Convergence of gradient flow

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

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

First draft : August 6, 2020
Last update : August 6, 2020

Gradient flow

- Let $f : \mathbb{R}^n \to \mathbb{R}$. Assume this f is a function of x in which is a function of time. i.e., $f\big(x(t)\big)$, where $x \in \mathbb{R}^n$ is a vector.

- Assuming f is continuously differentiable with respect to x and x is continuously differentiable w.r.t. t, then the gradient flow equation associated to this f is then
$$\frac{d}{dt}x(t) = -\nabla f\big(x(t)\big). \tag{1}$$
This equation is an Ordinary Differential Equation (ODE).

- A property of gradient flow is that the function f decreases along the flow. That is, the value $f\big(x(t)\big)$ decreases with time, which means the rate of change of $f\big(x(t)\big)$ is negative:
$$\frac{d}{dt}f\big(x(t)\big) \leq 0. \tag{2}$$
We can show this by definition.

Computing $\dfrac{d}{dt} f\big(x(t)\big)$

- In order to show $\dfrac{d}{dt} f\big(x(t)\big) \leq 0$, we first have to compute $\dfrac{d}{dt} f\big(x(t)\big)$. We use the chain rule:

$$\frac{d}{dt} f\big(x(t)\big) = \left\langle \frac{d}{dx(t)} f\big(x(t)\big)\ ,\ \frac{d}{dt} x(t) \right\rangle,$$

  where $\langle\ ,\ \rangle$ is the inner product.

- Explanation: why we need to take inner product?
  - First, the following "formula" is wrong

$$\frac{d}{dt} f\big(x(t)\big) = \frac{d}{dx(t)} f\big(x(t)\big) \frac{d}{dt} x(t)$$

  - The term $\frac{d}{dt} f\big(x(t)\big)$ is a scalar as the output of $f\big(x(t)\big)$ is a scalar.
  - x is a vector, hence $\frac{d}{dt} x(t)$ is a vector.
  - $\frac{d}{dx(t)} f\big(x(t)\big)$ is a derivative of a scalar with respect to vector, which is a vector
  - Therefore, the inner product is to make the chain rule "dimensionally correct".

The simple direct proof of $\frac{d}{dt} f\big(x(t)\big) \leq 0$

- By chain rule

$$\frac{d}{dt} f\big(x(t)\big) = \left\langle \frac{d}{dx(t)} f\big(x(t)\big), \frac{d}{dt} x(t) \right\rangle$$

- By the definition of gradient flow $\frac{d}{dt} x(t) = -\nabla f\big(x(t)\big)$

$$\frac{d}{dt} f\big(x(t)\big) = \left\langle \frac{d}{dx(t)} f\big(x(t)\big), -\nabla f\big(x(t)\big) \right\rangle$$

- Note that $\frac{d}{dx(t)} f\big(x(t)\big)$ is in fact $\nabla f\big(x(t)\big)$, they are just different notations of the same thing, so

$$\frac{d}{dt} f\big(x(t)\big) = \big\langle \nabla f\big(x(t)\big), -\nabla f\big(x(t)\big) \big\rangle = -\underbrace{\big\|\nabla f\big(x(t)\big)\big\|_2^2}_{\text{nonnegative}} \leq 0.$$

Convergence of gradient flow

- We showed that $f\big(x(t)\big)$ decreases monotonically with time. i.e., $\frac{d}{dt}f\big(x(t)\big) \leq 0$.
- If f is bounded below, i.e., there is a scalar B such that for any $x(t)$,

$$f\big(x(t)\big) \geq B > -\infty,$$

  then $f\big(x(t)\big)$ always converges (to somewhere), this is a standard result:

$$\text{bounded below} + \text{monotonically deceasing} = \text{convergence}.$$

- Note: it says $f\big(x(t)\big)$ converges, not $x(t)$. The sequence $x(t)$ may not converge when it oscillates forever (e.g. bouncing between two/more locations).
- This issue can be fixed by the sufficient conditions for convergence of gradient flow. One of which is based on Lojasiewicz inequality.

Lojasiewicz inequality

- In a simplified form, the inequality states that, for any two points x and y that are *close enough*, then

$$|f(x) - f(y)|^{1-\theta} \leq C\|\nabla f(x)\|,$$

  for some constants $\theta \in (0,1)$ and $C > 0$.
- It literally means that, given a point x, if y is close enough to x, then the magnitude of the difference in their function values, to a power between 0 and 1 which is controlled by an unknown parameter θ, is upper bounded by the size of the gradient at x scaled by a positive constant C.
- What it means: if y is close to x, as long as the norm of the gradient at x is not some crazy large value (i.e. super steep slope), then $f(y)$ has to be close to $f(x)$.
- Or in other words, the norm of the gradient at x bounds the distance between $f(y)$ and $f(x)$.
- This inequality holds for many commonly seen functions. "Practically speaking", we can assume this inequality is always true.

The Lojasiewicz inequlaity for strongly convex function

- If f is m-strongly convex, i.e., $f(x) - \frac{m}{2}\|x\|_2^2$ is convex (for $m > 0$), then the Lojasiewicz inequality becomes

$$f(x) - f^* \leq \frac{1}{2m}\|\nabla f(x)\|_2^2,$$

  where x^* is a (global[1]) minimizer of f. (Note: f may have multiple minimizers), and $f^* = f(x^*)$.

- Compared with $|f(x) - f(y)|^{1-\theta} \leq C\|\nabla f(x)\|$, we have
  - $\theta = 0$.
  - $f(x^*)$ is the minimum value of f, so it is smaller than any $f(x)$ and hence we can remove the absolute sign.
  - $C = \frac{1}{2m} > 0$ as $m > 0$.

- We can make use of this for the convergence of gradient flow.

[1]Global because it is convex function: all local minimizers of a convex function are global minimizer.

Convergence of gradient flow for strongly convex function ... 1/2

- For m-strongly convex f with Lojasiewicz inequality

$$f(x) - f^* \leq \frac{1}{2m}\|\nabla f(x)\|_2^2.$$

Rearrange it gives

$$-\|\nabla f(x)\|_2^2 \leq -2m\big(f(x) - f^*\big).$$

- From (2)

$$\frac{d}{dt}f(x) = -\big\|\nabla f(x)\big\|_2^2 \leq 0.$$

- Plug in the Lojasiewicz inequality gives

$$\frac{d}{dt}f(x) \leq -2m\big(f(x) - f^*\big).$$

Note that we hided the notation (t) for a clearer view.

Convergence of gradient flow for strongly convex function ... 2/2

- Since $\frac{d}{dt} f\big(x(t)\big) = \frac{d}{dt}\big(f(x) - f^*\big)$, so
$$\frac{d}{dt}\big(f(x) - f^*\big) \leq -2m\big(f(x) - f^*\big).$$
- Rearrange
$$\frac{1}{f(x) - f^*}\frac{d}{dt}\big(f(x) - f^*\big) \leq -2m.$$
The left hand size is in the form $\frac{1}{g(x)}\frac{dg(x)}{dx}$ that integration gives $\ln g(x)$. So after integration w.r.t. t from $t = 0$ to t,
$$\ln\left(\frac{f\big(x(t)\big) - f^*}{f\big(x_0\big) - f^*}\right) \leq -2mt,$$
where $x_0 = x(0)$.
- Let $\Delta F_0 = f\big(x_0\big) - f^*$, rearrange the inequality gives
$$f\big(x(t)\big) - f^* \leq e^{-2mt}\Delta F_0.$$
We arrive at exponential convergence of the gradient flow for strongly convex f.

Convergence of gradient flow for convex function ... 1/4

- Now consider the case f is convex but not strongly convex. Convexity of f means

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

  i.e., the 1st-order Taylor approximation of f at the point x is a global under-estimator of f.

- Rearrange the above inequality gives

$$f(y) - f(x) \geq -\langle \nabla f(x), x - y \rangle \tag{3}$$

Convergence of gradient flow for convex function ... 2/4

- The starting point of the convergence of gradient flow for convex function is the time derivative of the distance pf x to the minimizer

$$\frac{d}{dt}\|x(t)-x^*\|_2^2.$$

- Apply chain rule

$$\begin{aligned}
\frac{d}{dt}\|x(t)-x^*\|_2^2 &= \left\langle \frac{d}{d\big(x(t)-x^*\big)}\|x(t)-x^*\|_2^2 \ , \ \frac{d}{dt}\big(x(t)-x^*\big) \right\rangle \\
&= \left\langle 2\big(x(t)-x^*\big) \ , \ \frac{d}{dt}x(t) \right\rangle \\
&\overset{(1)}{=} 2\left\langle x(t)-x^* \ , \ -\nabla f\big(x(t)\big) \right\rangle \\
&\overset{(3)}{\leq} 2\Big(f^* - f\big(x(t)\big)\Big) = -2\Big(f\big(x(t)\big) - f^*\Big).
\end{aligned}$$

- We now get

$$\frac{d}{dt}\|x(t)-x^*\|_2^2 \leq -2\Big(f\big(x(t)\big) - f^*\Big).$$

Convergence of gradient flow for convex function ... 3/4

- Rearrange the inequality

$$f\big(x(t)\big) - f^* \leq \frac{-1}{2}\frac{d}{dt}\|x(t) - x^*\|_2^2.$$

We now integrate the whole equation from time $t = 0$ to time t.

- The right hand side gives

$$\begin{aligned}
& \tfrac{-1}{2}\Big(\|x(t) - x^*\|_2^2\Big) - \tfrac{-1}{2}\Big(\|x_0 - x^*\|_2^2\Big) \\
= \; & \tfrac{-1}{2}\Big(\|x(t) - x^*\|_2^2\Big) + \tfrac{1}{2}\Big(\|x_0 - x^*\|_2^2\Big) \\
= \; & \tfrac{-1}{2}\Big(\|x(t) - x^*\|_2^2\Big) + \tfrac{1}{2}R_0 \\
\leq \; & \tfrac{1}{2}R_0.
\end{aligned}$$

where $R_0 = \|x_0 - x^*\|_2^2$.

- The left hand side gives

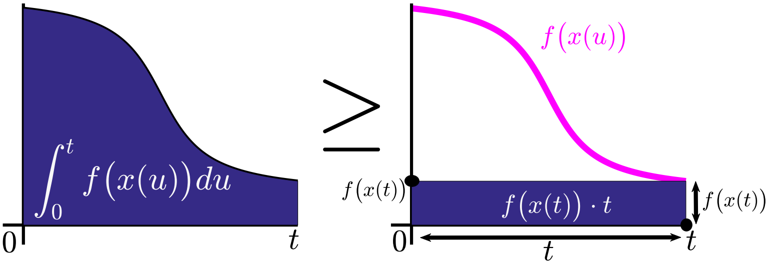
$$\begin{aligned}
\int_0^t \Big(f\big(x(u)\big) - f^*\Big)du &= \int_0^t f\big(x(u)\big)du - \int_0^t f^* du \\
&= \int_0^t f\big(x(u)\big)du - tf^*.
\end{aligned}$$

Convergence of gradient flow for convex function ... 4/4

- Combine the left-hand side and right hand side, now we have the inequality
$$\int_0^t f\big(x(u)\big)du - tf^* \leq \frac{1}{2}R_0.$$
- Divide the whole inequality by t
$$\frac{1}{t}\int_0^t f\big(x(u)\big)du - f^* \leq \frac{1}{2t}R_0.$$
- What we want to show is that
$$f\big(x(t)\big) - f^* \leq \frac{1}{2t}R_0.$$
- That is, we want to show
$$f(x(t)) \leq \frac{1}{t}\int_0^t f\big(x(u)\big)du,$$
which is true since f decreases with time (i.e. (2)). And the proof is completed.

A pictorial illustration of $f(x(t)) \le \frac{1}{t}\int_0^t f\big(x(u)\big)du$,

Note that $f(x(t)) \le \frac{1}{t}\int_0^t f\big(x(u)\big)du \iff f(x(t)) \cdot t \le \int_0^t f\big(x(u)\big)du$.



- The x-axis is the time-axis in u.
- The y-axis is $f(x(u))$.
- The curve $f(x(u))$ goes down as it is monotonically decreasing with time.

Last page - summary

- Gradient flow
$$\frac{d}{dt}x(t) = -\nabla f\big(x(t)\big).$$
- The function f decreases with time.
- For continuously differentiable f,
  - If f is m-strongly convex, then gradient flow converges as
  $$f\big(x(t)\big) - f^* \leq e^{-2mt}\Big(f(x_0) - f^*\Big).$$
  - If f is convex, then gradient flow converges as
  $$f\big(x(t)\big) - f^* \leq \frac{1}{2t}\|x_0 - x^*\|_2^2.$$

End of document