Convergence rate of gradient descent algorithm on β-smooth function

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

manshun.ang@umons.ac.be Homepage: angms.science

First draft: August 2, 2017
Last update : July 27, 2019

Overview

1. Gradient Descent Algorithm
2. Theorem 1. Distance between x_k and x^* decreases with k
3. Theorem 2. $f(x_k)$ and $f(x^*)$ decreases with k
4. Theorem 3. GD with constant step size converges at order $\mathcal{O}(1/k)$
5. Theorem 4. GD with diminishing step size converges at order $\mathcal{O}\left(\frac{\log k}{k}\right)$
6. Discussion and summary

Unconstrained optimization problem

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and β-smooth.

Goal : find the *minimizer* x^*.

The optimal value of the function at the point x^* is denoted as

$$f^* = f(x^*).$$

As f is convex, any local minima of f is the global minima of f.

Gradient Descent Algorithm (GD) - 1/2

GD finds the minimizer x^* by producing a sequence of points

$$\{x_k\}_{k\geq 0} = \{x_1, x_2, \ldots, x_k, \ldots\},$$

where k is the number of step, such that the objective function is monotonically decreasing

$$f(x_0) \geq f(x_1) \geq f(x_2) \geq \cdots \geq f(x_k) \geq \ldots \ ,$$

and the decision variable sequence approaches to the minimizer

$$\lim_{k\to\infty} x_k = x^*. \tag{1}$$

Practically equation (1) is useless as it means k has to be infinity to reach x^*. Instead, an ϵ-accuracy criterion is often used :

$$\text{Stop the algorithm if } \|x_k - x^*\|_2 \leq \epsilon$$

Gradient Descent Algorithm (GD) - 2/2

Given f that is convex and β-smooth, GD produces the sequence of points $\{x_k\}_{k\geq 0}$ by iterating the update equation

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

where t_k is the *step size* parameter.

Interpretation: move the current point x_k in the negative direction of the gradient with step size t_k.

Remark. The algorithm requires function f has to be *differentiable*. If f is non-differentiable, ∇f can be replaced by sub-gradient $g \in \partial f$ and the theory still holds.

Theorems on the GD Algorithm

1. If f is convex and β-smooth, then the distance between x_k and the optimizer x^* decreases with k under step sizes $t_k \leq \frac{1}{\beta}$.
2. The error $f(x_k) - f^*$ decreases with k under step sizes $t_k \leq \frac{1}{\beta}$ as

$$f(x_k) - f^* \leq \frac{\|x_0 - x^*\|^2}{\sum_{i=0}^{k-1} t_i \left(1 - \frac{\beta t_i}{2}\right)}$$

3. GD converges at order $\mathcal{O}\left(\frac{1}{k}\right)$ under constant step size $t_k = \frac{1}{\beta}$

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

4. Under diminishing step size $t_k = \frac{1}{k}$, GD converges as

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

The remaining pages are to prove these results.

Things we need for the proofs

Tool 1. If f is β-smooth, then for any two points $x, y \in \text{dom} f$,

$$0 \leq \left| f(y) - f(x) - \nabla f(x)^T (y - x) \right| \leq \frac{\beta}{2} \|y - x\|_2^2$$

See the details : up to page 8 here.

Tool 2. Sandwich theorem for β-smooth convex function.
If f is convex and β-smooth, then for any two points $x, y \in \text{dom} f$,

$$\frac{1}{2\beta} \|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y) - f(x) - \nabla f(x)^T (y - x) \leq \frac{\beta}{2} \|y - x\|_2^2$$

Interpretation: the error of first order Taylor approximation on f is bounded.
See the details : here.

Tool 3. If f is convex, then $\forall x, y \in \text{dom} f$

$$f(y) - f(x) \leq \nabla f(x)^T (y - x)$$

Theorem 1. Distance between x_k and x^* decreases with k

Theorem 1. Distance between x_k and x^* produced by GD update decreases with k for f is convex and β-smooth, step size is $t_k \leq \frac{1}{\beta}$.

Proof: Apply update $x_{k+1} = x_k - t_k \nabla f(x_k)$ on $\|x_{k+1} - x^*\|_2^2$:

$$\begin{aligned} \|x_{k+1} - x^*\|_2^2 &= \|x_k - t_k \nabla f(x_k) - x^*\|_2^2 \\ \text{[re-arrange]} &= \|(x_k - x^*) - t_k \nabla f(x_k)\|_2^2 \\ \text{[expand]} &= \|x_k - x^*\|_2^2 - 2t_k \nabla f(x_k)^T (x_k - x^*) + t_k^2 \|\nabla f(x_k)\|_2^2 \end{aligned}$$

That is, we have

$$\|x_{k+1} - x^*\|_2^2 = \|x_k - x^*\|_2^2 - 2t_k \nabla f(x_k)^T (x_k - x^*) + t_k^2 \|\nabla f(x_k)\|_2^2. \quad (2)$$

To handle the red term, we can use tool 2 by the fact f is convex and β-smooth : put $y = x^*$, $x = x_k$ in tool 2, and only consider the left part

$$\begin{aligned} \frac{1}{2\beta} \|\nabla f(x_k) - \nabla f(x^*)\|_2^2 &\leq f^* - f(x_k) - \nabla f(x_k)^T (x^* - x_k) \\ &= f^* - f(x_k) + \nabla f(x_k)^T (x_k - x^*) \end{aligned}$$

Next : simplify the expression

First-order optimality condition (FOC) : $\nabla f(x^*) = 0$

$$\frac{1}{2\beta}\|\nabla f(x_k)\|_2^2 \leq f^* - f(x_k) + \nabla f(x_k)^T(x_k - x^*)$$

(FOC is also the reason why we put $y = x^*$ but not $x = x^*$ in tool 2 as we want to keep the $\nabla f(x_k)$ term.)

The next step is tricky, we want to create a zero : rearrange the inequality,

$$f(x_k) - f^* \leq \nabla f(x_k)^T(x_k - x^*) - \frac{1}{2\beta}\|\nabla f(x_k)\|_2^2$$

By definition $f^* = \inf f \implies f(x_k) \geq f^*$ hence

$$0 \leq f(x_k) - f(x^*) \leq \nabla f(x_k)^T(x_k - x^*) - \frac{1}{2\beta}\|\nabla f(x_k)\|_2^2$$

Ignoring the middle

$$0 \leq \nabla f(x_k)^T(x_k - x^*) - \frac{1}{2\beta}\|\nabla f(x_k)\|_2^2$$

Rearrange

$$-\nabla f(x_k)^T(x_k - x^*) \leq -\frac{1}{2\beta}\|\nabla f(x_k)\|_2^2$$

We can now use this inequality on (2)

Use the inequality on (2) :

$$\begin{aligned}\|x_{k+1}-x^*\|_2^2 &= \|x_k-x^*\|_2^2 - 2t_k\nabla f(x_k)^T(x_k-x^*) + t_k^2\|\nabla f(x_k)\|_2^2 \\ &\leq \|x_k-x^*\|_2^2 - t_k\frac{1}{\beta}\|\nabla f(x_k)\|_2^2 + t_k^2\|\nabla f(x_k)\|_2^2 \\ &= \|x_k-x^*\|_2^2 + \left(-t_k\frac{1}{\beta}+t_k^2\right)\|\nabla f(x_k)\|_2^2\end{aligned}$$

That is, we finally have

$$\|x_{k+1}-x^*\|_2^2 \leq \|x_k-x^*\|_2^2 - t_k\left(\frac{1}{\beta}-t_k\right)\|\nabla f(x_k)\|_2^2$$

If $0 \leq t_k \leq \frac{1}{\beta}$, then $t_k\left(\frac{1}{\beta}-t_k\right) \geq 0$ and $t_k\left(\frac{1}{\beta}-t_k\right)\|\nabla f(x_k)\|_2^2$ is something non-negative, hence

$$\|x_{k+1}-x^*\|_2^2 \leq \|x_k-x^*\|_2^2 - \text{something larger than zero}$$

That is,

$$\|x_{k+1}-x^*\|_2^2 \leq \|x_k-x^*\|_2^2 \quad \square$$

Theorem 2. $f(x_k) - f^*$ decreases with k

Theorem 2. For convex and β-smooth f, GD with $t_k \leq \frac{1}{\beta}$ satisfies

$$f(x_k) - f^* \leq \frac{\|x_0 - x^*\|_2^2}{\sum_{i=0}^{k-1} t_i \left(1 - \frac{\beta t_i}{2}\right)}$$

Proof: Start with f is β-smooth. Put $y = x_{k+1} = x_k - t_k \nabla f(x_k)$, $x = x_k$ in tool 1 gives

$$\begin{aligned} f(x_{k+1}) - f(x_k) &= f\big(x_k - t_k \nabla f(x_k)\big) - f(x_k) \\ \text{[use tool 1]} &\leq \nabla f(x_k)^T \big(-t_k \nabla f(x_k)\big) + \frac{\beta}{2} t_k^2 \|\nabla f(x_k)\|_2^2 \\ \big(f(x_{k+1}) - f^*\big) - \big(f(x_k) - f^*\big) &= -t_k \left(1 - \frac{\beta}{2} t_k\right) \|\nabla f(x_k)\|_2^2 \end{aligned}$$

Let $f(x_k) - f^* = \delta_k$, rearrange

$$\delta_k - \delta_{k+1} \geq t_k \left(1 - \frac{\beta}{2} t_k\right) \|\nabla f(x_k)\|_2^2 \qquad (*)$$

Put $y = x_k$,$x = x^*$ in tool 3 gives $f(x_k) - f^* \leq \nabla f(x_k)^T (x_k - x^*)$, then take norm-squared, Cauchy-Schwarz, and recall $f(x_k) - f^* = \delta_k$:

$$\delta_k^2 \leq \|\nabla f(x_k)\|_2^2 \|x_k - x^*\|_2^2$$

Rearrange gives $\|\nabla f(x_k)\|_2^2 \geq \frac{\delta_k^2}{\|x_k - x^*\|_2^2}$, put it in equation (*)

$$\delta_k - \delta_{k+1} \geq t_k \left(1 - \frac{\beta}{2} t_k\right) \frac{\delta_k^2}{\|x_k - x^*\|_2^2}$$

By theorem 1: $\|x_k - x^*\|_2 \leq \|x_0 - x^*\|_2 \iff \frac{1}{\|x_k - x^*\|_2^2} \geq \frac{1}{\|x_0 - x^*\|_2^2}$

$$\delta_k - \delta_{k+1} \geq t_k \left(1 - \frac{\beta}{2} t_k\right) \frac{\delta_k^2}{\|x_0 - x^*\|_2^2}$$

Divide the whole inequality by $\delta_k \delta_{k+1}$

$$\frac{1}{\delta_{k+1}} - \frac{1}{\delta_k} \geq t_k \left(1 - \frac{\beta t_k}{2}\right) \frac{1}{\|x_0 - x^*\|^2} \frac{\delta_k}{\delta_{k+1}}$$

As $\frac{\delta_k}{\delta_{k+1}} = \frac{f(x_k) - f^*}{f(x_{k+1}) - f^*} \geq \frac{f(x_{k+1}) - f^*}{f(x_{k+1}) - f^*} = 1 \implies 1 \geq \frac{\delta_{k+1}}{\delta_k}$, by the logic $a > b > 0$ and $c > d > 0 \implies ac > bd$, multiply the inequality with $1 \geq \frac{\delta_{k+1}}{\delta_k}$ gives

$$\frac{1}{\delta_{k+1}} - \frac{1}{\delta_k} \geq t_k \left(1 - \frac{\beta t_k}{2}\right) \frac{1}{\|x_0 - x^*\|^2}$$

The inequality forms a *telescoping series*:

$$\begin{aligned}
\frac{1}{\delta_1} - \frac{1}{\delta_0} &\geq t_0 \left(1 - \frac{\beta t_0}{2}\right) \frac{1}{\|x_0 - x^*\|^2} \\
\frac{1}{\delta_2} - \frac{1}{\delta_1} &\geq t_1 \left(1 - \frac{\beta t_1}{2}\right) \frac{1}{\|x_0 - x^*\|^2} \\
\vdots &\geq \vdots \\
\frac{1}{\delta_k} - \frac{1}{\delta_{k-1}} &\geq t_{k-1} \left(1 - \frac{\beta t_{k-1}}{2}\right) \frac{1}{\|x_0 - x^*\|^2}
\end{aligned}$$

Sum all of them yields

$$\frac{1}{\delta_k} - \frac{1}{\delta_0} \geq \sum_{i=0}^{k-1} t_i \left(1 - \frac{\beta t_i}{2}\right) \frac{1}{\|x_0 - x^*\|^2}.$$

As $f^* = \inf f \implies \delta_0 = f(x_0) - f^* \geq 0 \implies \frac{1}{\delta_0} \geq 0$, so we have

$$\frac{1}{\delta_k} \geq \frac{1}{\|x_0 - x^*\|^2} \sum_{i=0}^{k-1} t_i \left(1 - \frac{\beta t_i}{2}\right)$$

Express δ_k back to $f(x_k) - f^*$, take reciprocal of the expression finishes the proof. □

Theorem 3. GD with constant step size

Theorem 3. GD converges at order $\mathcal{O}\left(\frac{1}{k}\right)$ with constant step size $t_k = \frac{1}{\beta}$. More precisely,

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

Proof: by Theorem 2

$$f(x_k) - f^* \leq \frac{\|x_0 - x^*\|^2}{\sum_{i=0}^{k-1} t_i \left(1 - \frac{\beta t_i}{2}\right)}.$$

This bound is tightened if $\sum_{i=0}^{k-1} t_i \left(1 - \frac{\beta t_i}{2}\right)$ is maximized.

i.e., all components $t_i \left(1 - \frac{\beta t_i}{2}\right)$ have to be maximized.

Let $g(t) = t\left(1 - \frac{\beta t}{2}\right)$, $\frac{\partial g}{\partial t} = 0$ gives $t = \frac{1}{\beta}$

So if $t = \frac{1}{\beta}$, $\sum_{i=0}^{k-1} t_i \left(1 - \frac{\beta t_i}{2}\right) = \sum_{i=0}^{k-1} \frac{1}{2\beta} = \frac{k}{2\beta}$ and

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

Theorem 4. GD with diminishing step size

Theorem 4. GD converges at order $\mathcal{O}\left(\frac{\log k}{k}\right)$ with diminishing step size $t_k = \frac{1}{k}$. More precisely,

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

How to prove (detail skipped): put $t_k = \frac{1}{k}$ into theorem 2 and apply Riemann Sum approximation. The key is to show

$$\sum \log p \left(1 - \frac{\beta}{2}\log p\right) \approx \frac{\beta k}{2\log k}$$

Note. The sum is now from 1 to k not 0 to $k-1$.

Discussion - On stopping condition

Assumes algorithm stops when $\|f(x_k) - f^*\| \leq \epsilon$ for a given ϵ. (e.g. 10^{-9})
Note that $\|.\|$ can be removed ($\because f^* = \inf f \implies f(x_k) - f^* \geq 0$)

By theorem 3: $f(x_k) - f^* \leq \frac{2\beta\|x_0 - x^*\|_2^2}{k}$, we can count the number of steps needed to reach ϵ-accuracy:

$$\frac{2\beta\|x_0 - x^*\|_2^2}{k} \leq \epsilon \quad \iff \quad k \geq \frac{2\beta\|x_0 - x^*\|_2^2}{\epsilon}$$

i.e. it takes at least $\frac{2\beta\|x_0 - x^*\|_2^2}{\epsilon} = \mathcal{O}\left(\frac{1}{\epsilon}\right)$ steps to converge

Two limitations of this theoretical result:

1. need to know β which is usually unknown
2. need to know x^* which is usually unknown (knowing x^* means you already solved the problem!)

Last page - summary

- Gradient Descent update $x_{k+1} = x_k - t_k \nabla f(x_k)$
- If f is convex and β-smooth, $\|x_k - x^*\|$ decreases with k with $t_k \leq \frac{1}{\beta}$.
- With $t_k \leq \frac{1}{\beta}$, the error $f(x_k) - f^* \leq \frac{\|x_0 - x^*\|^2}{\sum_{i=0}^{k-1} t_i \left(1 - \frac{\beta t_i}{2}\right)}$
- With $t_k = \frac{1}{\beta}$, GD converges : $f(x_k) - f^* \leq \frac{2\beta \|x_0 - x^*\|_2^2}{k}$
- With $t_k = \frac{1}{k}$, GD converges : $f(x_k) - f^* \leq \frac{2\log(k) \|x_0 - x^*\|_2^2}{\beta k}$

  (Detail not shown)

End of document