Lower bounds of convergence rate of gradient-based methods on convex problems

Optimal convergence rate of first order method on smooth convex f is $\mathcal{O}(1/k^2)$

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

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

First draft : April 22, 2020
Last update : April 23, 2020

Problem setup and convergence bounds

- Minimization problem

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

  - $\mathbf{x} \in \mathbb{R}^n$ is the optimization variable
  - $f : \mathbb{R}^n \to \mathbb{R}$ is convex
- The lower bounds of convergence rate of using gradient-based method to solve this problem are

$$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right) \qquad \text{if } f \text{ is } L\text{-Lipschitz.}$$

$$\mathcal{O}\left(\frac{1}{k^2}\right) \qquad \text{if } f \text{ is } L\text{-smooth.}$$

  That is, it takes at least these number of steps for any gradient method to reach an ϵ-accurate solution.
- How to derive these bounds?

Expanding the iterate of the gradient method

- The iterate of gradient descent (GD)

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k),$$

  where k is the iteration counter, $\alpha_k > 0$ is the step size, and $\nabla f(\mathbf{x}_k)$ is the gradient of f at the point $\mathbf{x}_k$.

- Expand the iterate on k

$$\begin{aligned}
\mathbf{x}_{k+1} &= \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k), \\
&= \mathbf{x}_{k-1} - \alpha \nabla f(\mathbf{x}_{k-1}) - \alpha \nabla f(\mathbf{x}_k), \\
&= \mathbf{x}_{k-1} - \sum_{j=0}^{1} \alpha_{k-j} \nabla f(\mathbf{x}_{k-j}), \\
&\vdots \\
&= \mathbf{x}_0 - \sum_{j=0}^{k} \alpha_{k-j} \nabla f(\mathbf{x}_{k-j}).
\end{aligned}$$

Black-box procedure

- As $\sum_j \alpha_{k-j} \nabla f(\mathbf{x}_{k-j})$ is the linear combination of $\nabla f(\mathbf{x}_0), \nabla f(\mathbf{x}_1), \ldots, \nabla f(\mathbf{x}_k)$, we can see that in general we have

$$\mathbf{x}_{k+1} \in \mathbf{x}_0 + \text{span}\Big\{\nabla f(\mathbf{x}_0), \nabla f(\mathbf{x}_1), \ldots, \nabla f(\mathbf{x}_k)\Big\}. \tag{1}$$

- Using $\mathbf{g}_i \in \partial f(\mathbf{x}_i)$ to genealize gradient to subgradient for handling nonsmooth case, we have the following Block-box procedure (Bbp) as a generalization of (1)

$$Bbp \ : \ \mathbf{x}_{k+1} \in \mathbf{x}_0 + \text{span}\Big\{\mathbf{g}_1, \mathbf{g}_1, \ldots, \mathbf{g}_k\Big\}.$$

- Idea: if we can derive a lower bound of the convergence rate of Bbp on minimizing f, we then can tell what is the lower bound on the convergence rate for all gradient-based methods.

- Such bound depends on two properties of f : smoothness and convexity

Theorem 1 : f is L_f-Lipschitz

- **Theorem 1** There exists a function f that is L-Lipschitz and convex such that any Bbp satisfies

$$\min_{t\in[1,k]} f(\mathbf{x}_t) - \min_{\mathbf{x}\in B_2(R)} f(\mathbf{x}) \geq \frac{RL}{2(1+\sqrt{k})},$$

for $R > 0$ and $k \leq n$ (recall n is the dimension of $\mathbf{x}$).

- Meaning : for all Bbp, there is always a function f, that the minimum of $f(\mathbf{x}_t)$ among the first k iterates, with k smaller than the dimension of $\mathbf{x}$, is lower bounded as $\mathcal{O}(\frac{1}{\sqrt{k}})$, for $\mathbf{x}$ inside a L_2 ball B_2 with radius R.

- Idea of the proof: build such a function f that, for all Bbp, we have

$$\text{span}\Big\{\mathbf{g}_1, \mathbf{g}_1, \ldots, \mathbf{g}_k\Big\} \subset \text{span}\Big\{\mathbf{e}_1, \ldots, \mathbf{e}_i\Big\},$$

where $\mathbf{e}_i$ is the ith standard basis vector. At iteration k ($\leq n$), there are at least $n-k$ coordinate of $\mathbf{x}$ are 0. This helps us to derive a bound on the error.

The proof of Theorem 1 ... (1/6)

- Consider the function

$$\begin{aligned} f(\mathbf{x}) &= \beta \max_{i\in[1,k]} x[i] + \frac{\alpha}{2}\|\mathbf{x}\|_2^2, \\ &= \beta \|\mathbf{x}[1:k]\|_\infty + \frac{\alpha}{2}\|\mathbf{x}\|_2^2, \end{aligned}$$

  where $\alpha, \beta \in \mathbb{R}$ are free parameters, $\mathbf{x}[1:k]$ is the first k component of $\mathbf{x}$.

- This f is α-strongly convex.

- This f is non-smooth (the $\|\cdot\|_\infty$ part is non-smooth)

- As we are considering the minimization of f by Bbp, which is gradient-based method, we need the gradient of f. Here f is non-smooth, so we use the subgradient.

The proof of Theorem 1 ... (2/6)

- For the L_∞ norm $\|\mathbf{z}\|_\infty$, its subdifferential can be expressed as a convex hull[1] :

$$\partial\|\mathbf{z}\|_\infty = \operatorname{conv}\Big\{\mathbf{w} \mid \|\mathbf{w}\|_1 \leq 1, \mathbf{z}^\top\mathbf{w} = \|\mathbf{z}\|_\infty\Big\}.$$

  using standard basis vector $\mathbf{e}$,

$$\partial\|\mathbf{x}[1:k]\|_\infty = \operatorname{conv}\Big\{\mathbf{e}_i \mid i \in \operatorname*{argmin}_{i\in[1,k]} x[i]\Big\}.$$

- The subdifferential of f is then

$$\partial f = \beta \operatorname{conv}\Big\{\mathbf{e}_i \mid i \in \operatorname*{argmin}_{i\in[1,k]} x[i]\Big\} + \alpha\mathbf{x}.$$

[1] See for example, proposition A.22 of the PhD thesis of Bertsekas.

The proof of Theorem 1 ... (3/6)

- What we have so far

$$\begin{aligned} f(\mathbf{x}) &= \beta\|\mathbf{x}(1:k)\|_\infty + \frac{\alpha}{2}\|\mathbf{x}\|_2^2, \\ \partial f &= \beta\,\mathrm{conv}\Big\{\mathbf{e}_i \mid i \in \operatorname*{argmin}_{i\in[1,k]} x[i]\Big\} + \alpha\mathbf{x}. \end{aligned}$$

- We now show the subgradient of f is Lipschitz. For $\mathbf{g}(\mathbf{x}) \in \partial f$ where $\|\mathbf{x}\|_2 \le R$, we have

$$\begin{aligned} \|\mathbf{g}\|_2 &= \Big\|\beta\,\mathrm{conv}\Big\{\mathbf{e}_i \mid i \in \operatorname*{argmin}_{i\in[1,k]} x[i]\Big\} + \alpha\mathbf{x}\Big\|_2 \\ &\le \beta \underbrace{\Big\|\mathrm{conv}\Big\{\mathbf{e}_i \mid i \in \operatorname*{argmin}_{i\in[1,k]} x[i]\Big\}\Big\|_2}_{=1} + \alpha \underbrace{\|\mathbf{x}\|_2}_{\le R} \\ &\le \beta + \alpha R, \end{aligned}$$

  where the first inequality sign comes from triangle inequality.
  Now we have that $\mathbf{g}$ is at most $(\beta+\alpha R)$-Lipschitz, or f is at most $(\beta+\alpha R)$-smooth on $B_2(R)$. We have $L_f = \beta + \alpha R$.
- Although α, β are free parameters, here we cannot assign their values.

The proof of Theorem 1 ... (4/6)

- Suppose the gradient is $\mathbf{g}_i = \beta \mathbf{e}_i + \alpha \mathbf{x}$ at iteration i, where i is also the first coorindate such that $x[i] = \|x[1:k]\|_\infty$. We have

$$\mathbf{x}_t \in \text{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_{t-1}\},$$

  assuming we start with $\mathbf{x}_0 = 0$ (for simplicity).

- Due to the structure of f, for $t \leq k$, we have

$$f(\mathbf{x}_t) \geq 0. \tag{2}$$

- Now, consider the minimizer of f with an analytic expression. Let $\mathbf{y} \in \mathbb{R}^n$ such that $y[i] = \frac{-\beta}{\alpha k}$ if $i \in [1, k]$, and $y[i] = 0$ otherwise. The vector $\mathbf{y}$ is a minimizer of f, and

$$f(\mathbf{y}) = \frac{-\beta^2}{\alpha k} + \frac{\alpha}{2} \frac{\gamma^2}{\alpha^2 k} = -\frac{\beta^2}{2\alpha k}. \tag{3}$$

The proof of Theorem 1 ... (5/6)

- Now we have

$$\begin{aligned} f(\mathbf{x}_t) - f(\mathbf{y}) &\overset{(2)}{\geq} 0 - f(\mathbf{y}) \\ &\overset{(3)}{=} 0 - \Big(- \frac{\beta^2}{2\alpha k} \Big) \\ &= \frac{\beta^2}{2\alpha k}. \end{aligned}$$

- Compare to what we want to prove :

$$\min_{t \in [1,k]} f(\mathbf{x}_t) - \min_{\mathbf{x} \in B_2(R)} f(\mathbf{x}) \geq \frac{RL}{2(1+\sqrt{k})},$$

  we can see that by selecting appropriate α, β such that $\mathbf{y} \in B_2(R)$, we finish the theorem.

- Here we need to solve for α, β such that

$$\frac{\beta^2}{2\alpha k} = \frac{RL}{2(1+\sqrt{k})}, \quad \|\mathbf{y}(\alpha, \beta)\|_2 = R.$$

The proof of Theorem 1 ... (6/6)

- From the definition of $\mathbf{y}$, we have

$$\|\mathbf{y}\|_2 = \left\| \underbrace{\frac{-\beta}{\alpha k}, \ldots, \frac{-\beta}{\alpha k}}_{k \text{ of them}}, 0, \ldots, 0 \right\|_2 = \sqrt{k\left(\frac{-\beta}{\alpha k}\right)^2} = \frac{\beta}{\alpha\sqrt{k}} = R.$$

- So from $\dfrac{\beta^2}{\alpha k} = \dfrac{RL}{1+\sqrt{k}}$,

$$\frac{\beta^2}{\alpha k} = \frac{\beta}{\alpha\sqrt{k}}\frac{\beta}{\sqrt{k}} = R\frac{\beta}{\sqrt{k}} = \frac{RL}{1+\sqrt{k}} \implies \beta = \frac{\sqrt{k}}{1+\sqrt{k}},$$

  and

$$\alpha = \frac{L}{R}\frac{1}{1+\sqrt{k}}.$$

- By selecting these α, β, the proof is completed. □

Theorem 2 : f is L_f-smooth

- **Theorem 2** There exists a function f that is L-smooth and convex such that any Bbp satisfies

$$\min_{t\in[1,k]} f(\mathbf{x}_t) - f(\mathbf{x}^*) \geq \frac{3L_f R_0^2}{32(1+k)^2},$$

for $R_0 = \|\mathbf{x}_0 - \mathbf{x}^*\|_2.$, $L_f > 0$, $k \leq \frac{n-1}{2}$.

- Meaning : no matter what gradient method you provide, there is always a function f that, when you apply your gradient method on minimizing such f, the convergence rate is lower bounded as $\mathcal{O}(1/k^2)$.

- The difference between Theorem 1 and Theorem 2 : both are lower bound on convergence rate. The difference is at the assumption of f : is it smooth or not. In Theorem 2, we assume the f is L_f-smooth.

- Idea of the proof: similar to the L_f-Lipschitz case.

Nesterov's worst function ... (1/2)

- The key of the proof is to build a special function f.
- Let $n = 2k + 1$ and $\mathbf{A} \in \mathbb{R}^{n \times n}$ as

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & & \dots & \\ -1 & 2 & -1 & \dots & \\ & -1 & 2 & \dots & \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ & & & \dots & 2 \end{bmatrix},$$

after some algebra, we have

$$\mathbf{x}^\top \mathbf{A} \mathbf{x} = x[1]^2 + \sum_{i=1}^{n-1} \big(x[i] - x[i+1]\big)^2 + x[n]^2.$$

Note that we have $0 \preceq \mathbf{A} \preceq 4\mathbf{I}$, we can define

$$f(\mathbf{x}) = \frac{1}{2}\frac{L_f}{4}\mathbf{x}^\top \mathbf{A} \mathbf{x} - \frac{L_f}{4}\mathbf{x}^\top \mathbf{e}_1 = \frac{L_f}{8}\mathbf{x}^\top \mathbf{A} \mathbf{x} - \frac{L_f}{4}\mathbf{x}^\top \mathbf{e}_1.$$

Nesterov's worst function ... (2/2)

- We have : $n = 2k + 1$ and

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & & \dots & \\ -1 & 2 & -1 & \dots & \\ & -1 & 2 & \dots & \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ & & & \dots & 2 \end{bmatrix}, \quad f(\mathbf{x}) = \frac{L_f}{8}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \frac{L_f}{4}\mathbf{x}^\top \mathbf{e}_1.$$

- The minimizer of f can be computed by Fermat's rule / first order optimality criterion : set $\nabla f(\mathbf{x}) = 0$, which gives $\mathbf{A}\mathbf{x}^* = \mathbf{e}_1$. As $\mathbf{A}$ is non-singular so $\mathbf{x}_i^* = \mathbf{A}^{-1}\mathbf{e}_1$, or equivalently

$$x^*[i] = 1 - \frac{i}{n+1}.$$

The optimal cost value is

$$f(\mathbf{x}^*) = \frac{-L_f}{8}\left(1 - \frac{1}{n+1}\right). \tag{4}$$

The proof of Theorem 2 ... (1/3)

- Let $n = 2k + 1$ and let f be the Nesterovs worst function.
- Using similar argument as in Theorem 1, we have

$$\mathbf{x}_k \in \text{span}\Big\{\mathbf{e}_1, \ldots, \mathbf{e}_{k-1}\Big\},$$

by assuming $\mathbf{x}_0 = 0$. We now have $x_k[i] = 0$ for $i \geq k$ for any Bbp.

- Let $\mathbf{A}_k$ be the first k rows and columns of $\mathbf{A}$. Let

$$\mathbf{x}_k^* = \underset{\mathbf{x}, x[i]=0, i \geq k}{\text{argmin}} \; f(\mathbf{x}),$$

where $\mathbf{x}_k^*$ is the solution of a smaller $k \times k$ system formed by $\mathbf{A}_k$, so we have $x_k^*[i] = 1 - \dfrac{i}{k+1}$ if $i < k$, and $x_k^*[i] = 0$ otherwise. This give cost value

$$f(\mathbf{x}_k^*) = \frac{-L_f}{8}\big(1 - \frac{1}{k+1}\big). \tag{5}$$

The proof of Theorem 2 ... (2/3)

- We now have

$$\begin{aligned} f(\mathbf{x}_k) - f(\mathbf{x}^*) &\geq f(\mathbf{x}_k^*) - f(\mathbf{x}^*) \\ &\overset{(4),(5)}{=} \frac{-L_f}{8}\left(1 - \frac{1}{k+1}\right) - \frac{-L_f}{8}\left(1 - \frac{1}{n+1}\right) \\ &= \frac{L_f}{8}\left(\tfrac{1}{k+1} - \frac{1}{n+1}\right) \\ &\overset{n=2k+1}{=} \frac{L_f}{8}\left(\frac{1}{k+1} - \frac{1}{2k+1+1}\right) = \frac{L_f}{8}\frac{1}{2(k+1)}. \end{aligned}$$

- Now we bound $R_0 = \|\mathbf{x}_0 - \mathbf{x}^*\|_2$

$$\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 = \|0 - \mathbf{x}^*\|_2^2 = \|\mathbf{x}^*\|_2^2 \overset{(4)}{=} \sum_{i=1}^{n}\left(1 - \frac{i}{n+1}\right)^2.$$

After some algebra, we have

$$\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 \leq \frac{n+1}{3} = \frac{2(k+1)}{3}.$$

or equivalently

$$k + 1 \geq \frac{3}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2$$

The proof of Theorem 2 ... (3/3)

- What we have so far

$$\begin{aligned} f(\mathbf{x}_k) - f(\mathbf{x}^*) &\geq \frac{L_f}{8}\frac{1}{2(k+1)} \\ k+1 &\geq \frac{3}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 = \frac{3}{2}R_0^2 \end{aligned}$$

- Finish the proof

$$\begin{aligned} f(\mathbf{x}_k) - f(\mathbf{x}^*) &\geq \frac{L_f}{8}\frac{1}{2(k+1)} \\ &= \frac{L_f}{8}\frac{1}{2(k+1)}\frac{k+1}{k+1} \\ &= \frac{L_f}{8}\frac{1}{2(k+1)^2}(k+1) \\ &\geq \frac{L_f}{8}\frac{1}{2(k+1)^2}\frac{3}{2}R_0^2 \\ &= \frac{3L_f R_0^2}{32(k+1)^2}. \quad \square \end{aligned}$$

Last page - summary

The lower bounds on convergence rate of gradient-based method on minimizing a convex f

	f is convex	f is strongly convex
f is L-Lipschitz	$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$	$\mathcal{O}\left(\frac{1}{k}\right)$
f is L-smooth	$\mathcal{O}\left(\frac{1}{k^2}\right)$	$\mathcal{O}\left(\frac{1}{k^2}\right)$

We didn't show the proofs for the cases in the second column, the proofs are similar.

End of document