Optimizing (L_0, L_1)-Smooth Functions by Gradient Methods

Anton Rodomanov (CISPA, Germany)

(joint work with D. Vankov, A. Nedich, L. Sankar, and S. Stich)

3 December 2024
Research Seminar at WIAS
Berlin, Germany

Outline

Motivation

Classical Theory for Gradient Descent

Optimization problem: $f^* := \min_{x \in \mathbb{R}^d} f(x)$, where f is smooth.

Gradient Descent (GD):

$$x_{k+1} = x_k - \eta \nabla f(x_k), \qquad k \geq 0.$$

The standard assumption for analyzing GD is that f is Lipschitz-smooth:

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|, \qquad \forall x, y \in \mathbb{R}^d,$$

which is equivalent to the boundedness of the second derivative:

$$\boxed{\|\nabla^2 f(x)\| \leq L,} \qquad \forall x \in \mathbb{R}^d.$$

Under this assumption, the theory suggests choosing the stepsize

$$\eta = \frac{1}{L}$$

which ensures the good convergence rate of the method.

Are All Smooth Functions Lipschitz-Smooth?

Many smooth functions arising in applications are not Lipschitz-smooth...

For example, $f(x) = |x|^p$ for $p > 2$ or $f(x) = e^x$.

How do we solve optimization problems involving such functions?

Relative Smoothness [Bauschke et al. 2017; Lu et al. 2018]

Instead of Lipschitz-smoothness, we can consider relative smoothness:

$$\boxed{\nabla^2 f(x) \preceq L \nabla^2 \rho(x),} \qquad x \in \mathbb{R}^d,$$

where ρ is a certain convex "reference function".

Then, we can apply the Bregman GD / Mirror Descent:

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^d} \{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + L \beta_\rho(x_k, x) \},$$

where $\beta_\rho(x, y) := \rho(y) - \rho(x) - \langle \nabla \rho(x), y - x \rangle$ is the Bregman distance generated by ρ.

Example: $f(x) = \frac{1}{4} \| Ax - b \|^4 + \frac{1}{2} \| Cx - d \|^2$ is smooth relative to $\rho(x) = \frac{1}{4} \| x \|^4 + \frac{1}{2} \| x \|^2$.

This is a very powerful technique but requires fixing the reference function ρ in advance.

(L_0, L_1)-Smooth Functions [J. Zhang et al. 2020]

In this work, we concentrate instead on another interesting smoothness assumption referred to as (L_0, L_1)-smoothness:

$$\boxed{\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|,} \qquad \forall x \in \mathbb{R}^d.$$

Original motivation: Empirical study of loss functions in Neural Networks for Natural Language Processing (NLP) problems.

NB: f is L-smooth $\iff$ f is $(L, 0)$-smooth.

Basic example: Any polynomial $f(x) = \sum_{i=0}^d a_i x^i$ $(a_i \in \mathbb{R})$ of degree $d \geq 3$ is (L_0, L_1)-smooth but not Lipschitz-smooth.

Indeed, $f'(x) = \sum_{i=1}^d i a_i x^{i-1}$, $f''(x) = \sum_{i=2}^d i(i-1) a_i x^{i-2}$. Therefore $\frac{|f''(x)|}{|f'(x)|} \to 0$ as $|x| \to \infty$, while $|f''(x)|$ is bounded on any compact interval.

Clipped GD

A popular algorithm that provably works for (L_0, L_1)-smooth functions is the Clipped GD:

$$x_{k+1} = x_k - \eta_k \nabla f(x_k), \qquad \eta_k = \min\Big\{\eta, \frac{\gamma}{\|\nabla f(x_k)\|}\Big\},$$

where $\eta = \Theta(\frac{1}{L_0})$ and $\gamma = \Theta(\frac{1}{L_1})$.

- [J. Zhang et al. 2020] showed that, to find an ϵ-stationary point ($\|\nabla f(\bar{x})\| \leq \epsilon$), Clipped GD needs at most $O(\frac{L_0 F_0}{\epsilon^2} + \frac{L_1^2 F_0}{L_0})$ gradient computations, where $F_0 := f(x_0) - f^*$.
- [Koloskova et al. 2023] further improved it up to $O(\frac{L_0 F_0}{\epsilon^2} + \frac{L_1 F_0}{\epsilon})$.

CF: Standard GD for L-smooth functions has complexity of $O(\frac{LF_0}{\epsilon^2})$.

Motivation for This Work

- Further study of (L_0, L_1)-class: main inequalities and properties.
- Why does Clipped GD work for this class? How "natural" is this method and is there any good interpretation for it?
- What is the efficiency of gradient methods when our problem is additionally convex? Can we improve upon the previously known algorithms/results?

(L_0, L_1)-Smooth Functions

Basic Examples

Recall the definition: $\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|$.

Examples:

1. (exponent) $f(x) = e^x$ is (L_0, L_1)-smooth with $L_0 = 0$ and $L_1 = 1$.
2. (logistic function) $f(x) = \ln(1 + e^x)$ is (L_0, L_1)-smooth with arbitrary $L_1 \in [0, 1]$ and $L_0 = \frac{1}{4}(1 - L_1)^2$.
3. (power of Euclidean norm) $f(x) = \frac{1}{p}\|x\|^p$, where $p > 2$, is (L_0, L_1)-smooth with arbitrary $L_1 > 0$ and $L_0 = (\frac{p-2}{L_1})^{p-2}$.

NB: For the same function, the choice of (L_0, L_1) may not be unique.

Calculus of (L_0, L_1)-Smooth Functions

In general, the class is not closed under summation or affine substitution of the arguments. Nevertheless, the class is still closed under some operations.

1. If f_i is $(L_{0,i}, L_{1,i})$-smooth for each $1 \leq i \leq n$, then $f(x) = \sum_{i=1}^n f_i(x_i)$, where $x \equiv (x_1, \ldots, x_n)$, is (L_0, L_1)-smooth with $L_0 = \max_{1 \leq i \leq n} L_{0,i}$ and $L_1 = \max_{1 \leq i \leq n} L_{1,i}$.
2. If f is (L_0, L_1)-smooth and g is L-smooth and M-Lipschitz, then $f + g$ is (L_0', L_1')-smooth with $L_0' = L_0 + ML_1 + L$ and $L_1' = L_1$.
3. If $h(x) = f(\langle a, x \rangle + b)$ and f is (L_0, L_1)-smooth, then h is (L_0', L_1')-smooth with $L_0' = \|a\|^2 L_0$ and $L_1' = \|a\| L_1$.

Main Inequalities

Theorem. Function f is (L_0, L_1)-smooth iff any of the following inequalities holds for any $x, y \in \mathbb{R}^d$:

$$\|\nabla f(y) - \nabla f(x)\| \leq (L_0 + L_1\|\nabla f(x)\|)\frac{e^{L_1\|y-x\|} - 1}{L_1},$$

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \leq (L_0 + L_1\|\nabla f(x)\|)\frac{\phi(L_1\|y - x\|)}{L_1^2},$$

where $\phi(t) := e^t - t - 1$.

CF: These bounds are tighter than those from (B. Zhang et al. 2020; Li et al. 2024).

Lower Bound for Convex Functions

Theorem. Let f be a convex (L_0, L_1)-smooth function. Then, for any $x, y \in \mathbb{R}^d$, we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_0 + L_1 \|\nabla f(y)\|}{L_1^2} \phi_* \Big(\frac{L_1 \|\nabla f(y) - \nabla f(x)\|}{L_0 + L_1 \|\nabla f(y)\|} \Big),$$

where $\phi_*(\gamma) = (1+\gamma)\ln(1+\gamma) - \gamma$ $(\geq \frac{\gamma^2}{2+\gamma})$ is conjugate to ϕ.

Corollary 1:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\|\nabla f(y) - \nabla f(x)\|^2}{2(L_0 + L_1 \|\nabla f(y)\|) + L_1 \|\nabla f(y) - \nabla f(x)\|}.$$

Corollary 2:

$$f(x) - f^* \geq \frac{\|\nabla f(x)\|^2}{2L_0 + 3L_1 \|\nabla f(x)\|}.$$

Gradient Descent (GD)

Minimizing Upper Bound

Natural idea: Minimize the upper bound on the objective:

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + (L_0 + L_1 \|\nabla f(x)\|) \frac{\phi(L_1 \|y - x\|)}{L_1^2},$$

where $\phi(t) = e^t - t - 1$.

The optimal point $y^* = T(x)$ is the result of the gradient step:

$$T(x) = x - r^* \frac{\nabla f(x)}{\|\nabla f(x)\|}, \qquad r^* = \frac{1}{L_1} \ln\Big(1 + \frac{L_1 \|\nabla f(x)\|}{L_0 + L_1 \|\nabla f(x)\|}\Big),$$

resulting in the following bound on improving the function value:

$$\begin{aligned} f(x) - f(T(x)) &\ge \max_{r \ge 0} \Big\{ \|\nabla f(x)\| r - \frac{L_0 + L_1 \|\nabla f(x)\|}{L_1^2} \phi(L_1 r) \Big\} \\ &= \frac{L_0 + L_1 \|\nabla f(x)\|}{L_1^2} \phi_* \Big(\frac{L_1 \|\nabla f(x)\|}{L_0 + L_1 \|\nabla f(x)\|} \Big). \end{aligned}$$

Optimal Stepsize

Thus, the point y^* minimizing the upper bound on the objective is the result of the gradient step

$$T(x) = x - \eta^* \nabla f(x),$$

where the optimal stepsize is given by

$$\boxed{\eta^* = \frac{1}{L_1\|\nabla f(x)\|} \ln\left(1 + \frac{L_1\|\nabla f(x)\|}{L_0 + L_1\|\nabla f(x)\|}\right).}$$

The corresponding progress in decreasing the objective is

$$f(x) - f(T(x)) \geq \frac{L_0 + L_1\|\nabla f(x)\|}{L_1^2} \phi_*\Big(\frac{L_1\|\nabla f(x)\|}{L_0 + L_1\|\nabla f(x)\|}\Big) =: \Delta(x).$$

Simplified Stepsize

The function ϕ_* satisfies $\frac{\gamma^2}{2+\gamma} \leq \phi_*(\gamma) \leq \frac{\gamma^2}{2}$.

From this estimate, it follows that $\Delta(x) \sim \frac{\|\nabla f(x)\|^2}{L_0 + L_1 \|\nabla f(x)\|}$. More precisely:

$$\frac{\|\nabla f(x)\|^2}{2L_0 + 3L_1\|\nabla f(x)\|} \leq \Delta(x) \leq \frac{\|\nabla f(x)\|^2}{2(L_0 + L_1\|\nabla f(x)\|)}.$$

Thus, the gurantee for the optimal stepsize can be simplified:

$$f(x) - f(T(x)) \geq \frac{\|\nabla f(x)\|^2}{2L_0 + 3L_1\|\nabla f(x)\|}.$$

We can obtain the same guarantee by using the simplified stepsize

$$\boxed{\eta_{\mathrm{si}} = \frac{1}{L_0 + \frac{3}{2}L_1\|\nabla f(x)\|}.}$$

Clipping Stepsize

Note that our simplified stepsize is essentially the clipping stepsize:

$$\eta_{\mathrm{si}} \sim \frac{1}{L_0 + L_1\|\nabla f(x)\|} \sim \frac{1}{\max\{L_0, L_1\|\nabla f(x)\|\}} = \min\Big\{\frac{1}{L_0}, \frac{1}{L_1\|\nabla f(x)\|}\Big\}.$$

For the clipping stepsize

$$\boxed{\eta_{\mathrm{cl}} = \min\Big\{\frac{1}{2L_0}, \frac{1}{3L_1\|\nabla f(x)\|}\Big\},}$$

we can show a similar bound on the function progress as before:

$$f(x) - f(T(x)) \geq \frac{\|\nabla f(x)\|^2}{2(2L_0 + 3L_1\|\nabla f(x)\|)}.$$

Various Stepsize Choices: Summary

We have shown that the gradient step

$$T(x) = x - \eta(x)\nabla f(x)$$

is a natural operation minimizing the upper bound on the objective.

The following three stepsizes are equivalent (up to absolute constants) in terms of the objective progress:

1. (Optimal stepsize) $\eta^*(x) = \frac{1}{L_1\|\nabla f(x)\|}\ln(1 + \frac{L_1\|\nabla f(x)\|}{L_0 + L_1\|\nabla f(x)\|})$.
2. (Simplified stepsize) $\eta_{\mathrm{si}}(x) = \frac{1}{L_0 + \frac{3}{2}L_1\|\nabla f(x)\|}$.
3. (Clipping stepsize) $\eta_{\mathrm{cl}}(x) = \min\{\frac{1}{2L_0}, \frac{1}{3L_1\|\nabla f(x)\|}\}$.

These stepsizes satisfy $\eta_{\mathrm{cl}}(x) \le \eta_{\mathrm{si}}(x) \le \eta^*(x)$ and all ensure that

$$\boxed{f(x) - f(T(x)) \ge \frac{\|\nabla f(x)\|^2}{c(2L_0 + 3L_1\|\nabla f(x)\|)},}$$

where $c = 1$ for the first two choices and $c = 2$ for the third one.

GD: Convergence to Stationary Point

Consider now GD

$$x_{k+1} = x_k - \eta(x_k)\nabla f(x_k), \qquad k \geq 0,$$

where $\eta(\cdot)$ is one of the stepsize formulas considered before.

Theorem. For any given $\epsilon > 0$, to reach $\min_{0 \leq i \leq k-1} \|\nabla f(x_i)\| \leq \epsilon$, it suffices to make the following number of iterations:

$$k \geq \frac{(2c)L_0 F_0}{\epsilon^2} + \frac{(3c)L_1 F_0}{\epsilon},$$

where $F_0 = f(x_0) - f^*$, $c = 1$ for the optimal and simplified stepsizes, and $c = 2$ for the clipping stepsize.

CF: This coincides with the best-known rate for the clipped GD from (Koloskova et al. 2023).

Efficiency on Convex Functions

Consider the same method but now additionally assume that f is convex.

Theorem. Let $F_0 := f(x_0) - f^*$. Then, $f(x_k) - f^* \leq \epsilon$ for any given $0 \leq \epsilon < F_0$ whenever

$$k \geq (2c)\frac{L_0 R^2}{\epsilon} + (3c) L_1 R \ln \frac{F_0}{\epsilon} =: k(\epsilon),$$

where $R := \|x_0 - x^*\|$ and $c \in \{1, 2\}$ depending on the stepsize strategy. Furthermore, the distance $\|x_k - x^*\|$ decreases monotonically.

NB: In the worst case, $F_0 \leq \frac{L_0 R^2}{2} \exp(L_1 R)$ and $k(\epsilon) \leq c(2 + \frac{3}{e})\frac{L_0 R^2}{\epsilon} + c(3 + \frac{1}{e}) L_1^2 R^2$.

CF: The previous best-known result for the method from (Li et al. 2024) was enjoying the much worse estimate of $O(\frac{(L_0 + L_1\|\nabla f(x_0)\|)R^2}{\epsilon})$.

Other Algorithms

Normalized Gradient Method

We can also consider the Normalized Gradient Method (NGM):

$$x_{k+1} = x_k - \frac{\beta_k}{\|\nabla f(x_k)\|} \nabla f(x_k), \qquad k \geq 0.$$

Theorem. Consider NGM run for K iterations with constant coefficients:

$$\beta_k = \frac{\hat{R}}{\sqrt{K}}, \qquad 0 \leq k \leq K - 1.$$

Then, for any given $\epsilon > 0$, we have $\min_{0 \leq k \leq K} f(x_k) - f^* \leq \epsilon$ whenever

$$K + 1 \geq \max\left\{ \frac{L_0 \bar{R}^2}{\epsilon}, \frac{4}{9} L_1^2 \bar{R}^2 \right\},$$

where $\bar{R} := \frac{R^2}{\hat{R}} + \hat{R}$ and $R := \|x_0 - x^*\|$.

NB: We can also use time-varying coefficients $\beta_k = \frac{\hat{R}}{\sqrt{k+1}}$. The complexity is the same up to an extra logarithmic factor.

Gradient Method with Polyak Stepsize

Another interesting method is GM with Polyak Stepsize:

$$x_{k+1} = x_k - \frac{f(x_k) - f^*}{\|\nabla f(x_k)\|^2} \nabla f(x_k), \qquad k \geq 0.$$

It also achieves the same complexity (up to absolute constants).

Fast Gradient Method (FGM)

Main idea

1. In the region $Q := \{x : \|\nabla f(x)\| \le \Delta\}$, the function f is essentially standard L-smooth with $L = L_0 + L_1\Delta$.
2. If we could stay inside Q (defined e.g., by $\Delta = \|\nabla f(x_0)\|$), then by running the standard FGM, we can expect the following complexity to find an ϵ-solution: $O(\sqrt{\frac{LR^2}{\epsilon}}) = O(\sqrt{\frac{(L_0+L_1\Delta)R^2}{\epsilon}})$.
3. However, we cannot guarantee that FGM stays in Q.
4. But we can ensure that the iterates remain in the initial sublevel set, $\mathcal{F}_0 := \{x : f(x) \le f(x_0)\}$ on which

$$\psi(\|\nabla f(x)\|) \le f(x) - f^* \le f(x_0) - f^* := F_0,$$

where $\psi(\gamma) := \frac{\gamma^2}{2L_0+3L_1\gamma}$. This means that, for any $x \in \mathcal{F}_0$,

$$\|\nabla f(x)\| \le \psi^{-1}(F_0) =: \Delta \le \sqrt{2L_0F_0} + 3L_1F_0,$$

and so

$$L \le L_0 + L_1\psi^{-1}(F_0) \le 2L_0 + \tfrac{7}{2}L_1^2F_0.$$

Monotone FGM

Algorithm AGMsDR($x_0, T(\cdot), L, K$) [Nesterov et al. 2021]

1: $v_0 = x_0$, $A_0 = 0$.
2: **for** $k = 0, 1, \ldots, K-1$ **do**
3: $\quad y_k = \operatorname{argmin}_y \{ f(y) : y = v_k + \beta(x_k - v_k),\ \beta \in [0,1] \}$.
4: $\quad x_{k+1} = T(y_k)$.
5: $\quad$ Find $a_{k+1} > 0$ from $L a_{k+1}^2 = A_k + a_{k+1}$. Set $A_{k+1} = A_k + a_{k+1}$.
6: $\quad v_{k+1} = v_k - a_{k+1} \nabla f(y_k)$.
return x_K.

This method works for any $T(\cdot)$ such that

$$f(y) - f(T(y)) \geq \frac{1}{2L} \|\nabla f(y)\|^2, \qquad \forall y \in \mathcal{F}_0.$$

In our case, $T(y) = y - \eta(y) \nabla f(y)$, where $\eta(\cdot)$ is one of the stepsize strategies considered earlier.

Efficiency Bounds

Theorem. To ensure that $f(x_k) - f^* \leq \epsilon$ for any given $\epsilon > 0$, AGMsDR needs at most the following number of gradient-oracle calls:

$$O\left(m\sqrt{\frac{(L_0 + L_1^2 F_0)R^2}{\epsilon}}\right)$$

where m is the complexity of finding y_k at each iteration.

NB: This is much better than the previous best result for the method from (Li et al. 2024): $O\Big((L_1^2 R^2 + \frac{L_1^2 F_0}{L_0} + 1)\sqrt{\frac{L_0 R^2 + F_0}{\epsilon}}\Big)$.

Two-stage acceleration procedure

1. Run GD to find x_0 such that $F_0 \equiv f(x_0) - f^* \leq \frac{L_0}{5L_1^2}$.
2. Run AGMsDR from x_0.

Efficiency: $O(L_1^2 R^2 + m\sqrt{\frac{L_0 R^2}{\epsilon}})$.

Experiments

Experiments

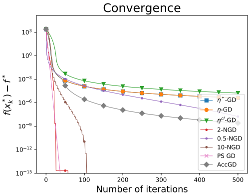
We use the following test problem:

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{p} \|x\|^p \right\}.$$

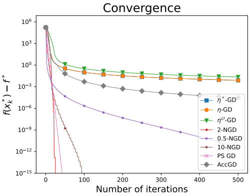
The initial point x_0 is chosen such that $\|x_0\| = R$ with $R = 10$.
We choose

$$L_1 = 1, \qquad L_0 = \left(\frac{p-2}{L_1} \right)^{p-2}.$$

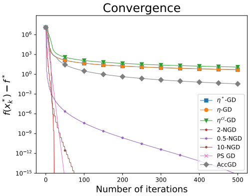
Comparison between different methods:



(a) $p = 4$



(b) $p = 6$

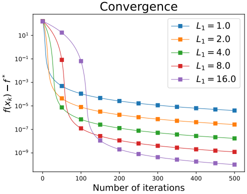


(c) $p = 8$

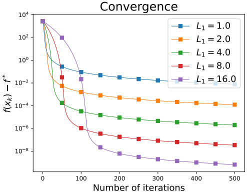
Experiments – II

Recall that $L_1 > 0$ can be arbitrary for the same problem.

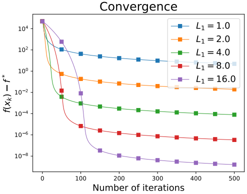
GD with optimal stepsize for different choices of L_1:



(a) $p = 4$



(b) $p = 6$



(c) $p = 8$

Conclusions

Conclusions

- We have seen that GD is a natural method for (L_0, L_1)-smooth functions, obtained by minimizing the upper bound on the objective.
- The clipping stepsize is a simplification of the corresponding optimal stepsize ensuring the same bound on the function progress.
- In the convex case, we have obtained complexities of $O(\frac{L_0 R^2}{\epsilon} + L_1 R \ln \frac{F_0}{\epsilon})$ and $O(m\sqrt{\frac{L_0 R^2}{\epsilon}} + L_1^2 R^2)$ for the basic and accelerated methods, respectively.

Open questions: Acceleration of first stage? Removing line search? Lower bounds? Alternative smoothness assumptions?

Paper (arXiv:2410.10800)

Optimizing (L_0, L_1)-Smooth Functions by Gradient Methods
D. Vankov, A. Rodomanov, A. Nedich, L. Sankar, S. Stich

Thank you!

References I

- H. H. Bauschke, J. Bolte, and M. Teboulle. A Descent Lemma Beyond Lipschitz Gradient Continuity: First-Order Methods Revisited and Applications. **Mathematics of Operations Research**, 42(2):330–348, 2017.
- A. Koloskova, H. Hendrikx, and S. U. Stich. Revisiting Gradient Clipping: Stochastic Bias and Tight Convergence Guarantees. In **International Conference on Machine Learning**, pages 17343–17363. PMLR, 2023.
- H. Li, J. Qian, Y. Tian, A. Rakhlin, and A. Jadbabaie. Convex and Non-convex Optimization Under Generalized Smoothness. **Advances in Neural Information Processing Systems**, 36, 2024.
- H. Lu, R. M. Freund, and Y. Nesterov. Relatively Smooth Convex Optimization by First-Order Methods, and Applications. **SIAM Journal on Optimization**, 28(1):333–354, 2018.

References II

- Y. Nesterov, A. Gasnikov, S. Guminov, and P. Dvurechensky. Primal–dual accelerated gradient methods with small-dimensional relaxation oracle. **Optimization Methods and Software**, 36(4):773–810, 2021.
- B. Zhang, J. Jin, C. Fang, and L. Wang. Improved Analysis of Clipping Algorithms for Non-convex Optimization. **Advances in Neural Information Processing Systems**, 33:15511–15521, 2020.
- J. Zhang, T. He, S. Sra, and A. Jadbabaie. Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity. In **8th International Conference on Learning Representations**, 2020.