Acceleration in Convex Optimization

rkm0959 (Gyumin Roh)

March 28th

Outline

Table of Contents

Foreword

How to listen to this talk

- a lot of equations : **do not focus on them**
- I will mention what parts of the equations are meaningful
- focus on **ideas**, not routine calculations
- this is a summary of recent survey paper on acceleration
- but this talk also contains some of my thoughts

The Problem

Consider $f : \mathbb{R}^d \to \mathbb{R}$ and the minimization problem

$$f_* = \min_{x \in \mathbb{R}^d} f(x) = f(x_*)$$

Furthermore, assume that

- minimizer x_* exists
- f is convex, differentiable
- f is L-smooth

Our goal : find a $x \in \mathbb{R}^d$ such that

$$f(x) - f_* < \epsilon$$

The Definitions

Convexity

f is convex if for all $x, y \in \mathbb{R}^d$ and $\theta \in [0, 1]$,

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y)$$

If f is differentiable, f is convex if for all $x, y \in \mathbb{R}^d$,

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

The Definitions

Smoothness

f is L-smooth ($L \geq 0$) if for all $x, y \in \mathbb{R}^d$,

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

Strong Convexity

f is μ-strongly convex ($\mu \geq 0$) if for all $x, y \in \mathbb{R}^d$,

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

The Definitions

Function Class

Define $\mathcal{F}_{\mu,L}$ the set of μ-strongly convex, L-smooth functions.
This naturally generalizes to all cases in

$$0 \leq \mu < L \leq +\infty$$

We also define the *condition number* $\kappa = L/\mu$.

Table of Contents

Gradient Descent

$O(1/N)$ Convergence of Gradient Descent

Let f be a function in $\mathcal{F}_{0,L}$. Let $x_0 \in \mathbb{R}^d$ and

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$$

then, for all $N \in \mathbb{N}$ we have

$$f(x_N) - f_* \leq \frac{L||x_0 - x_*||^2}{2N}$$

Lyapunov Function

The idea is to build a *Lyapunov Function* that decreases over iteration. For Gradient Descent, the following works.

Lyapunov Function for Gradient Descent

For any $x_k \in \mathbb{R}^d$, $k \geq 0$ with $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$,

$$\begin{aligned}\phi_{k+1} &= (k+1)(f(x_{k+1}) - f_*) + \frac{L}{2}||x_{k+1} - x_*||^2 \\ &\leq k(f(x_k) - f_*) + \frac{L}{2}||x_k - x_*||^2 = \phi_k\end{aligned}$$

Proof

The idea is to calculate a weighted sum of the inequalities we know from convexity and L-smoothness appropriately. For the above,

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

Calculate $(1) + (2) \cdot (k+1)$ and work - this shows

$$(k+1)(f(x_{k+1}) - f_*) + \frac{L}{2}||x_{k+1} - x_*||^2$$

$$\leq k(f(x_k) - f_*) + \frac{L}{2}||x_k - x_*||^2 - \frac{k}{2L}||\nabla f(x_k)||^2$$

Strong Convexity

$O((1-1/\kappa)^N)$ Convergence of Gradient Descent

f is a L-smooth, μ-strongly convex function. Let $x_0 \in \mathbb{R}^d$ and

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$$

then, for all $N \in \mathbb{N}$ we have

$$f(x_N) - f_* \leq \frac{\mu||x_0 - x_*||^2}{2((1-1/\kappa)^{-N} - 1}$$

Proof : Lyapunov Functions

Table of Contents

Nesterov's AGM

Initialize with $x_0 = y_0 = z_0$. The iteration is

$$
\begin{aligned}
x_{k+1} &= y_k - \frac{1}{L}\nabla f(y_k) \\
z_{k+1} &= z_k - \frac{A_{k+1} - A_k}{L}\nabla f(y_k) \\
y_{k+1} &= \frac{A_{k+1}}{A_{k+2}} x_{k+1} + \left(1 - \frac{A_{k+1}}{A_{k+2}}\right) z_{k+1}
\end{aligned}
$$

where we will specify the values of $\{A_k\}$ later.

Lyapunov Analysis

Lyapunov Function for AGM

If $A_{k+1} \geq A_k \geq 0$ and $(A_{k+1} - A_k)^2 = A_{k+1}$, we have

$$\begin{aligned}\phi_{k+1} &= A_{k+1}(f(x_{k+1}) - f_*) + \frac{L}{2}||z_{k+1} - x_*||^2 \\ &\leq A_k(f(x_k) - f_*) + \frac{L}{2}||z_k - x_*||^2 = \phi_k\end{aligned}$$

Proof

We know the drill by now. Our ingredients are :

- $f_* \geq f(y_k) + \langle \nabla f(y_k), x_* - y_k \rangle$ - (1)
- $f(x_k) \geq f(y_k) + \langle \nabla f(y_k), x_k - y_k \rangle$ - (2)
- $f(x_{k+1}) \leq f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k \rangle + \frac{L}{2}||x_{k+1} - y_k||^2$ - (3)

$(1) \times (A_{k+1} - A_k) + (2) \times A_k + (3) \times A_{k+1}$ shows

$$A_{k+1}(f(x_{k+1}) - f_*) + \frac{L}{2}||z_{k+1} - x_*||^2$$

$$\leq A_k(f(x_k) - f_*) + \frac{L}{2}||z_k - x_*||^2 - \frac{A_{k+1} - (A_{k+1} - A_k)^2}{2L}||\nabla f(y_k)||^2$$

The Convergence of AGM

It can be easily shown that $A_N \geq N^2/4$ with $A_0 = 0$, which gives

$O(1/N^2)$ Convergence of AGM (Nesterov 1983)

For all $N \in \mathbb{N}$ and initial point $x_0 \in \mathbb{R}^d$ we have

$$f(x_N) - f_* \leq \frac{2L||x_0 - x_*||^2}{N^2}$$

Facts : Alternative Forms

It can also be shown that the method is equivalent to

$$a_k = A_{k+1} - A_k = \sqrt{A_{k+1}}$$
$$x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$$
$$y_{k+1} = x_{k+1} + \frac{a_k - 1}{a_{k+1}}(x_{k+1} - x_k)$$

and also, it can be proved that

$$\frac{a_k - 1}{a_{k+1}} = \frac{k}{k+3} + o(1/k)$$

The asymptotic version shows acceleration as well.

Strong Convexity

We will show only the simple form of the asymptotic version

$$x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$$

$$y_{k+1} = x_{k+1} + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}(x_{k+1} - x_k)$$

$O((1-1/\sqrt{\kappa})^N)$ Convergence of AGM

For all $N \in \mathbb{N}$ and initial point $x_0 \in \mathbb{R}^d$ we have

$$f(x_N) - f_* \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^N \frac{\mu + L}{2}||x_0 - x_*||^2$$

Summary and Beyond

The accelerated algorithms show the speedups of

- Smooth Convex : $O(1/N)$ to $O(1/N^2)$
- Smooth Strongly Convex : $O((1-1/\kappa)^N)$ to $O((1-1/\sqrt{\kappa})^N)$

The natural questions to ask are

- Can we go faster? What is possible/impossible?
- Why does this phenomenon happen?

Due to limited time, we focus on the first question.
For the second question, check out some examples below.

- Estimate Sequences (Nesterov)
- Linear Coupling (Allen-Zhu)
- Geometric Descent (Bubeck)

Table of Contents

First Order Methods

We assume that our iterative algorithm calculates

$$x_k \in x_0 + \text{span}\{\nabla f(x_0), \cdots, \nabla f(x_{k-1})\}$$

Our goal is to find x such that

$$f(x) - f_* < \epsilon$$

We also assume that our only knowledge on the function f is

- The function class f belongs to, i.e. parameters μ, L
- "First-Order Black Box" : evaluation of functions/gradients

Roadmap

How do we find algorithms with faster convergence?

- use **tight inequalities**
- what exactly is "tight", and how do we prove "tightness"?

How do we prove no better algorithms exist?

- "adversarial thinking" : find a "counter" function f
- how do we find f systematically?

These ideas will soon be incorporated together.

The Bottom Line

$O(1/N^2)$ is the best (Nesterov)

For any integer $k \leq (d-1)/2$ and any first-order algorithm with given x_0, there exists a function $f \in \mathcal{F}_{0,L} : \mathbb{R}^d \to \mathbb{R}$ such that

$$f(x_k) - f_* \geq \frac{3L||x_0 - x_*||^2}{32(k+1)^2}$$

Proof : WLOG $x_0 = 0$. We construct our function as

$$f(x) = \frac{L}{4}\left(\frac{1}{2}x^T A x - e_1^T x\right)$$

where $e_1 = (1, 0, 0, \cdots, 0)^T$ and $A = \text{tridiag}(-1, 2, -1)$.

Strong Convexity

Another Result by Nesterov

In the same context in the strongly convex setting,

$$f(x_k) - f_* \geq \frac{\mu}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} ||x_0 - x_*||^2$$

Proof is similar, with

$$f(x) = \frac{\mu(\kappa-1)}{4}\left(\frac{1}{2}x^T A x - e_1^T x\right) + \frac{\mu}{2}||x||^2$$

The Interpolation Idea

Interpolation

A set $\{(x_i, g_i, f_i)\}$ is $\mathcal{Q}$-*interpolable* if there exists $f \in \mathcal{Q}$ such that

$$f_i = f(x_i), \quad g_i = \nabla f(x_i)$$

The intuitive way to think is

- $f \in \mathcal{Q}$ gives us some inequalities on x_i, g_i, f_i
- interpolation discusses the opposite direction

Interpolation on $\mathcal{F}_{\mu,L}$

Understanding $\mathcal{F}_{\mu,L}$

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable. $f \in \mathcal{F}_{\mu,L}$ iff

$$\begin{aligned} f(x) \geq & f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2L} ||\nabla f(x) - \nabla f(y)||^2 \\ & + \frac{\mu}{2(1 - \mu/L)} ||(x - y) - \frac{1}{L}(\nabla f(x) - \nabla f(y))||^2 \end{aligned}$$

for all $x, y \in \mathbb{R}^n$. This is also the interpolation condition for $\mathcal{F}_{\mu,L}$ in a sense that the set $\{(x_i, g_i, f_i)\}$ is $\mathcal{F}_{\mu,L}$-interpolable iff

$$\begin{aligned} f(x_i) \geq & f(x_j) + \langle g_j, x_i - x_j \rangle + \frac{1}{2L} ||g_i - g_j||^2 \\ & + \frac{\mu}{2(1 - \mu/L)} ||(x_i - x_j) - \frac{1}{L}(g_i - g_j)||^2 \end{aligned}$$

How to use Interpolation

We can use this "tight" inequality for Lyapunov analysis

- we can use it for *designing* the Lyapunov function
- we can use it for *proving* the necessary inequality

We can also use this for "adversarial" thinking

- we had to design a the function f entirely before
- now we can give $\{(x_i, g_i, f_i)\}$ with the interpolation condition

Performance Estimation Problem (Drori, Teboulle 2014)

A toy example for understanding

How well does Gradient Descent decrease the gradient size?

$$
\begin{aligned}
\text{maximize} \quad & ||\nabla F(x_1)||^2 \\
\text{subject to} \quad & F \in \mathcal{F}_{\mu,L} \\
& x_1 = x_0 - \gamma \nabla F(x_0) \\
& ||\nabla F(x_0)||^2 \leq R^2
\end{aligned}
$$

Performance Estimation Problem (Drori, Teboulle 2014)

We only work with two points, so change variables.
We also add the interpolation condition to the problem.

How well does Gradient Descent decrease the gradient size?

$$\begin{aligned} \text{maximize} \quad & ||g_1||^2 \\ \text{subject to} \quad & \{(x_i, g_i, f_i)\}_{i=0,1} \text{ is } \mathcal{F}_{\mu,L}\text{-interpolable} \\ & x_1 = x_0 - \gamma g_0 \\ & ||g_0||^2 \leq R^2 \end{aligned}$$

- note that new problem does not deal with $F \in \mathcal{F}_{\mu,L}$
- the new problem can be written as a SDP
- leads to computer-assisted proofs (dual problem)

Performance Estimation Problem (Drori, Teboulle 2014)

Consider the first-order algorithm

$$x_k = x_{k-1} - \sum_{i=0}^{k-1} h_{k,i} \nabla f(y_i)$$

and the problem

$$\begin{aligned} \text{maximize} \quad & \frac{f(x_N) - f_*}{||x_0 - x_*||^2} \\ \text{subject to} \quad & x_0 \in \mathbb{R}^d, f \in \mathcal{F}_{0,L} \end{aligned}$$

We wish to minimize the optimal value over $h_{k,i} \in \mathbb{R}$.

Drori, Teboulle does this with PEP formulation, numerically.

Optimized Gradient Method (Kim, Fessler 2016)

Kim, Fessler obtains the **explicit** solution to the minimax problem.

Initialize $x_0 = y_0 = z_0$ and $\theta_{-1,N} = 0$. For $k = 0, 1, \cdots N-1$,

$$\theta_{k,N} = \frac{1 + \sqrt{4\theta_{k-1,N}^2 + 1}}{2}$$

$$y_k = \left(1 - \frac{1}{\theta_{k,N}}\right) x_k + \frac{1}{\theta_{k,N}} z_k$$

$$x_{k+1} = y_k - \frac{1}{L} \nabla f(y_k)$$

$$z_{k+1} = z_k - \frac{2\theta_{k,N}}{L} \nabla f(y_k)$$

Define $\theta_{N,N} = \frac{1+\sqrt{8\theta_{N-1,N}^2+1}}{2}$.

Output : $y_N = \left(1 - \frac{1}{\theta_{N,N}}\right) x_N + \frac{1}{\theta_{N,N}} z_N$.

Lyapunov Analysis of OGM

Keynotes

- proof always uses "tight" inequalities from interpolation
- Lyapunov Function has $f(y_k) - f_* - \frac{1}{2L}||\nabla f(y_k)||^2$
- "last step modification" : analyze twice

Convergence of OGM

OGM satisfies

$$f(y_N) - f_* \leq \frac{L||x_0 - x_*||^2}{2\theta_{N,N}^2} \leq \frac{L||x_0 - x_*||^2}{(N+1)^2}$$

which is approximately $\sqrt{2}$-times optimal in terms of cost.

Information Theoretic Exact Method (Taylor, Drori 2021)

Similar idea, for $\mathcal{F}_{\mu,L}$. Reduces to OGM for $\mu = 0$.
Initialize $x_0 = z_0$ with $A_0 = 0$. Let $q = 1/\kappa = \mu/L$.

$$A_{k+1} = \frac{(1+q)A_k + 2\left(1 + \sqrt{(1+A_k)(1+qA_k)}\right)}{(1-q)^2}$$

$$\tau_k = 1 - A_k/((1-q)A_{k+1})$$

$$\delta_k = ((1-q)^2 A_{k+1} - (1+q)A_k)/(2(1+q+qA_k))$$

$$y_k = x_k + \tau_k(z_k - x_k)$$

$$x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$$

$$z_{k+1} = (1-q\delta_k)z_k + q\delta_k y_k - \frac{\delta_k}{L}\nabla f(y_k)$$

Lyapunov Analysis of ITEM

Keynotes

- proof always uses "tight" inequalities from interpolation
- Lyapunov Function has

$$f(y_k) - f_* - \frac{1}{2L}||\nabla f(y_k)||^2 - \frac{\mu}{2(1-\mu/L)}||y_k - \frac{1}{L}\nabla f(y_k) - x_*||^2$$

$O((1-1/\sqrt{\kappa})^{2N})$ Convergence of ITEM

ITEM satisfies

$$\begin{aligned} ||z_N - x_*||^2 &\leq \frac{1}{1+qA_n}||x_0 - x_*||^2 \\ &\leq \frac{(1-\sqrt{q})^{2N}}{(1-\sqrt{q})^{2N}+q}||x_0 - x_*||^2 \end{aligned}$$

The End : Lower Bounds meet Upper Bounds

OGM is not improvable (Drori, 2017)

For any black-box first-order method with at most N calls to the first-order oracle, if $d \geq N+1$, $\exists f \in \mathcal{F}_{0,L}(\mathbb{R}^d)$ such that

$$f(x_N) - f_* \geq \frac{||x_0 - x_*||^2}{2\theta_{N,N}^2}$$

ITEM is not improvable (Taylor, Drori 2021)

With the same context, with $d \geq 2N+1$, $\exists f \in \mathcal{F}_{0,L}(\mathbb{R}^d)$ such that

$$||x_N - x_*||^2 \geq \frac{||x_0 - x_*||^2}{1 + qA_N}$$

Table of Contents

References

The Survey Paper : https://arxiv.org/abs/2101.09545

Nesterov's Contributions

- A method for solving the convex programming problem with convergence rate $O(1/k^2)$
- Introductory lectures on convex optimization: A basic course

References

Performance Estimation Problem

- Performance of first-order methods for smooth convex minimization: a novel approach (Drori, Teboulle)
- Convex Interpolation and Performance Estimation of First-order Methods for Convex Optimization (Taylor)
- Smooth Strongly Convex Interpolation and Exact Worst-case Performance of First-order Methods
- Performance Estimation Toolbox (PESTO): automated worst-case analysis of first-order optimization methods
- Operator Splitting Performance Estimation: Tight contraction factors and optimal parameter selection

References

Optimized Gradient Method

- Optimized first-order methods for smooth convex minimization (Kim, Fessler)
- The exact information-based complexity of smooth convex minimization (Drori)
- Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions (Taylor, Bach)

Information Theoretic Exact Method

- An optimal gradient method for smooth (possibly strongly) convex minimization (Taylor, Drori)

References

Interpretation of Acceleration

- Introductory lectures on convex optimization: A basic course
- Understanding the Acceleration Phenomenon via High-Resolution Differential Equations
- A geometric alternative to Nesterov's accelerated gradient descent
- Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent