Convergence rates in optimization

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

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

First draft : November 7, 2019
Last update : November 10, 2019

Set up in this document

Consider a minimization problem

$$\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x})$$

where

- f is lower bounded : $\inf f > -\infty$.
- Optimizer $\mathbf{x}^*$ exists

Let

- $\mathcal{X}$ be the set of $\mathbf{x}^*$: $\mathcal{X}$ can have size 1 or more (more than one optimizer)
- $f^* = f(\mathbf{x}^*)$ be the optimal cost value for the problem
- $\mathbf{x}^k$ be the variable $\mathbf{x}$ produced by an algorithm at iteration k
- $f^k = f(\mathbf{x}^k)$ be the cost value of $\mathbf{x}$ at iteration k
- ϵ be the user-defined precision value

Different definitions of convergence

- (Convergence to the minimizer $\mathbf{x}^*$)
  In the worst case, how many iteration is needed to make sure $\mathbf{x}$ is ϵ-close to the minimizer $\mathbf{x}^*$

  $$d(\mathbf{x}^k, \mathbf{x}^*) = \|\mathbf{x}^k - \mathbf{x}^*\| \leq \epsilon$$

  where d is a distance function.
  This refer to the convergence to **the** minimizer $\mathbf{x}^*$.
  The **the** is important : it does not specify *how many* minimizer there are (can be 1 or more)

- (Convergence to the set of minimizer)
  In the worst case, how many iteration is needed to make sure $\mathbf{x}$ is ϵ-close to the set of minimizer

  $$d(\mathbf{x}^k, \mathcal{X}) \leq \epsilon$$

  Again, there is no specification on the size of $\mathcal{X}$, $\mathcal{X}$ can be singleton (so $\mathbf{x}^*$ is the unique global minimizer) or set with multiple elements

Different definitions of convergence

- (Zero-th order convergence on cost value)
  In the worst case, how many iteration is needed to make sure the cost value f^k is ϵ-close to the optimal value

  $$|f(\mathbf{x}^k) - f^*| \leq \epsilon$$

  (Actually the absolute sign can be removed as $f^* = \inf_{\mathbf{x} \in \text{dom} f} f$)
- (1st order convergence)
  In the worst case, how many iteration is needed to make sure the gradient of the cost f at $\mathbf{x}$ is ϵ-close to zero

  $$\|\nabla f(\mathbf{x}^k)\| \leq \epsilon$$

  - For convex f, this also means convergence to the global minimizer $\mathbf{x}^*$
  - For non-convex f, this only means convergence to a local minim or a saddle point
- 2nd/higher order convergence : not in this document

Sub-linear convergence rate

- Sub-linear convergence means k is order $\mathcal{O}\left(\frac{1}{\epsilon^p}\right)$
- The equation

$$\|\mathbf{x}^k - \mathbf{x}^*\| \leq \frac{1}{k^{\frac{1}{p}}} \|\mathbf{x}_0 - \mathbf{x}^*\|$$

where p is a constant, implies sub-linear convergence.

Proof. Let $R_0 = \|\mathbf{x}_0 - \mathbf{x}^*\|$, which is the distance between initial point $\mathbf{x}_0$ and the optimizer $\mathbf{x}^*$. To achieve ϵ-close accuracy, we have

$$\frac{R_0}{k^{\frac{1}{p}}} \leq \epsilon$$

Make k the subject we have

$$k \geq \frac{R_0}{\epsilon^p}$$

which is order $\mathcal{O}\left(\frac{1}{\epsilon^p}\right)$, after ignoring the constant R_0

Linear convergence rate

- Linear convergence means k is order $\mathcal{O}\left(\log \frac{1}{\epsilon}\right)$
- The equation
$$\|\mathbf{x}^k - \mathbf{x}^*\| \leq q\|\mathbf{x}^{k-1} - \mathbf{x}^*\| \tag{1}$$
where $q < 1$, implies linear convergence

Note

- q needs to be smaller than 1 otherwise it does not converge
- For $q = 1$, it does not converge nor diverge, for $q > 1$ it diverges
- Equation (1) is a recursion and holds for all k, so we have

$$\begin{aligned}
\|\mathbf{x}^1 - \mathbf{x}^*\| &\leq q\|\mathbf{x}^0 - \mathbf{x}^*\| \\
\|\mathbf{x}^2 - \mathbf{x}^*\| &\leq q\|\mathbf{x}^1 - \mathbf{x}^*\| \\
\|\mathbf{x}^3 - \mathbf{x}^*\| &\leq q\|\mathbf{x}^2 - \mathbf{x}^*\| \\
&\vdots \\
\|\mathbf{x}^k - \mathbf{x}^*\| &\leq q\|\mathbf{x}^{k-1} - \mathbf{x}^*\|
\end{aligned}$$

$\|\mathbf{x}^k - \mathbf{x}^*\| \leq q\|\mathbf{x}^{k-1} - \mathbf{x}^*\|, q < 1$ implies linear convergence rate

Apply recursion

$\|\mathbf{x}^k - \mathbf{x}^*\| \leq q\|\mathbf{x}^{k-1} - \mathbf{x}^*\| \leq q^2\|\mathbf{x}^{k-2} - \mathbf{x}^*\| \leq \cdots \leq q^k\|\mathbf{x}^0 - \mathbf{x}^*\| = q^k R_0$

To achieve ϵ-close accuracy, we need $q^k R_0 \leq \epsilon$, or

$$q^k \leq \frac{\epsilon}{R_0}$$

Make k the subject by taking log

$$k \log q \leq \log \frac{\epsilon}{R_0} = \log \epsilon - \log R_0$$

Divide $\log q$, note that $q < 1$ so $\log q$ is negative, flip the inequality sign

$$k \geq \frac{\log \epsilon}{\log q} - \frac{\log R_0}{\log q}$$

As $\log q$ is negative, $\log q = -|\log q|$, together with $\log \epsilon = -\log \dfrac{1}{\epsilon}$, we have

$$k \geq \frac{1}{|\log q|} \log \frac{1}{\epsilon} + \frac{\log R_0}{|\log q|} = c_1 \log \frac{1}{\epsilon} + c_2$$

Linear convergence rate

- For linear convergence rate

$$k \geq c_1 \log \frac{1}{\epsilon} + c_2$$

  k is of order $\mathcal{O}\left(\log \frac{1}{\epsilon}\right)$ *in the long run* : as $c_1 = \frac{1}{|\log q|}$ can be (very) small and $c_2 = \frac{\log R_0}{|\log q|}$ can be (very) large

- It also means that it is possible that linear convergence rate is not observed during the *first few iterations* : the first few iteration can be slow
- Compared with sub-linear convergence rate of order $\mathcal{O}\left(\frac{1}{\epsilon}\right)$, there is an additional log in linear convergence rate

Quadratic convergence rate

- Quadratic convergence means k is order $\mathcal{O}\left(\log\log\frac{1}{\epsilon}\right)$
- The equation

$$\|\mathbf{x}^k - \mathbf{x}^*\| \leq q\|\mathbf{x}^{k-1} - \mathbf{x}^*\|^2 \tag{2}$$

  implies quadratic convergence

Note

- q here does not need to be smaller than 1
- Equation (2) is a recursion and holds for all k :

$$\|\mathbf{x}^{k-1} - \mathbf{x}^*\| \quad \leq \quad q\|\mathbf{x}^{k-2} - \mathbf{x}^*\|^2$$

  implies

$$\|\mathbf{x}^{k-1} - \mathbf{x}^*\|^2 \quad \leq \quad q^2\|\mathbf{x}^{k-2} - \mathbf{x}^*\|^4$$

$\|\mathbf{x}^k - \mathbf{x}^*\| \leq q\|\mathbf{x}^{k-1} - \mathbf{x}^*\|^2$ implies quadratic convergence rate

Apply recursion

$$
\begin{aligned}
\|\mathbf{x}^k - \mathbf{x}^*\| &\leq \leq q\|\mathbf{x}^{k-1} - \mathbf{x}^*\|^2 \\
&\leq qq^2\|\mathbf{x}^{k-2} - \mathbf{x}^*\|^4 \\
&\vdots \\
&\leq qq^2 \dots q^{2^k}\|\mathbf{x}^0 - \mathbf{x}^*\|^{2^k} \\
&= q^{1+2+4+\cdots+2^k} R_0^{2^k} \\
&= q^{\frac{1-2^k}{1-2}} R_0^{2^k} \\
&= q^{2^k-1} R_0^{2^k}
\end{aligned}
$$

To achieve ϵ-close accuracy, we need $q^{2^k-1}R_0^{2^k} \leq \epsilon$, what we want to do now is to make k the subject. Here we cannot just throw $R_0^{2^k}$ to the other side since it contains k, we need to do some trick :

$$
q^{2^k-1}R_0^{2^k} = \frac{1}{q}q^{2^k}R_0^{2^k} = c(qR_0)^{2^k}
$$

where $c = q$ is just a constant.

$\|\mathbf{x}^k - \mathbf{x}^*\| \leq q\|\mathbf{x}^{k-1} - \mathbf{x}^*\|^2$ implies quadratic convergence rate

From $q^{2^k-1}R_0^{2^k} \leq \epsilon$, we have $(qR_0)^{2^k} \leq \dfrac{\epsilon}{c}$. Take log

$$2^k \log qR_0 \leq \log \frac{\epsilon}{c} = \log \epsilon - \log c = \log \epsilon + c$$

Assume $qR_0 \leq 1$, so $\log qR_0$ is negative. Divide the whole equation by $\log qR_0$, flip the inequality sign

$$2^k \geq \frac{\log \epsilon}{\log qR_0} + c$$

where all terms independent of ϵ are absorbed by c.
As $\log qR_0$ is negative so $\log qR_0 = -|\log qR_0|$, together with $\log \epsilon = -\log \dfrac{1}{\epsilon}$ we have

$$2^k \geq \frac{1}{|\log qR_0|} \log \frac{1}{\epsilon} + c$$

Take log again, and all terms independent of ϵ are absorbed by c

$$k \geq c_1 \log \log \frac{1}{\epsilon} + c_2$$

Quadratic convergence rate

- For quadratic convergence rate

$$k \geq c_1 \log\log\frac{1}{\epsilon} + c_2$$

  this requires $qR_0 < 1$, as $R_0 = \|\mathbf{x}^0 - \mathbf{x}^*\|$, $qR_0 < 1$ implies that $\|\mathbf{x}^0 - \mathbf{x}^*\| < \frac{1}{q}$, meaning that $\mathbf{x}^0$ has to be *sufficiently close* to $\mathbf{x}^*$. In other words, the initial point cannot be too far away from the minimizer. This is the typical requirement of Second order method (e.g. Newton's method)

- Compared with linear convergence rate, quadratic convergence rate has one more log

Remarks on the rate

Says method A has a better convergence rate than method B.
Does it mean method A is always better than method B?
The answer is : NO.

- The rates are the worst case convergence rate : it does not means method A with a better rate *always* run faster than method B.
- Big-O notation means that the rates are *in the long run*, it is possible method B is faster than method A *in the first few iterations*
  - Some problems takes days to run 1 iteration, or only a few iterations are affordable, hence it is possible B is better.
  - It is possible the constants in the rate are not insignificant and can quite affect rate
- Therefore it is important to always try different algorithms on the problem to solve, *do not believe blindly* that a method with better rate is always better

Remarks on the big-O notation

In fact the big-O notation is wrongly used in optimization.

In general, big-Ω should be used instead of big-O for lower bound.
i.e., we should have

$$k \geq \Omega(\frac{1}{\epsilon})$$

instead of

$$k \geq \mathcal{O}(\frac{1}{\epsilon})$$

The proper use of big-O is for upper bound, not for lower bound.

However people don't care as it already became popular.

Last page - summary

- $\|\mathbf{x}^k - \mathbf{x}^*\| \leq \dfrac{1}{k^{\frac{1}{p}}}\|\mathbf{x}_0 - \mathbf{x}^*\|$ implies sub-linear convergence rate $\mathcal{O}\left(\frac{1}{\epsilon}\right)$
- $\|\mathbf{x}^k - \mathbf{x}^*\| \leq q\|\mathbf{x}^{k-1} - \mathbf{x}^*\|$ implies linear convergence rate $\mathcal{O}\left(\log\frac{1}{\epsilon}\right)$, where $q < 1$
- $\|\mathbf{x}^k - \mathbf{x}^*\| \leq q\|\mathbf{x}^{k-1} - \mathbf{x}^*\|^2$ implies quadratic convergence rate $\mathcal{O}\left(\log\log\frac{1}{\epsilon}\right)$, where $q\|\mathbf{x}^0 - \mathbf{x}^*\| < 1$

End of document