Bregman Divergence

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

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

First draft: June 6, 2017
Last update: November 25, 2020

Two equivalent definitions of convexity

A (differentiable) function $f(x)$ is convex if :

- $\mathrm{dom}\ f$ is a convex set, and
- Two inequalities
  - Jensen's inequality:

    $$f\big(\lambda y + (1-\lambda)x\big) \le \lambda f(y) + (1-\lambda) f(x),$$

    $\forall x, y \in \mathrm{dom}\ f$, $0 \le \lambda \le 1$
  - the 1st-order Taylor approximation of f at the point x is a global under-estimator of f for all y:

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

    $\forall x, y \in \mathrm{dom}\ f$.
- **Theorem**. The two inequalities are equivalent

$$f\big(\lambda y + (1-\lambda)x\big) \leq \lambda f(y) + (1-\lambda) f(x) \implies f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

Proof. From $f\big(\lambda y + (1-\lambda)x\big) \leq \lambda f(y) + (1-\lambda) f(x)$, rearrange

$$\begin{aligned} f\big(x + \lambda(y-x)\big) &\leq f(x) + \lambda\big(f(y) - f(x)\big) \\ \iff \quad \frac{f\big(x + \lambda(y-x)\big) - f(x)}{\lambda} &\leq f(y) - f(x) \\ \iff \quad f(y) &\geq \frac{f\big(x + \lambda(y-x)\big) - f(x)}{\lambda} + f(x). \end{aligned}$$

Let $g(\lambda) = f\big(x + \lambda(y-x)\big)$, then $f(x) = g(0)$ and

$$f(y) \geq \frac{g(\lambda) - g(0)}{\lambda} + f(x).$$

Note that we only replace one $f(x)$ by $g(0)$. Take limit $\lambda \to 0$ on both side

$$f(y) \geq \lim_{\lambda \to 0} \frac{g(\lambda) - g(0)}{\lambda} + f(x) = \nabla_\lambda g(\lambda)\Big|_{\lambda=0} + f(x) \tag{1}$$

For $g(\lambda) = f\big(x + \lambda(y-x)\big)$, by chain rule,

$$\begin{aligned} \nabla_\lambda g(\lambda) &= \Big\langle \nabla_x f\big(x + \lambda(y-x)\big), y - x \Big\rangle \\ \nabla_\lambda g(\lambda)\Big|_{\lambda=0} &= \Big\langle \nabla_x f(x), y - x \Big\rangle. \end{aligned} \tag{2}$$

Put (2) into (1) finishes the proof. □

$$f\big(\lambda y + (1-\lambda)x\big) \leq \lambda f(y) + (1-\lambda) f(x) \iff f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

Proof. Let $z = \lambda x + (1-\lambda)y$, $0 \leq \lambda \leq 1$, so $\forall x, y \in \mathrm{dom}\ f$, $z \in \mathrm{dom}\ f$.
By $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$,

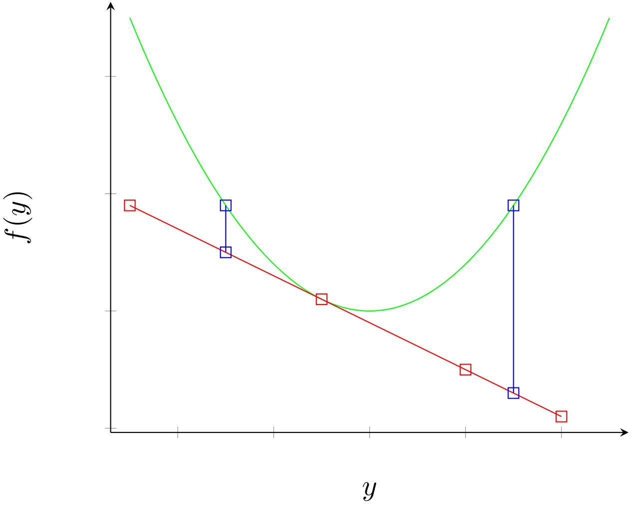
$$\begin{aligned} f(x) &\geq f(z) + \langle \nabla f(z), x - z \rangle &(3)\\ f(y) &\geq f(z) + \langle \nabla f(z), y - z \rangle &(4)\end{aligned}$$

Consider $\lambda(3) + (1-\lambda)(4)$

$$\lambda f(x) + (1-\lambda) f(y) \geq f(z) + \underbrace{\Big\langle \nabla f(z), \lambda(x - z) + (1-\lambda)(y - z) \Big\rangle}_{=0 \text{ for } z=\lambda x + (1-\lambda)y}. \quad (5)$$

Put $z = \lambda x + (1-\lambda)y$ into (5) finishes the proof. $\square$

1st-order Taylor approximation of a convex function is an global under-estimator : $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$.
Consider $e(x) := f(y) - f(x) - \nabla f(x)^\top (y - x)$, which is the error made by the Taylor approximation (the blue lines).
This $e(x)$ is the Bregman divergence.



Bregman Divergence B

For a (strictly) convex (and differentiable) function f, the Bregman divergence between two point p, $q \in \mathrm{dom}\ f$ is defined as :

$$B(p,q) = f(p) - f(q) - \langle \nabla f(q), p - q \rangle$$

- $B \geq 0$, it is nonnegative (as shown in previous diagram)
- $B(p,q)$ is convex in p (in q may not!)
- $B(p,q)$ is linear
- $B(p,q) \neq B(q,p)$
- $B(p,q)$ does not satisfy the triangle inequality

Other more advanced properties

- $B_f(p,q) = B_{f^*}\big(\nabla f(p), \nabla f(q)\big)$, f^* is convex conjugate of f
- (Banerjee et al. 2005) given a random vector, the mean vector minimizes the expected distance from the random vector. That distance is a Bregman divergence.

Other issues of B

Given a function f (not necessarily differentiable), the generalized Bregman divergence between two point p, $q \in \mathrm{dom}\ f$ is defined as:

$$B(p, q) = f(p) - f(q) - \langle g, p - q \rangle$$

where g is a sub-gradient of f at q such that $f(p) \geq f(q) + \langle g, p - q \rangle \ \forall p$

Example of B: squared L_2 norm of difference between two vectors

$$\begin{aligned}
\|x - y\|_2^2 &= \|x\|_2^2 + \|y\|_2^2 - 2x^\top y \\
&= \|y\|_2^2 - \|x\|_2^2 - 2x^\top y + 2\|x\|_2^2 \\
&= \|y\|_2^2 - \|x\|_2^2 - 2x^\top y + 2x^\top x \\
&= \|y\|_2^2 - \|x\|_2^2 - 2x^\top (y - x) \\
&= f(y) - f(x) - 2\nabla f(x)(y - x)
\end{aligned}$$

where $f(x) = \|x\|_2^2$. Hence $\|x - y\|_2^2$ is the B of $\|x\|_2^2$.
Other example: negative entropy with KL divergence as B.

Last page - summary

- Equivalent definitions of convexity
  - $f\big(\lambda y + (1-\lambda)x\big) \leq \lambda f(y) + (1-\lambda) f(x)$, $\forall x, y \in \text{dom } f$, $0 \leq \lambda \leq 1$
  - $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$, $\forall y \in \text{dom } f$
- Bregman Divergence as the gap between 1st order Taylor approximation and original function
- $\|x-y\|_2^2$ is the Bregman divergence of $\|x\|_2^2$

End of document