, while Patel and Berahas (2022) develop a framework using stopping times and Lu and Mei (2023) use line-search and a modified update. Finally, Park et al. (2021) leverage the local smoothness constants along the optimization path to ensure convergence.

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

2. Directional smoothness

A common view of GD for L-smooth functions is that it minimizes the quadratic upper-bound,

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

However, this viewpoint gives rates which depend on the global, worst-case growth of f. This is both counter-intuitive and undesirable: the iterates of gradient descent depend only on local properties of f, so the analysis should show the conditioning GD "sees" on the optimization path $\{x_{k+1} := x_k - \eta_k \nabla f(x_k)\}$. Towards this goal, assume f is both differentiable and absolutely continuous and define the point-wise directional smoothness as,

$$D(y, x) := \frac{2\|\nabla f(y) - \nabla f(x)\|_2}{\|y - x\|_2}. \tag{3}$$

Point-wise smoothness is a local estimate of L and satisfies $D(y,x) \leq 2L$. However, when f is convex and C^2, $D(y,x)$ also gives a surprising alternative to the smoothness upper-bound.

Lemma 1 (Point-wise Directional Smoothness) *If f is convex and twice-differentiable, then*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{D(x,y)}{2}\|y - x\|_2^2. \tag{4}$$

Eq. (4) is purely local. However, it is weaker than the standard quadratic upper-bound by a factor of two and requires $f \in C^2$. As an alternative, we define the path-wise directional smoothness,

$$A(x,y) := \sup_{t \in [0,1]} \frac{\langle \nabla f(x + t(y-x)) - \nabla f(x), y - x \rangle}{t\|x - y\|^2}, \tag{5}$$

and show that it exactly verifies the quadratic upper-bound with no additional assumptions.

Lemma 2 (Path-wise Directional Smoothness) *The path directional smoothness satisfies*

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

Path smoothness is tighter than point-wise smoothness since $A(y,x) \leq L$, but not easily computed because it depends on the chord between x and y. We refer to both concepts (Eq. (3) and (5)) as directional smoothness and use the notation $M(y,x)$ to stand-in for either quantity. Substituting the GD update into directional smoothness gives a descent lemma which reflects only local geometry,

$$f(x_{k+1}) \leq f(x_k) - \eta_k \left(1 - \frac{\eta_k M(x_{k+1}, x_k)}{2}\right) \|\nabla f(x_k)\|_2^2. \tag{7}$$

See Lemma 6 for proof. If $\eta_k < 2/M(x_{k+1}, x_k)$, then GD is guaranteed to decrease f and we call η_k *adapted* to the directional smoothness. However, finding a sequence of adapted step-sizes is not straightforward. For instance, computing the standard $1/L$ analogue requires solving the non-linear equation $\eta_k = 1/M(x_{k+1}(\eta_k), x_k)$ which is non-trivial. We tackle this problem in Section 5.

Relation to Smoothness: As mentioned, if f is L-smooth, then $M(y,x)$ is globally bounded with $D(y,x) \leq 2L$ and $A(y,x) \leq L$. This second inequality follows from one application of Cauchy-Schwarz and mirrors the standard proof of the smoothness upper-bound (see Nesterov et al. (2018, Theorem 2.1.5)). However, since L is a global quantity, the directional smoothness is often much smaller than predicted by these bounds (Malitsky and Mishchenko, 2020).

In order for $M(y,x)$ to be well-defined, we only need the weaker assumption that f is locally smooth. A function is locally smooth if for every compact set $\mathcal{S}$ there exists $L_{\mathcal{S}} \geq 0$ such that f is $L_{\mathcal{S}}$-smooth on $\mathcal{S}$. Although our bounds on the directional smoothness also hold with $L_{\mathcal{S}}$, we expect $D(y,x)$ and $A(y,x)$ to smaller in general because they only depend on a small subset of $\mathcal{S}$.

3. Path-Dependent convergence rates

Now we leverage directional smoothness to derive new guarantees for gradient descent. We emphasize that the following results are *sub-optimality bounds*, rather than convergence rates; a sequence

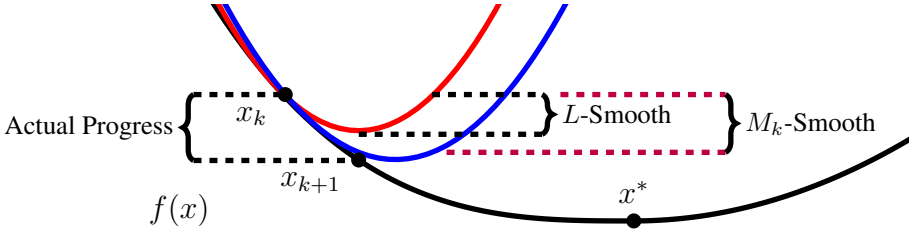


Figure 1: Illustration of GD with step-size $\eta_k = 1/L$. Even though the step-size exactly minimizes the upper-bound from L-smoothness, M_i directional smoothness better predicts the progress of the actual gradient step. Our rates improve on L-smoothness because they use this tighter bound.

of adapted step-sizes is required to convert our propositions into full a convergence theory. As a trade-off, we obtain bounds reflecting 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 following directional strong convexity constant:

$$\mu(y, x) = \inf_{t \in [0,1]} \frac{\langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle}{t\|x - y\|^2}. \tag{8}$$

If f is convex, then $\mu(y, x)$ verifies the standard lower-bound from strong convexity. Moreover, $\mu(y, x) \geq \mu$ when f is strongly convex (Lemma 7). We prove two bounds for convex functions using directional strong convexity. For simplicity, let $M_i := M(x_{i+1}, x_i)$ and $\mu_i := \mu_i(x_{i+1}, x_i)$.

Proposition 3 *If f is convex, then GD with step-size sequence $\{\eta_k\}$ satisfies,*

$$\delta_k \leq \left[\prod_{i \in \mathcal{G}} (1 + \eta_i \lambda_i \mu_i)\right] \delta_0 + \sum_{i \in \mathcal{B}} \left[\prod_{j > i, j \in \mathcal{G}} (1 + \eta_j \lambda_j \mu_i)\right] \frac{\eta_i \lambda_i}{2} \|\nabla f(x_i)\|_2^2 \tag{9}$$

where $\lambda_i = \eta_i M_i - 2$, $\mathcal{G} = \left\{i : \eta_i < \frac{2}{M(x_{i+1}, x_i)}\right\}$, $\mathcal{B} = [k] \setminus \mathcal{G}$, and $\delta_i = f(x_i) - f(x^*)$.

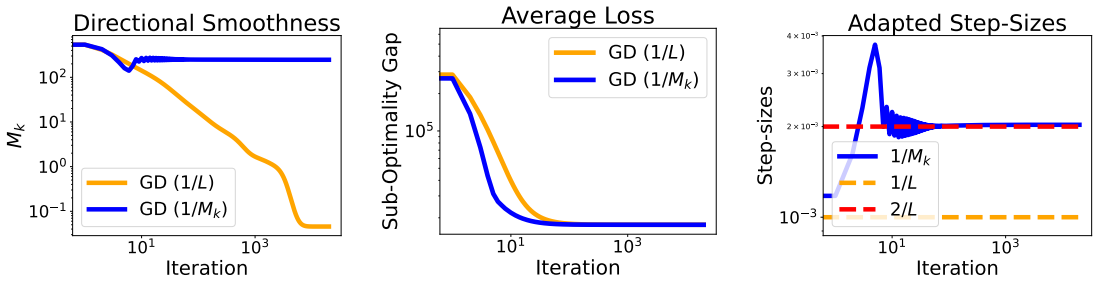
The analysis splits iterations into good steps where η_k is adapted to the directional smoothness, and bad steps $\mathcal{B}$ where the step-size is too large and GD may increase the optimality gap. In the case where f is L-smooth and μ-strongly convex, using the step-size sequence $\eta_k = 1/L$ gives,

$$f(x_{k+1}) - f(x^*) \leq \prod_{i=0}^{k} \left(1 - \frac{\mu_i \left(2 - M_i/L\right)}{L}\right) \left[f(x_0) - f(x^*)\right], \tag{10}$$

where $\mu_i \left(2 - M_i/L\right) \geq \mu$. Eq. (9) gives a tighter rate for GD under standard assumptions and step-sizes by localizing to the convergence path. We give a more elegant bound in Appendix B, which does not divide k into good steps and bad steps. In exchange, the requirement for η_k to be adapted to the directional smoothness changes to $\eta_k \leq 1/M(x_{k+1}, x_k)$. We conclude this section with a bound for when there is no lower curvature, meaning $\mu_i = 0$.

Proposition 4 *Let $\bar{x}_k = \sum_{i=0}^k \eta_i x_{i+1} / \sum_{i=0}^k \eta_i$. If f is convex, then GD satisfies,*

$$f(\bar{x}_k) - f(x^*) \leq \frac{\|x_0 - x^*\|_2^2 + \sum_{i=0}^k \eta_i^2 (\eta_i M_i - 1) \|\nabla f(x_i)\|_2^2}{2 \sum_{i=0}^k \eta_i}. \tag{11}$$



(a) (b) (c)

Figure 2: Results from linear regression with $L = 1000$ and Hessian skew for 20000 steps of gradient descent: (a) shows the directional smoothness over training trajectory, (b) shows the sub-optimality gap over training trajectory, and (c) shows the step-sizes over training trajectory.

This rate, which is at least as tight as the standard analysis, is key to our results in the next sections.

4. The quadratic case

In the past sections, we showed how to obtain tighter convergence rates for GD when using step-sizes adapted to the local smoothness. Now we show that selecting adapted step-sizes is straightforward when f is a convex quadratic. Suppose $f(x) = x^\top Bx/2 - c^\top x$, where B is a positive semi-definite matrix. Using Proposition 4 with step-sizes $\eta_i = 1/M(x_{i+1}(\eta_i), x_i)$ for each i yields

$$f(\bar{x}_k) - f_* \leq \frac{\|x_0 - x_*\|^2}{2\sum_{i=0}^k \eta_i} = \frac{\|x_0 - x_*\|^2}{2\sum_{i=0}^k \frac{1}{M(x_{i+1},x_i)}} \leq \frac{\|x_0 - x_*\|^2}{2(k+1)} \frac{\sum_{i=0}^k M(x_{i+1}, x_i)}{k+1}, \tag{12}$$

which depends solely on the average directional smoothness along the trajectory.

It is not clear how to solve $\eta_i = 1/M(x_{i+1}, x_i)$ in general, given that x_{i+1} is a function of η_i. However, if f is quadratic as above and M_i is the point-wise smoothness, then Lemma 9 shows $D(x_{i+1}, x_i) = 2\|B\nabla f(x_i)\|/\|\nabla f(x_i)\|$ and thus $\eta_i = 1/D(x_{i+1}, x_i)$ can be computed easily. This step-size was first suggested by Dai and Yang (2006), who show it approximates the Cauchy step-size and converges to $\frac{2}{L}$. Interestingly, this matches recent results on the edge-of-stability (Ahn et al., 2022; J. Cohen et al., 2021). To our knowledge, no prior non-asymptotic convergence rate exists for this step-size, meaning our work gives it new theoretical justification.

Figure 2 compares the performance of GD with adapted step-sizes and with a fixed step-size for a synthetic linear regression problem with Hessian skew (Pan et al., 2022). The results shown are averaged over ten different random initializations. We find that adapted step-sizes speed-up optimization by taking advantage of the directional smoothness, which drops significantly during optimization when using a fixed step-size. Interestingly, the adapted step-sizes can be much larger than even $2/L$, especially in the beginning of training when they oscillate around $2/L$. This reflects the theoretical connections to edge-of-stability (J. M. Cohen et al., 2022).

5. Exponential search

Our goal is to move beyond quadratics and take advantage of our tighter rates for general convex functions. We consider a fixed horizon k and denote by $x_i(\eta)$ the sequence of iterates obtained by gradient descent from initialization x_0 using a fixed step-size η. Define the adaptedness criterion

$$\psi(\eta) = \frac{\sum_{i=0}^{k} \|\nabla f(x_i(\eta))\|^2}{\sum_{i=0}^{k} M(x_{i+1}(\eta), x_i(\eta))\|\nabla f(x_i(\eta))\|^2}, \tag{13}$$

and suppose that η that satisfies $\frac{\psi(\eta)}{2} \leq \eta \leq \psi(\eta)$. Using these bounds in Proposition 4 yields

$$f(\bar{x}_k) - f_* \leq \frac{\|x_0 - x_*\|^2}{k} \left[\frac{\sum_{i=0}^{k} M(x_{i+1}, x_i)\|\nabla f(x_i)\|^2}{\sum_{i=0}^{k} \|\nabla f(x_i)\|^2} \right], \tag{14}$$

This is 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 bound.

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 bisection search. Although we adapt their technique to our setting, our approach differs from Carmon and Hinder (2022) in that they start with a small step-size and narrow upwards, while we start with a large step-size and narrow downwards. We give more details in Algorithm 1, which we prove obtains the following guarantee.

Theorem 5 *Assume the objective f is convex and L-smooth. Then Algorithm 1 with $\eta_0 > 0$ requires at most $2T(\log\log \frac{2\eta_0}{L^{-1}} \vee 1)$ iterations of gradient descent and in the last run it outputs a step-size η and point $\hat{x}_T = \frac{1}{T}\sum_{t=0}^{T-1} x_t(\eta)$ such that exactly one of the following holds:*

$$\textit{Case 1:} \quad \eta = \eta_0 \quad \textit{and} \quad f(\hat{x}_T) - f_* \leq \frac{\|x_0 - x_*\|^2}{2T\eta_0}$$

$$\textit{Case 2:} \quad \eta \neq \eta_0 \quad \textit{and} \quad f(\hat{x}_T) - f_* \leq \frac{\|x_0 - x_*\|^2}{2T} \left[\frac{\sum_{i=0}^{k} M(x'_{i+1}, x'_i)\|\nabla f(x'_i)\|^2}{\sum_{i=0}^{k} \|\nabla f(x'_i)\|^2} \right],$$

where $x'_1, x'_2, \ldots$ are the iterates generated by GD with step-size η' satisfying $\eta' \in [\eta, 2\eta]$.

The proof of Theorem 5 is provided in the appendix. Observe that we only get a $\log\log$ dependence on the global smoothness constant, while obtaining a convergence rate that scales with the weighted average of the directional smoothness constants along a very close trajectory.

6. Conclusion

We present new sub-optimality bounds for gradient descent under two novel measures of local gradient variation which we call directional smoothness. Our results hold for any sequence of step-sizes and improve over standard analyses when the step-size sequence is adapted to the directional smoothness. Although finding adapted step-sizes is challenging, we show that for convex quadratics the sequence $\eta_k = 1/M(x_{k+1}, x_k)$ can be computed with a single Hessian-vector product. We tackle the general case with an algorithm based on exponential search; our approach gives a weighted-version of the path-dependent convergence rate with no need for adapted step-sizes.

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.

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

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.

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

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.

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.

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*.

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.

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

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

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.

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*.

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.

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, 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.

Appendix A. Directional Smoothness: Proofs

Lemma 1 (Point-wise Directional Smoothness) *If f is convex and twice-differentiable, then*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{D(x,y)}{2}\|y - x\|_2^2. \tag{4}$$

Proof Taylor's theorem and the integral form of the remainder imply

$$\begin{aligned}
f(y) &= f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \left\langle y - x, \nabla^2 f(x + t(y-x))(y-x) \right\rangle (1-t) dt \\
&= f(x) + \langle \nabla f(x), y - x \rangle - \int_0^1 t \left\langle y - x, \nabla^2 f(x + t(y-x))(y-x) \right\rangle dt \\
&\qquad + \int_0^1 \left\langle y - x, \nabla^2 f(x + t(y-x))(y-x) \right\rangle dt \\
&\leq f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \left\langle y - x, \nabla^2 f(x + t(y-x))(y-x) \right\rangle dt,
\end{aligned}$$

where we have used the fact that $\left\langle y - x, \nabla^2 f(x + t(y-x))(y-x) \right\rangle \geq 0$ for all $t \in [0,1]$ by convexity of f. The fundamental theorem of calculus now implies

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

where the last two steps use Cauchy-Schwarz inequality and the definition of $D(x,y)$. ■

Lemma 2 (Path-wise Directional Smoothness) *The path directional smoothness satisfies*

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

Proof Starting again from the fundamental theorem of calculus,

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

which completes the proof. ■

Lemma 6 *One step of gradient descent with step-size $\eta_k > 0$ makes progress as*

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

Proof Starting from Eq. (3), we have

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

■

Appendix B. Path-Dependent Convergence Rates: Proofs

Lemma 7 *If f is convex, then for any $y, x \in \mathbb{R}^d$,*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu(y, x)}{2} \|y - x\|_2^2. \tag{15}$$

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

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 $\mu(y, x) = 0$. Now assume that f is μ strongly convex. As a standard consequence of strong-convexity, we obtain:

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

■

Proposition 3 *If f is convex, then GD with step-size sequence $\{\eta_k\}$ satisfies,*

$$\delta_k \leq \left[\prod_{i\in\mathcal{G}}(1+\eta_i\lambda_i\mu_i)\right]\delta_0 + \sum_{i\in\mathcal{B}}\left[\prod_{j>i,j\in\mathcal{G}}(1+\eta_j\lambda_j\mu_i)\right]\frac{\eta_i\lambda_i}{2}\|\nabla f(x_i)\|_2^2 \tag{9}$$

where $\lambda_i = \eta_i M_i - 2$, $\mathcal{G} = \left\{i : \eta_i < \frac{2}{M(x_{i+1},x_i)}\right\}$, $\mathcal{B} = [k]\setminus\mathcal{G}$, and $\delta_i = f(x_i) - f(x^*)$.

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 Eq. (7),

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

where we used that directional strong convexity gives

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

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

Proposition 8 *Let $\Delta_i = \|x_i - x_0\|_2^2$. If f is convex, then GD with step-size sequence $\{\eta_k\}$ satisfies,*

$$\Delta_k \leq \left[\prod_{i=0}^{k}(1-\mu_i\eta_i)\right]\Delta_0 + \sum_{i=0}^{k}\left[\prod_{j>i}(1-\mu_i\eta_j)\right]\eta_i^2\left(M_i\eta_i - 1\right)\|\nabla f(x_k)\|_2^2. \tag{16}$$

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

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

Re-arranging this expression allows us to deduce a rate with error terms depending on the local smoothness,

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

■

Proposition 4 *Let $\bar{x}_k = \sum_{i=0}^k \eta_i x_{i+1} / \sum_{i=0}^k \eta_i$. If f is convex, then GD satisfies,*

$$f(\bar{x}_k) - f(x^*) \leq \frac{\|x_0 - x^*\|_2^2 + \sum_{i=0}^k \eta_i^2(\eta_i M_i - 1)\|\nabla f(x_i)\|_2^2}{2\sum_{i=0}^k \eta_i}. \tag{11}$$

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

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

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

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

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

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

■

Appendix C. The Quadratic Case: Proofs

Lemma 9 *Let B be a positive semi-definite matrix and suppose that*

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

Then the point-wise directional smoothness is given by

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

Proof

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

■

Appendix D. Exponential Search: Proofs

Proof [Proof of Theorem 5] This analysis follows (Carmon and Hinder, 2022). First, instantiate Eq. (11) from Proposition 4 with $\eta_i = \eta$ for all i to obtain

$$f(\bar{x}_k) - f_* \leq \frac{\|x_0 - x_*\|^2}{2\eta k} + \frac{\eta\left[\eta\sum_{i=0}^k M(x_{i+1}, x_i)\|\nabla f(x_i)\|^2 - \sum_{i=0}^k \|\nabla f(x_i)\|^2\right]}{2k}. \tag{17}$$

Now, observe that if we get a "Lucky strike" and $\phi(\eta_{\text{hi}}) = \phi(\eta_0) \leq 0$, then specializing Eq. (17) for $\eta = \eta_0$ we get

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

This covers the first case of Theorem 5.

Procedure *ExponentialSearch*(x, L_0)
  for $k \leftarrow 1, 2, 3, \ldots$ **do**
    $\eta_{\text{out}} \leftarrow \text{RootFindingBisection}(x, 2^{-2^k}\eta_0, \eta_0)$ **If** $\eta_{\text{out}} < \infty$ **Return** η_{out}.
  end
end
Procedure *RootFindingBisection*$(x, \eta_{\text{lo}}, \eta_{\text{hi}})$
  Define $\phi(\eta) = \eta - \psi(\eta)$ where $\psi(\eta)$ is given in (13)
  /* One access to ϕ requires T descent steps. */
  If $\phi(\eta_{\text{hi}}) \leq 0$ **Return** η_{hi}. /* Lucky strike. */
  If $\phi(\eta_{\text{lo}}) > 0$ **Return** ∞.
  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 $\eta_{\text{lo}} = \eta_{\text{mid}}$.
    /* Invariant: $\phi(\eta_{\text{hi}}) > 0$, and $\phi(\eta_{\text{lo}}) \leq 0$. */
  end
  Return η_{lo}.
end

Algorithm 1: Gradient descent with exponential search.

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 stepsize η_{lo} satisfies $\frac{\eta_{\text{hi}}}{2} \leq \eta_{\text{lo}} \leq \eta_{hi}$ and $\phi(\eta_{\text{lo}}) \leq 0$. Specializing Eq. (17) for $\eta = \eta_0$ and using that $\phi(\eta_{\text{lo}}) \leq 0$ we get

$$\begin{aligned} f(\bar{x}_k) - f_* &\leq \frac{\|x_0 - x_*\|^2}{2\eta_{\text{lo}} k} + \frac{\eta_{\text{lo}} \sum_{i=0}^k M(x_{i+1}(\eta_{\text{lo}}), x_i(\eta_{\text{lo}}))\|\nabla f(x_i(\eta_{\text{lo}}))\|^2}{2k} \cdot \phi(\eta_{\text{lo}}) \\ &\leq \frac{\|x_0 - x_*\|^2}{2\eta_{\text{lo}} k}. \end{aligned} \tag{18}$$

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

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

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

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

Plugging this into Eq. (18) we obtain

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

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