Efficient Gradient Methods for Functions with Super-Quadratic Curvature

Anton Rodomanov (CISPA, Germany)

Joint work with Nikita Doikov

18 June 2026
Optimization Seminar at UCLouvain
Louvain-la-Neuve, Belgium

Outline

Outline

Gradient Descent

Consider the unconstrained optimization problem

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable convex function. Assume this problem admits a solution x^*.

Gradient Descent (GD): Choose $x_0 \in \mathbb{R}^n$ and iterate

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

where $M_k > 0$ are step-size coefficients.

Classic Theory

Main assumption: f is Lipschitz-smooth:

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

Standard convergence result: If $\boxed{M_k \equiv L}$, then

$$f(x_{k+1}) - f^* \leq \frac{LR^2}{2(k+1)},$$

where R is the distance from the initial point to the solution:

$$R := \|x_0 - x^*\|.$$

To get $f(x_{k+1}) - f^* \leq \epsilon$, it suffices to make the following number of steps:

$$N_{\text{GD}}(\epsilon) = \frac{LR^2}{2\epsilon}.$$

Key Inequalities in Classic Theory

- Step-size coefficients M_k must be sufficiently large so that

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{M_k}{2}\|x_{k+1} - x_k\|^2 \qquad (*)$$
$$\equiv f(x_k) - \frac{1}{2M_k}\|\nabla f(x_k)\|^2.$$

  This is certainly true whenever $M_k \geq L$.

- If a GD step satisfies (*), then

$$\frac{1}{M_k}[f(x_{k+1}) - f^*] + \frac{1}{2}\|x_{k+1} - x^*\|^2 \leq \frac{1}{2}\|x_k - x^*\|^2.$$

  Consequently,

$$f(x_{k+1}) - f^* \leq \frac{R^2}{2\sum_{i=0}^{k} \frac{1}{M_i}} \leq \frac{M_{\max,k} R^2}{k+1},$$

  where $M_{\max,k} := \max_{0 \leq i \leq k} M_i$.

Line Search

Step-size coefficients can be automatically selected by the algorithm.

Line Search: Given $\hat{M}_k > 0$, find smallest integer $i_k \geq 0$ such that

$$M_k = 2^{i_k} \hat{M}_k, \qquad x_{k+1} = x_k - \frac{1}{M_k} \nabla f(x_k)$$

satisfy inequality (*). Set $\hat{M}_{k+1} = \frac{1}{2} M_k$.

Number of oracle queries: In the worst case,

$$N_{\text{GD-LS}}(\epsilon) = \frac{MR^2}{\epsilon} + \log_2 \frac{4M}{\hat{M}_0},$$

where $M = \max\{2L, \hat{M}_0\}$.

- For a small $\hat{M}_0$, this is the same worst-case efficiency as for the constant step-size version, up to an additive logarithmic factor.
- In practice, much better performance since usually $M_k \ll L$.

Main Observation

Recall the main inequality (*):

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

which, in turn, ensures the key recurrence:

$$\frac{1}{M_k}[f(x_{k+1}) - f^*] + \frac{1}{2}\|x_{k+1} - x^*\|^2 \leq \frac{1}{2}\|x_k - x^*\|^2.$$

Note: $\|x_{k+1} - x^*\| \leq \|x_k - x^*\|$. This means that

$$x_k \in \boxed{B(x^*, R) := \{x : \|x - x^*\| \leq R\},} \qquad \forall k \geq 0,$$

where $R = \|x_0 - x^*\|$.

Do we actually need to assume that f is globally Lipschitz-smooth? Maybe we just need Lipschitz-smoothness on $B(x^*, R)$?

This Talk

Indeed, it is enough to assume smoothness only on $B(x^*, R)$:

$$\|\nabla f(x) - \nabla f(y)\| \leq L_*(R)\|x - y\|, \qquad \forall x, y \in B(x^*, R),$$

and all the previous results are still valid with L replaced by $L_*(R)$!

The same is also true for the Fast Gradient Method (FGM)!

Outline

The Setting

Problem:

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable convex function. Assume this problem admits a solution x^*.

Main assumption: f is $L_*(R)$-smooth on $B(x^*, R)$:

$$\|\nabla f(x) - \nabla f(y)\| \leq L_*(R)\|x - y\|, \qquad \forall x, y \in B(x^*, R),$$

where

$$R := \|x_0 - x^*\|,$$

and $x_0 \in \mathbb{R}^n$ is a given starting point.

Main Property of Gradient Step

Gradient step: $\boxed{T_M(x) := x - \frac{1}{M}\nabla f(x),}$ $x \in \mathbb{R}^n$, $M > 0$.

Main property

If $x \in B(x^*, R)$ and $M \geq L_*(R)$, then $T_M(x) \in B(x^*, R)$.

Proof: Let $T := T_M(x)$. Then,

$$\|T - x^*\|^2 = \|x - x^*\|^2 - \frac{2}{M}\langle \nabla f(x), x - x^* \rangle + \frac{1}{M^2}\|\nabla f(x)\|^2 \leq \|x - x^*\|^2$$

since

$$\langle \nabla f(x), x - x^* \rangle \geq \frac{1}{2L_*(R)}\|\nabla f(x)\|^2 \geq \frac{1}{2M}\|\nabla f(x)\|^2. \quad \square$$

Key Inequality

Let f be convex and $L_{\bar{x}}(r)$-smooth on $B(\bar{x}, r)$. Then, for any $x \in B(\bar{x}, r)$,

$$\begin{aligned}\langle \nabla f(\bar{x}) - \nabla f(x), \bar{x} - x \rangle &\geq f(\bar{x}) - f(x) - \langle \nabla f(x), \bar{x} - x \rangle \\ &\geq \frac{1}{2L_{\bar{x}}(r)} \|\nabla f(\bar{x}) - \nabla f(x)\|^2.\end{aligned}$$

GD with Constant Step-Size Coefficients

Basic Version: $x_{k+1} = T_M(x_k)$, $k \geq 0$, where $M \equiv L_*(R)$.

In this method, $x_k \in B(x^*, R)$ for all k.

Efficiency estimate: To get $f(x_{k+1}) - f^* \leq \epsilon$, it suffices to perform

$$N_{\mathrm{GD}}(\epsilon) = \frac{L_*(R)R^2}{2\epsilon}$$

steps.

Gradient Descent with Line Search (GD-LS)

1: Choose $\hat{M}_0 > 0$.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Find the smallest integer $i_k \geq 0$ such that, for

$$M_k = 2^{i_k} \hat{M}_k, \qquad x_{k+1} = T_{M_k}(x_k)$$

the following inequality is satisfied:

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

4: Set $\hat{M}_{k+1} = \frac{1}{2} M_k$.

In this method, $x_k \in B(x^*, R)$ for all k.

Efficiency estimate: At most the following number of oracle queries:

$$N_{\text{GD-LS}}(\epsilon) = \frac{M_*(R) R^2}{\epsilon} + \log_2 \frac{4 M_*(R)}{\hat{M}_0},$$

where $M_*(R) = \max\{2L_*(R), \hat{M}_0\}$.

Fast Gradient Method (FGM)

$$v_0 = x_0, \qquad A_0 = 0,$$
$$M_k a_{k+1}^2 = A_k + a_{k+1}, \qquad A_{k+1} = A_k + a_{k+1},$$
$$y_k = \frac{A_k x_k + a_{k+1} v_k}{A_{k+1}}, \qquad x_{k+1} = T_{M_k}(y_k),$$
$$v_{k+1} = v_k - a_{k+1} \nabla f(y_k),$$

where M_k is such that

$$f(x_{k+1}) \leq f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k \rangle + \frac{M_k}{2} \|x_{k+1} - y_k\|^2. \qquad (*)$$

In the basic constant step-size version,

$$M_k \equiv L_*(R).$$

Then, (*) is satisfied and we get $x_k, y_k, v_k \in B(x^*, R)$, and

$$N_{\text{FGM}}(\epsilon) = \sqrt{\frac{2L_*(R)R^2}{\epsilon}}.$$

Can also use a standard line search!

Outline

Motivating Example

Problem: $\min_{x\in\mathbb{R}^n}\left\{f(x) := \frac{1}{2}\|A_1x - b_1\|^2 + \frac{1}{4}\|A_2x - b_2\|^4\right\}.$

Note:

$$\nabla^2 f(x) = A_1^T A_1 + A_2^T\left[\|A_2x - b_2\|^2 I + 2(A_2x - b_2)(A_2x - b_2)^T\right]A_2.$$

Hence, f is not Lipschitz-smooth (on the entire space) and the classical theory of gradient methods (GMs) does not apply.

However, $\nabla^2 f(x)$ has polynomial growth:

$$\boxed{\|\nabla^2 f(x)\| \leq p(\|x\|)}$$

for some non-decreasing polynomial $p(\cdot)$. Indeed,

$$\begin{aligned}\|\nabla^2 f(x)\| &\leq \|A_1\|^2 + 3\|A_2\|^2\|A_2x - b_2\|^2\\ &\leq \|A_1\|^2 + 6\|A_2\|^2(\|A_2\|^2\|x\|^2 + \|b_2\|^2).\end{aligned}$$

Existing advice [Lu et al. 2018]: Use Bregman Method (BM).

But the new theory of GMs does apply! Let's compare the two approaches.

RGH Functions: Definition

Extension of the previous idea:

Function with Radially Growing Hessian (RGH)

A twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is called RGH around $\bar{x} \in \mathbb{R}^n$ with modulus $L_{\bar{x}}(\cdot)$ if, for any $r \geq 0$ and any $x \in \mathbb{R}^n$,

$$\|x - \bar{x}\| \leq r \quad \Longrightarrow \quad \|\nabla^2 f(x)\| \leq L_{\bar{x}}(r).$$

Note:

- Equivalent to requiring f be $L_{\bar{x}}(r)$-smooth on $B(\bar{x}, r)$ for any $r \geq 0$.
- Any twice continuously differentiable function is RGH with

$$L_{\bar{x}}(r) = \max_{x \in B(\bar{x}, r)} \|\nabla^2 f(x)\|.$$

RGH Functions: Basic Examples

- $f(\tau) = |\tau|^p$, $p > 2 \implies L_0(r) = p(p-1)r^{p-2}$.
- $f(\tau) = e^{\tau} \implies L_0(r) = e^r$.
- $f(\tau) = e^{\tau^2} \implies L_0(r) = 2(1+2r^2)e^{r^2}$.

Radial function:

$$f(x) = \psi(\|x - \bar{x}\|),$$

where $\psi\colon \mathbb{R}_+ \to \mathbb{R}$ is twice differentiable such that

$$\psi'(0) = 0, \qquad \psi''(\cdot) \text{ is non-negative and non-decreasing.}$$

Then,

$$L_{\bar{x}}(r) = \psi''(r).$$

- $f(x) = \|x - \bar{x}\|^p$, $p > 2 \implies L_{\bar{x}}(r) = p(p-1)r^{p-2}$.
- $f(x) = e^{\|x-\bar{x}\|^2} \implies L_{\bar{x}}(r) = 2(1+2r^2)e^{r^2}$.

RGH Functions: Calculus Rules

- $f(x) = cg(x)$, $c \in \mathbb{R} \implies L_{\bar{x}}(r) = |c| L_{g,\bar{x}}(r)$.
- $f(x) = f_1(x) + f_2(x) \implies L_{\bar{x}}(r) = L_{1,\bar{x}}(r) + L_{2,\bar{x}}(r)$.
- $f(x) = g(Ax + b) \implies L_{\bar{x}}(r) = \alpha^2 L_{g,\bar{y}}(\alpha r + \Delta)$, where $\alpha := \|A\|$ and $\Delta := \|A\bar{x} + b - \bar{y}\|$.

Examples:

- $f(x) = \frac{1}{2}\|A_1 x - b_1\|^2 + \frac{1}{p}\|A_2 x - b_2\|^p$, $p > 2$:

$$L_{\bar{x}}(r) = \|A_1\|^2 + (p-1)\|A_2\|^2(\|A_2\| r + \|A_2\bar{x} - b_2\|)^{p-2}.$$

- $f(x) = \sum_{i=1}^m e^{\langle a_i, x\rangle + b_i}$:

$$L_{\bar{x}}(r) = \sum_{i=1}^m \alpha_i^2 e^{\alpha_i r + \Delta_i},$$

where $\alpha_i = \|a_i\|$ and $\Delta_i = |\langle a_i, \bar{x}\rangle + b_i|$.

Minimizing RGH Function

Problem:

$$\min_{x \in \mathbb{R}^n} f(x),$$

where f is a convex RGH function around x_0 with modulus $L(\cdot)$:

$$\|x - x_0\| \leq r \quad \Longrightarrow \quad \|\nabla^2 f(x)\| \leq L(r).$$

Denote $R := \|x_0 - x^*\|$.

Minor regularity assumption: $L(\cdot)$ is non-decreasing and differentiable.

- Monotonicity is natural from the definition.
- Differentiability ensures BM is well-defined and helps avoid technicalities that are irrelevant to the main ideas.
- Hold for all previous examples and are preserved by calculus rules.

Note: Since $L(\cdot)$ is non-decreasing, our RGH assumption is equivalent to

$$\|\nabla^2 f(x)\| \leq L(\|x - x_0\|), \qquad \forall x \in \mathbb{R}^n.$$

Bregman Method [Bauschke et al. 2017; Lu et al. 2018]

Relative smoothness: $\nabla^2 f(x) \preceq L_d \nabla^2 d(x)$, $\forall x \in \mathbb{R}^n$.

Bregman distance: $\beta_d(x, y) := d(y) - d(x) - \langle \nabla d(x), y - x \rangle$.

Bregman Method (BM):

$$\boxed{x_{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + L_d \beta_d(x_k, x) \},} \quad k \geq 0.$$

Complexity bound:

$$N_{\text{BM}}(\epsilon) = \frac{L_d R_d^2}{2\epsilon},$$

where $\frac{1}{2} R_d^2 := \beta_d(x_0, x^*) \equiv d(x^*) - d(x_0) - \langle \nabla d(x_0), x^* - x_0 \rangle$.

Note: W.l.o.g., we can assume that

$$L_d = 1, \qquad d(x_0) = 0, \qquad \nabla d(x_0) = 0.$$

Then,

$$\boxed{N_{\text{BM}}(\epsilon) = \frac{d(x^*)}{\epsilon}.}$$

Choosing Prox-Function – I

Goal: Find d such that $\nabla^2 f(x) \preceq \nabla^2 d(x)$, $\forall x \in \mathbb{R}^n$.

Our assumption: $\nabla^2 f(x) \preceq L(\|x - x_0\|) I$, $\forall x \in \mathbb{R}^n$.

Natural idea: Find d such that $L(\|x - x_0\|) I \preceq \nabla^2 d(x)$, $\forall x \in \mathbb{R}^n$.

Natural choice: Radial function

$$\boxed{d(x) = \psi(\|x - x_0\|)}$$

for a twice differentiable $\psi : \mathbb{R}_+ \to \mathbb{R}$.

Requirements on ψ:

- $\boxed{\psi'(0) = 0}$ so that d is differentiable at x_0. Then, $\nabla d(x_0) = 0$.
- $\boxed{\psi(0) = 0}$ so that $d(x_0) = 0$.
- Among all possible ψ, choose the one which minimizes $N_{\mathrm{BM}}(\epsilon)$, i.e.,

$$d(x^*) = \psi(\|x^* - x_0\|) = \psi(R).$$

Choosing Prox-Function – II

Derivatives of prox-function:

$$\nabla d(x) = \frac{\psi'(r_x)}{r_x}(x - x_0),$$

$$\nabla^2 d(x) = \frac{\psi'(r_x)}{r_x} I + \left[\psi''(r_x) - \frac{\psi'(r_x)}{r_x}\right] u_x u_x^T,$$

where $r_x := \|x - x_0\|$, $u_x := \frac{x - x_0}{r_x}$ and $\nabla^2 d(x_0) = \psi''(0) I$.

Goal: Ensure that $\nabla^2 d(x) \succeq L(r_x) I$, $\forall x \in \mathbb{R}^n$, i.e., for any $r > 0$,

$$\boxed{\frac{\psi'(r)}{r} \geq L(r)} \qquad \text{and} \qquad \psi''(r) \geq L(r).$$

Solution: The smallest such a function with $\psi(0) = \psi'(0) = 0$ is

$$\boxed{\psi(r) = \int_0^r \tau L(\tau) d\tau.}$$

Note: $\psi'(r) = rL(r)$ and $\psi''(r) = L(r) + rL'(r) \geq L(r)$.

Choosing Prox-Function: Result

Result:

$$\boxed{d(x) = \psi(\|x - x_0\|), \qquad \psi(r) = \int_0^r \tau L(\tau) d\tau.}$$

Note: For functions with polynomially-growing Hessian,

$$L(r) = \sum_{i=0}^{p} \alpha_i r^i,$$

where $\alpha_i \geq 0$, this recovers the recommendation from [Lu et al. 2018]:

$$\psi(r) = \sum_{i=0}^{p} \frac{\alpha_i}{i+2} r^{i+2}.$$

In particular, for $L(r) = \alpha_0 + \alpha_2 r^2$, we get $\psi(r) = \frac{\alpha_0}{2} r^2 + \frac{\alpha_2}{4} r^4$.

Implementing BM

- To implement one step of the method, we need to solve

$$\nabla d(x_{k+1}) = \nabla d(x_k) - \nabla f(x_k).$$

- In our case, $\nabla d(x) = \frac{\psi'(r_x)}{r_x}(x - x_0) = L(r_x)(x - x_0)$, so this is

$$L(r_{k+1})(x_{k+1} - x_0) = L(r_k)(x_k - x_0) - \nabla f(x_k) =: \Delta_k,$$

  where $r_k := \|x_k - x_0\|$.

- We first find r_{k+1} from the non-linear equation

$$L(r)r = \|\Delta_k\|,$$

  and then obtain

$$x_{k+1} = x_0 + \frac{1}{L(r_{k+1})}\Delta_k = x_k - \left[1 - \frac{L(r_k)}{L(r_{k+1})}\right](x_k - x_0) - \frac{1}{L(r_{k+1})}\nabla f(x_k).$$

BM: Complexity Estimate

Recall:

$$N_{\mathrm{BM}}(\epsilon) = \frac{\psi(R)}{\epsilon},$$

where $\psi(R) = \int_0^R \tau L(\tau) d\tau$.

Note: $\boxed{\psi(R) \simeq L(R)R^2}$ in the sense that

$$\psi(R) \leq L(R) \int_0^R \tau d\tau = \tfrac{1}{2} L(R)R^2,$$

$$\psi(R) \geq \int_{\frac{R}{2}}^R \tau L(\tau) d\tau \geq L(\tfrac{R}{2}) \int_{\frac{R}{2}}^R \tau d\tau = \tfrac{3}{8} L(\tfrac{R}{2})R^2.$$

Conclusion:

$$\boxed{N_{\mathrm{BM}}(\epsilon) \simeq \frac{L(R)R^2}{\epsilon}.}$$

Comparison with GD

For the "classical" GD,

$$x_{k+1} = x_k - \frac{1}{M_k}\nabla f(x_k), \quad k \geq 0,$$

with $M_k = L_*(R)$, we have, according to our new theory,

$$\boxed{N_{\text{GD}}(\epsilon) = \frac{L_*(R)R^2}{\epsilon},}$$

where $L_*(R)$ is the smoothness constant on $B(x^*, R)$.

Note: $B(x^*, R) \subseteq B(x_0, 2R)$ and $B(x_0, R) \subseteq B(x^*, 2R)$. Hence,

$$L_*(R) \leq L(2R), \qquad L(R) \leq L_*(2R).$$

Thus: $\boxed{N_{\text{GD}}(\epsilon) \simeq N_{\text{BM}}(\epsilon).}$

Comparison with GD: Discussion

- Thus, GD has the same efficiency guarantee as BM.
- However, GD is simpler:
  - No need to solve non-linear equations.
  - Needs only a scalar $L_*(R)$ in contrast to an entire function $L(\cdot)$ in BM.
- Knowing $L(\cdot)$ is hard. E.g., $f(x) = \frac{1}{2}\|A_1x - b_1\|^2 + \frac{1}{4}\|A_2x - b_2\|^4$ has

$$L(r) = \|A_1\|^2 + 3\|A_2\|^2(\|A_2\|r + \|A_2x_0 - b_2\|)^2.$$

  Depends on the function's structure and problem's data; computing $\|A_1\|$ and $\|A_2\|$ is not easy.
- Knowing $L_*(R)$ is even more difficult. But we can use line search, preserving the same efficiency guarantee (up to an additive log-term). This is not the case for BM: line search is possible, but can we prove anything "good"?
- GD can be accelerated $\Rightarrow$ FGM. For BM, this is an open question.

Outline

Conclusions

- Classical GD and FGM methods do not need the objective to be smooth on the entire space, but only on the restricted set $B(x^*, R)$, where $R = \|x_0 - x^*\|$.
- This also works for methods with line search. The resulting algorithms are completely standard.
- Same results also hold for composite optimization problems.
- Possible application: minimization of RGH functions. No need to apply the sophisticated Bregman Method based on relative smoothness.

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, May 1, 2017. ISSN: 0364-765X. DOI: 10.1287/moor.2016.0817. URL: https://pubsonline.informs.org/doi/10.1287/moor.2016.0817.

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, Jan. 1, 2018. ISSN: 1052-6234. DOI: 10.1137/16M1099546. URL: https://epubs.siam.org/doi/10.1137/16M1099546.