Sandwich theorem for β-smooth convex function

Andersen Ang

ECS, Uni. Southampton, UK
andersen.ang@soton.ac.uk

Homepage angms.science

Version: April 6, 2023
First draft: June 6, 2017

Content

$$\frac{1}{2\beta}\|\nabla f(\boldsymbol{x})-\nabla f(\boldsymbol{y})\|_2^2 \;\leq\; f(\boldsymbol{y})-f(\boldsymbol{x})-\nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x}) \;\leq\; \frac{\beta}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2.$$

Convex β-smooth function

- Setup: given a function $f : \mathbb{R}^n \to \mathbb{R}$ that is differentiable
- f is convex if f and $\mathrm{dom} f$ $(= \mathbb{R}^n)$ are convex.
- f is β-smooth if ∇f is β-Lipschitz. I.e., for any $\boldsymbol{a}, \boldsymbol{b} \in \mathrm{dom} f$, we have a constant β such that:

$$\|\nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{b})\| \leq \beta \|\boldsymbol{a} - \boldsymbol{b}\|.$$

  Details here.

- **Theorem**. If f is convex β-smooth, then $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$\frac{1}{2\beta}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2 \;\leq\; \left| f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) \right| \;\leq\; \frac{\beta}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2. \tag{1}$$

  - $f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) =$ 1st-order Taylor series of f at $\boldsymbol{x}$
  - $f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x})$ is the Bregman divergence (error made by the approximation)

  It means: for a convex β-smooth f, its 1st-order Taylor approximation has bounded error.

- This PDF: prove (1).

Prerequisite: directional derivative

- Given a function $f : \mathbb{R}^n \to \mathbb{R}$ as $f(\boldsymbol{x}) = f(x_1, x_2, ..., x_n)$.
  - We are in simple Euclidean space $\mathbb{R}^n$
  - Input of f is a vector $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ in $\mathbb{R}^n$
  - Output of f is a scalar in $\mathbb{R}$

- **Directional derivative** of f along a vector $\boldsymbol{v} = (v_1, v_2, ..., v_n)$, denoted as $D_{\boldsymbol{v}} f(\boldsymbol{x})$, is defined as

$$D_{\boldsymbol{v}} f(\boldsymbol{x}) \coloneqq \lim_{h \to 0} \frac{f(\boldsymbol{x} + h\boldsymbol{v}) - f(\boldsymbol{x})}{h}.$$

  - $\boldsymbol{v}$ specifies a direction where $\boldsymbol{x}$ is moving, and h is the stepsize
  - $\boldsymbol{v}$ is possibly not a unit vector

- **Theorem**: let $\nabla : \mathbb{R} \to \mathbb{R}^n$ be a scalar to vector mapping and let ∇f be the gradient of f, then

$$D_{\boldsymbol{v}} f(\boldsymbol{x}) = \nabla f(\boldsymbol{x})^\top \boldsymbol{v}.$$

(Point,direction) to (point,point) form

$$D_{\boldsymbol{v}} f(\boldsymbol{x}) \coloneqq \lim_{h \to 0} \frac{f(\boldsymbol{x} + h\boldsymbol{v}) - f(\boldsymbol{x})}{h} = \nabla f(\boldsymbol{x})^\top \boldsymbol{v}$$

- Now instead of $(\boldsymbol{x}, \boldsymbol{v})$, we are given two points $(\boldsymbol{x}, \boldsymbol{y})$. Let $\boldsymbol{v} = \boldsymbol{y} - \boldsymbol{x}$ be the direction pointing from $\boldsymbol{x}$ to $\boldsymbol{y}$. The directional derivative is then

$$D_{\boldsymbol{v}} f(\boldsymbol{x}) = D_{\boldsymbol{y}-\boldsymbol{x}} f(\boldsymbol{x}) \coloneqq \lim_{h \to 0} \frac{f(\boldsymbol{x} + h(\boldsymbol{y} - \boldsymbol{x})) - f(\boldsymbol{x})}{h} = \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}).$$

- Now let $h \in [0, 1]$ and we consider the line segment $[\boldsymbol{x}, \boldsymbol{y}]$
  - At one end we have $\boldsymbol{x} = \boldsymbol{x} + 0 \cdot (\boldsymbol{y} - \boldsymbol{x})$ $\quad h = 0$
  - At another end we have $\boldsymbol{y} = \boldsymbol{x} + 1 \cdot (\boldsymbol{y} - \boldsymbol{x})$ $\quad h = 1$

  This means
  - $D_{\boldsymbol{v}} f(\boldsymbol{x}) = D_{\boldsymbol{v}} \big(\boldsymbol{x} + h(\boldsymbol{y} - \boldsymbol{x})\big)\Big|_{h=0} = \nabla f\big(\boldsymbol{x} + h(\boldsymbol{y} - \boldsymbol{x})\big)^\top (\boldsymbol{y} - \boldsymbol{x})\Big|_{h=0}$
  - $D_{\boldsymbol{v}} f(\boldsymbol{y}) = D_{\boldsymbol{v}} \big(\boldsymbol{x} + h(\boldsymbol{y} - \boldsymbol{x})\big)\Big|_{h=1} = \nabla f\big(\boldsymbol{x} + h(\boldsymbol{y} - \boldsymbol{x})\big)^\top (\boldsymbol{y} - \boldsymbol{x})\Big|_{h=1}$

- Why do these? Because now we have

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) = \int_0^1 \nabla f\big(\boldsymbol{x} + h(\boldsymbol{y} - \boldsymbol{x})\big)^\top (\boldsymbol{y} - \boldsymbol{x}) dh$$

  to prove the upper bound in the sandwitch theorem.

Proving the upper bound $\left|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x})\right| \leq \frac{\beta}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2$ ... 1/3

- Start with the integral

$$\begin{aligned}
f(\boldsymbol{y}) - f(\boldsymbol{x}) &= \int_0^1 \nabla f\big(\boldsymbol{x} + h(\boldsymbol{y}-\boldsymbol{x})\big)^\top(\boldsymbol{y}-\boldsymbol{x})dh \\
&= \int_0^1 \Big[\nabla f\big(\boldsymbol{x} + h(\boldsymbol{y}-\boldsymbol{x})\big) + \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\Big]^\top(\boldsymbol{y}-\boldsymbol{x})dh \\
&= \int_0^1 \Big[\nabla f\big(\boldsymbol{x} + h(\boldsymbol{y}-\boldsymbol{x})\big) - \nabla f(\boldsymbol{x})\Big]^\top(\boldsymbol{y}-\boldsymbol{x})dh + \int_0^1 \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x})dh \\
&= \int_0^1 \Big[\nabla f\big(\boldsymbol{x} + h(\boldsymbol{y}-\boldsymbol{x})\big) - \nabla f(\boldsymbol{x})\Big]^\top(\boldsymbol{y}-\boldsymbol{x})dh + \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x}) \underbrace{\int_0^1 dh}_{=1}
\end{aligned}$$

- Now we have

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x}) = \int_0^1 \Big[\nabla f\big(\boldsymbol{x} + h(\boldsymbol{y}-\boldsymbol{x})\big) - \nabla f(\boldsymbol{x})\Big]^\top(\boldsymbol{y}-\boldsymbol{x})dh.$$

Compare it with the bound we see that we put absolute value sign next.

Proving the upper bound $\left|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x})\right| \le \frac{\beta}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2$ ... 2/3

- Put absolute value sign on

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x}) = \int_0^1 \left[\nabla f\big(\boldsymbol{x}+h(\boldsymbol{y}-\boldsymbol{x})\big) - \nabla f(\boldsymbol{x})\right]^\top (\boldsymbol{y}-\boldsymbol{x})dh$$

gives

$$0 \le \left|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x})\right| = \left|\int_0^1 \left[\nabla f\big(\boldsymbol{x}+h(\boldsymbol{y}-\boldsymbol{x})\big) - \nabla f(\boldsymbol{x})\right]^\top (\boldsymbol{y}-\boldsymbol{x})dh\right|$$

- By $\left|\int_a^b f(x)dx\right| \le \int_a^b |f(x)|dx$

$$0 \le \left|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x})\right| \le \int_0^1 \left|\left[\nabla f\big(\boldsymbol{x}+h(\boldsymbol{y}-\boldsymbol{x})\big) - \nabla f(\boldsymbol{x})\right]^\top (\boldsymbol{y}-\boldsymbol{x})\right|dh$$

- By $|\boldsymbol{a}^\top\boldsymbol{b}| \le \|\boldsymbol{a}\|\cdot\|\boldsymbol{b}\|$

$$\begin{aligned} 0 \le \left|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x})\right| &\le \int_0^1 \left\|\nabla f\big(\boldsymbol{x}+h(\boldsymbol{y}-\boldsymbol{x})\big) - \nabla f(\boldsymbol{x})\right\| \cdot \|\boldsymbol{y}-\boldsymbol{x}\|dh \\ &= \|\boldsymbol{y}-\boldsymbol{x}\| \int_0^1 \left\|\nabla f\big(\boldsymbol{x}+h(\boldsymbol{y}-\boldsymbol{x})\big) - \nabla f(\boldsymbol{x})\right\|dh \end{aligned}$$

Proving the upper bound $\left|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x})\right| \;\leq\; \frac{\beta}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2$ ... 3/3

- f is β-smooth $\iff$ $\|\nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{b})\| \leq \beta\|\boldsymbol{a}-\boldsymbol{b}\|$, so

$$0 \;\leq\; \left|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x})\right| \;\leq\; \|\boldsymbol{y}-\boldsymbol{x}\| \int_0^1 \beta\big\|h(\boldsymbol{y}-\boldsymbol{x})\big\|dh \;\leq\; \beta\big\|\boldsymbol{y}-\boldsymbol{x}\big\|^2 \int_0^1 |h|dh$$

- $\int_0^1 |t|dt = \frac{1}{2}$

$$0 \;\leq\; \left|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x})\right| \;\leq\; \frac{\beta}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2.$$

- f is convex $\iff$ the Bregman divergence $f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x})$ is nonnegative so

$$0 \;\leq\; f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x}) \;\leq\; \frac{\beta}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2. \quad \square \tag{2}$$

  Note: if f is non-convex, $0 \leq f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x})$ does not hold and we need the absolute sign.

- (2) gives

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) \;\leq\; \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x}) + \frac{\beta}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2. \tag{\#}$$

Proving the lower bound $\frac{1}{2\beta}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2 \;\leq\; f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x})$ ... 1/2

$$
\begin{aligned}
f(\boldsymbol{y}) - f(\boldsymbol{x}) &= f(\boldsymbol{z}) - f(\boldsymbol{x}) - \big[f(\boldsymbol{z}) - f(\boldsymbol{y})\big] && \text{not so trivial} \\
&\geq \nabla f(\boldsymbol{x})^\top(\boldsymbol{z}-\boldsymbol{x}) - \big[f(\boldsymbol{z}) - f(\boldsymbol{y})\big] && f \text{ is cvx} \\
&\geq \nabla f(\boldsymbol{x})^\top(\boldsymbol{z}-\boldsymbol{x}) - \Big[\nabla f(\boldsymbol{y})^\top(\boldsymbol{z}-\boldsymbol{y}) + \frac{\beta}{2}\|\boldsymbol{z}-\boldsymbol{y}\|_2^2\Big] && \text{by } (\#) \\
&= \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x}+\boldsymbol{z}-\boldsymbol{y}) - \nabla f(\boldsymbol{y})^\top(\boldsymbol{z}-\boldsymbol{y}) - \frac{\beta}{2}\|\boldsymbol{z}-\boldsymbol{y}\|_2^2 \\
&= \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x}) + \Big(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\Big)^\top(\boldsymbol{z}-\boldsymbol{y}) - \frac{\beta}{2}\|\boldsymbol{z}-\boldsymbol{y}\|_2^2
\end{aligned}
$$

We now have

$$\Big(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\Big)^\top(\boldsymbol{z}-\boldsymbol{y}) - \frac{\beta}{2}\|\boldsymbol{z}-\boldsymbol{y}\|_2^2 \;\leq\; f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x})$$

Compare with the bound, we need to form $\frac{1}{2\beta}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2$ in the left-hand-side

Proving the lower bound $\frac{1}{2\beta}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2 \;\leq\; f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x})$ ... 2/2

$$\Big(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\Big)^\top(\boldsymbol{z}-\boldsymbol{y}) - \frac{\beta}{2}\|\boldsymbol{z}-\boldsymbol{y}\|_2^2 \;\leq\; f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x})$$

When we introduce $\boldsymbol{z}$, $\boldsymbol{z}$ is free so we can set any value for $\boldsymbol{z}$.

A very tricky step: let $\boldsymbol{z} = \boldsymbol{y} - \frac{1}{\beta}\big(\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\big)$

$$\begin{aligned}
\boldsymbol{z}-\boldsymbol{y} &= -\frac{1}{\beta}\big(\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\big) \\
\big(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\big)^\top(\boldsymbol{z}-\boldsymbol{y}) &= \frac{1}{\beta}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2 \\
-\frac{\beta}{2}\|\boldsymbol{z}-\boldsymbol{y}\|_2^2 &= -\frac{1}{2\beta}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2 \\
\big(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\big)^\top(\boldsymbol{z}-\boldsymbol{y}) - \frac{\beta}{2}\|\boldsymbol{z}-\boldsymbol{y}\|_2^2 &= \frac{1}{2\beta}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|^2 \quad \square
\end{aligned}$$

Application - proving gradient is coercive

- Consider the lower bound of the theorem on $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{y}, \boldsymbol{x}$

$$\frac{1}{2\beta}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|^2 \;\leq\; f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{y} - \boldsymbol{x})$$

$$\frac{1}{2\beta}\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|^2 \;\leq\; f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^\top(\boldsymbol{x} - \boldsymbol{y})$$

  Sum them up

$$\frac{1}{\beta}\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|^2 \;\leq\; \big(\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\big)^\top(\boldsymbol{y} - \boldsymbol{x}) \qquad (\dagger)$$

  this inequality is called co-coercivity of gradient.

- Remark: apply Cauchy-Schwartz inequality $|ab| \leq |a||b|$ on (†) gives the Lipschitz continuity of ∇f: $\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\| \leq \beta\|\boldsymbol{y} - \boldsymbol{x}\|$.
  So f is β-smooth $\iff$ ∇f is Lipschitz continuous $\iff$ ∇f is coercive.

Last page - summary

1. For a β-smooth f, $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom} f$,

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

2. Sandwich theorem for β-smooth convex function: if f is convex and β-smooth, $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom} f$,

$$\frac{1}{2\beta} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2 \;\leq\; f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) \;\leq\; \frac{\beta}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2.$$

   Implication: 1st-order Taylor series has bounded error.

3. Application: proving gradient is co-coercive

$$\frac{1}{\beta} \|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|^2 \;\leq\; \Big(\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\Big)^\top (\boldsymbol{y} - \boldsymbol{x}).$$

End of document