Gradient Descent Algorithm

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

Optimization Problem

Given a function $f : \mathrm{dom} f = \mathbf{Q} \to \mathbb{R}$, a minimization problem ask for

- the optimal value f^*

$$f^* = \min_{\mathbf{x} \in \mathbf{Q}} f(\mathbf{x}).$$

- the optimizer $\mathbf{x}$ that minimizes the function f

$$\mathbf{x}^* = \arg\min_{\mathbf{x} \in \mathbf{Q}} f(\mathbf{x}).$$

- $\mathbf{x}$: optimization variable (a.k.a. decision variable)
- $\mathrm{dom} f = \mathbf{Q}$: the domain of the function f
- $\mathbf{Q}$ is also called the constraint set
- A solution $\mathbf{x}$ is feasible if $\mathbf{x} \in \mathbf{Q}$
- A solution $\mathbf{x}$ is infeasible if $\mathbf{x} \notin \mathbf{Q}$

Unconstrained Optimization Problem

When $\mathbf{Q} = \mathbb{R}^n$:

$$\min_{\mathbf{x}} f(\mathbf{x}),$$

all $\mathbf{x}$ are feasible for the problem.

Remarks :

- More accurately, "min" should be replaced by "inf"

$$f^* = \inf_{\mathbf{x} \in \mathbf{Q}} f(\mathbf{x})$$

- For unbounded problem, optimal value f^* can be $-\infty$. (e.g. minimizes $-\|\mathbf{x}\|_2^2$)
- For bounded problems, $f^* > -\infty$

Solving unconstrained optimization problem

A way to solve is to come up with an algorithm that produces a sequence

$$\{\mathbf{x}_k\}_{k\in\mathbb{N}} : \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k, ...$$

where $k > 0$ is the iteration number, such that

$$\lim_{k\to\infty} f(\mathbf{x}_k) = f^*.$$

ϵ-**accuracy**: we stop the algorithm when the function value is ϵ-close to the optimal value f^* :

$$0 \leq |f(\mathbf{x}_k) - f^*| \leq \epsilon.$$

- As $f^* \leq f(\mathbf{x})$ for all $\mathbf{x}$, hence the absolute value sign can be removed.
- ϵ is user-defined.
- This stopping condition requires the knowledge of f^*, which is usually not known in advance.

Iterative descent algorithm

One iterative algorithm is called *Descent Algorithm* which iterates

$$\mathbf{x}_{k+1} = \mathbf{x}_k + t_k \Delta_k$$

- k : iteration counter
- $\mathbf{x}_k$: the current decision variable
- $\mathbf{x}_{k+1}$: the decision variable of the next iteration
- $t_k \geq 0$: the step size
- $\Delta_k \in \mathbb{R}^n$: the descent direction
  - $\|\Delta_k\| \neq 1$, which is not "direction" in the common sense

"Descent" is defined as the monotonic non-increasing of function value per iteration :

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) \ \forall k.$$

Iterative descent algorithm

In conceptual level, iterative algorithm can be generalized as

---

Algorithm 1: General framework of descent algorithm

Result: A solution $\mathbf{x}$ that approimately solve $\min_{\mathbf{x}\in\mathbf{Q}} f(\mathbf{x})$

Initialization pick initial point $\mathbf{x}_0 \in \mathbf{Q}$ and initial parameters θ (if any)

while *stopping condition is not met* **do**

  Find a descent direction Δ_k in the form $\Delta_k = p(\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_k; \theta)$

  Pick a step size t_k in the form $t_k = q(\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_k; \Delta_k, \theta)$

  Update $\mathbf{x}_{k+1}$ in the form $r(\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_k; t_k, \Delta_k, \theta)$

  Update θ in the form $s(\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_k, \mathbf{x}_{k+1}; t_k, \Delta_k, \theta)$

end

Questions are

- How to initialize? (How to pick $\mathbf{x}_0$?)
- How to define stopping condition? (What does "converge" mean?)
- How to pick descent direction? (How to pick Δx_k?)
- How to select the step size? (How to pick t_k?)
- How to update? (How to get $\mathbf{x}_{k+1}$ from the available information?)

Theorem on the descent direction

Theorem. If f is convex and differentiable[1], the descent algorithm requires the descent direction Δx_k satisfies the following inequality

$$\nabla f(\mathbf{x}_k)^\top \Delta_k \leq 0$$

Proof: f is convex : $\forall a, b \in \mathrm{dom} f$ we have $f(a) \geq f(b) + \nabla f(b)^\top (a - b)$.
Put $a = \mathbf{x}_{k+1}$, $b = \mathbf{x}_k$,

$$f(\mathbf{x}_{k+1}) \geq f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^\top (\mathbf{x}_{k+1} - \mathbf{x}_k)$$

By definition of update $\mathbf{x}_{k+1} = \mathbf{x}_k + t_k \Delta_k$ thus $\mathbf{x}_{k+1} - \mathbf{x}_k = t_k \Delta_k$ and

$$f(\mathbf{x}_{k+1}) \quad \geq \quad f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^\top t_k \Delta_k = f(\mathbf{x}_k) + t_k \nabla f(\mathbf{x}_k)^\top \Delta_k$$

Re-arrange, we have $f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \geq t_k \nabla f(\mathbf{x}_k)^\top \Delta_k$.
By definition of step size $t_k \geq 0$ and the descent condition $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k)$, we have

$$0 \geq f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \geq t_k \nabla f(\mathbf{x}_k)^\top \Delta_k. \quad \square$$

[1] If f is not differentiable, replaces ∇f by sub-gradient, the theorem still holds.

Gradient descent direction

How to pick Δ_k from $\nabla f(\mathbf{x}_k)^\top \Delta_k \leq 0$?

One way to do so is

$$\Delta_k = -\nabla f(\mathbf{x}_k),$$

since $\nabla f(\mathbf{x}_k)^\top \Delta_k = -\|\nabla f(\mathbf{x}_k)\|_2^2 \leq 0$.

This is the descent direction used in **Gradient Descent** (GD).

It can be shown (next page), GD with a **small enough** step size t_k will converge (to a point).

Notice that the theorem on descent direction does not require things like

- f is convex
- all stationary points are the local minimum of f

GD converges to a point with step size $t \leq \frac{2}{\beta}$

Theorem For a β-smooth function f with optimal value $f^* > -\infty$, GD with step sizes $t_k \leq \frac{2}{\beta}$ converges to a stationary point.

Proof. f is β-smooth implies $\forall a, b \in \mathrm{dom} f$ we have

$$f(b) \leq f(a) + \nabla f(a)^\top (b - a) + \frac{\beta}{2} \|a - b\|_2^2.$$

Put $a = \mathbf{x}_k$, $b = \mathbf{x}_{k+1}$:

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^\top (\mathbf{x}_{k+1} - \mathbf{x}_k) + \frac{\beta}{2} \|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2^2.$$

By definition of GD update $\mathbf{x}_{k+1} - \mathbf{x}_k = -t \nabla f(\mathbf{x}_k)$

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq -t \|\nabla f(\mathbf{x}_k)\|_2^2 + \frac{\beta}{2} t^2 \|\nabla f(\mathbf{x}_k)\|_2^2 = -t \left(1 - \frac{\beta t}{2}\right) \|\nabla f(\mathbf{x}_k)\|_2^2$$

Note $0 \leq t \leq \frac{2}{\beta} \implies t\left(1 - \frac{\beta t}{2}\right) \geq 0$. Rearrange and let $p = \frac{1}{t\left(1 - \frac{\beta t}{2}\right)}$, we get

$$\|\nabla f(x_k)\|_2^2 \leq p\big(f(x_k) - f(x_{k+1})\big),$$

which forms a telescoping series

GD converges to a point with step size $t \leq \frac{2}{\beta}$ ... 2

Telescoping, and let $f_k = f(\mathbf{x}_k)$

$$\begin{array}{cc} k=0 & \|\nabla f_0\|_2^2 \leq p(f_0 - f_1) \\ k=1 & \|\nabla f_1\|_2^2 \leq p(f_1 - f_2) \\ \vdots & \vdots \\ k=K-1 & \|\nabla f_{K-1}\|_2^2 \leq p(f_{K-1} - f_K) \end{array}$$

The sum yields $\sum_{i=0}^{K-1} \|\nabla f_i\|_2^2 \leq p\left(f_0 - f_K\right)$. Take limit to $K \to \infty$:

$$\begin{array}{rcll} \sum_{i=0}^{\infty} \|\nabla f(\mathbf{x}_i)\|_2^2 & \leq & p\left(f(\mathbf{x}_0) - \lim_{K\to\infty} f(\mathbf{x}_K)\right) & \\ & = & p\Big(f(\mathbf{x}_0) - f^*\Big) & \\ & \leq & p\left(f(\mathbf{x}_0) + \infty\right) & \text{by } f^* > -\infty \\ & = & \infty & \end{array}$$

An infinite sum being finite $\implies$ $\|\nabla f(\mathbf{x}_k)\| \to 0$ as $k \to \infty$
i.e. sequence produced by GD algorithm converges to a point.

We know that a point $\mathbf{x}$ is a (1st-order) stationary point if $\|\nabla f(\mathbf{x}_k)\| = 0$, so we further conclude that GD converges to a stationary point of f. $\square$

Last page

Summary :

- Minimization problem
- Descent algorithm iterates $\mathbf{x}_{k+1} = \mathbf{x}_k + t_k \Delta_k$
- Descent algorithm requires $\nabla f(\mathbf{x}_k)^\top \Delta_k \leq 0$
- Gradient Descent algorithm picks $\Delta_k = -\nabla f(\mathbf{x}_k)$
- For β-smooth function, the sequence $\{\mathbf{x}_k\}_{k \in \mathbb{N}}$ produced by GD with sufficient small step size ($t_k \leq \frac{2}{\beta}$) converges to a stationary point of f

Not discussed :

- The convergence of $\{\mathbf{x}_k\}_{k \in \mathbb{N}}$ to a (1st-order) stationary point of f.
- Convergence rate of GD on different f.
- Convergence rate of GD under different step size

End of document