Bregman Divergence = 1st-order Taylor error

Andersen Ang
U.Southampton UK
angms.science

January 23, 2025

1st draft
June 6, 2017

Content

Contents

Two equivalent definitions of convexity

- Setup: given a function $f : \mathbb{R}^n \to \mathbb{R}$ that is convex and differentiable

- If f is convex, then
  - $\mathrm{dom} f (= \mathbb{R}^n)$ is a convex set, and
    Jensen's inequality: $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom} f$, $\lambda \in [0, 1]$

$$f\big(\lambda \boldsymbol{y} + (1-\lambda)\boldsymbol{x}\big) \;\leq\; \lambda f(\boldsymbol{y}) + (1-\lambda) f(\boldsymbol{x}). \qquad (\#)$$

    1st-order Taylor series of f at $\boldsymbol{x}$ is a global under-estimator: $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom} f$

$$f(\boldsymbol{y}) \;\geq\; f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle. \qquad (\dagger)$$

- **Theorem**. The $(\#)$ and $(\dagger)$ are equivalent. I.e., $(\#) \iff (\dagger)$.

- This PDF: prove this theorem and show $(\dagger)$ is the Bregman divergence of f.

- You want more definitions then click here

$$f\big(\lambda \boldsymbol{y} + (1-\lambda)\boldsymbol{x}\big) \le \lambda f(\boldsymbol{y}) + (1-\lambda) f(\boldsymbol{x}) \quad \Longrightarrow \quad f(\boldsymbol{y}) \ge f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$$

- **Proof**. Starting from Jansen inequality $f\big(\lambda \boldsymbol{y} + (1-\lambda)\boldsymbol{x}\big) \le \lambda f(\boldsymbol{y}) + (1-\lambda) f(\boldsymbol{x})$

$$\begin{aligned}
& f\big(\lambda \boldsymbol{y} + (1-\lambda)\boldsymbol{x}\big) && \le && \lambda f(\boldsymbol{y}) + (1-\lambda) f(\boldsymbol{x}) \\
\Longleftrightarrow \quad & f\big(\boldsymbol{x} + \lambda(\boldsymbol{y} - \boldsymbol{x})\big) && \le && f(\boldsymbol{x}) + \lambda\Big(f(\boldsymbol{y}) - f(\boldsymbol{x})\Big) \\
\Longleftrightarrow \quad & f\big(\boldsymbol{x} + \lambda(\boldsymbol{y} - \boldsymbol{x})\big) - f(\boldsymbol{x}) && \le && \lambda\Big(f(\boldsymbol{y}) - f(\boldsymbol{x})\Big) \\
\Longleftrightarrow \quad & \frac{f\big(\boldsymbol{x} + \lambda(\boldsymbol{y} - \boldsymbol{x})\big) - f(\boldsymbol{x})}{\lambda} && \le && f(\boldsymbol{y}) - f(\boldsymbol{x}) \\
\Longleftrightarrow \quad & f(\boldsymbol{y}) && \ge && f(\boldsymbol{x}) + \frac{f\big(\boldsymbol{x} + \lambda(\boldsymbol{y} - \boldsymbol{x})\big) - f(\boldsymbol{x})}{\lambda}.
\end{aligned}$$

- For convex function $\lambda \in [0, 1]$ includes 0 so we are allowed to take $\lambda \to 0$
  Take limit $\lambda \to 0$ on both side and by definition of directional derivative

$$f(\boldsymbol{y}) \;\ge\; f(\boldsymbol{x}) + \lim_{\lambda \to 0} \frac{f\big(\boldsymbol{x} + \lambda(\boldsymbol{y} - \boldsymbol{x})\big) - f(\boldsymbol{x})}{\lambda} \;=\; f(\boldsymbol{x}) + \Big\langle \nabla f(\boldsymbol{x}),\; \boldsymbol{y} - \boldsymbol{x} \Big\rangle.$$

The big picture of the proof of $\Longrightarrow$

f is convex $\Longrightarrow$ Jansen inequality $\boxed{f\big(\lambda \boldsymbol{y} + (1-\lambda)\boldsymbol{x}\big) \leq \lambda f(\boldsymbol{y}) + (1-\lambda) f(\boldsymbol{x})}$

do algebra $\Updownarrow$

$$\boxed{f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \frac{f\big(\boldsymbol{x} + \lambda(\boldsymbol{y}-\boldsymbol{x})\big) - f(\boldsymbol{x})}{\lambda}}$$

Jansen for $\lambda \in [0,1]$ includes the case $\lim_{\lambda \to 0}$ $\Big\downarrow$

$$\boxed{f(\boldsymbol{y}) \;\geq\; f(\boldsymbol{x}) + \lim_{\lambda \to 0} \frac{f\big(\boldsymbol{x} + \lambda(\boldsymbol{y}-\boldsymbol{x})\big) - f(\boldsymbol{x})}{\lambda}}$$

directional derivative $\Updownarrow$

$$\boxed{f(\boldsymbol{y}) \;\geq\; f(\boldsymbol{x}) + \Big\langle \nabla f(\boldsymbol{x}),\; \boldsymbol{y} - \boldsymbol{x} \Big\rangle.}$$

$$f\big(\lambda \boldsymbol{y} + (1-\lambda)\boldsymbol{x}\big) \leq \lambda f(\boldsymbol{y}) + (1-\lambda) f(\boldsymbol{x}) \quad \Longleftarrow \quad f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$$

Proof Let $\boldsymbol{z} = \lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}$ with $0 \leq \lambda \leq 1$.
so $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom} f$ we have $\boldsymbol{z} \in \mathrm{dom} f$ because $\mathrm{dom} f$ is a convex set.

By convexity of f:

$$f(\boldsymbol{x}) \geq f(\boldsymbol{z}) + \langle \nabla f(\boldsymbol{z}), \boldsymbol{x} - \boldsymbol{z} \rangle \tag{1}$$
$$f(\boldsymbol{y}) \geq f(\boldsymbol{z}) + \langle \nabla f(\boldsymbol{z}), \boldsymbol{y} - \boldsymbol{z} \rangle \tag{2}$$

Consider $\lambda(1) \ + \ (1-\lambda)(2)$

$$\lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) \ \geq \ f(\boldsymbol{z}) + \underbrace{\Big\langle \nabla f(\boldsymbol{z}), \lambda(\boldsymbol{x} - \boldsymbol{z}) + (1-\lambda)(\boldsymbol{y} - \boldsymbol{z}) \Big\rangle}_{=0 \text{ for } \boldsymbol{z} = \lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}}. \tag{3}$$

Put $\boldsymbol{z} = \lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}$ into (3) finishes the proof. □

Contents

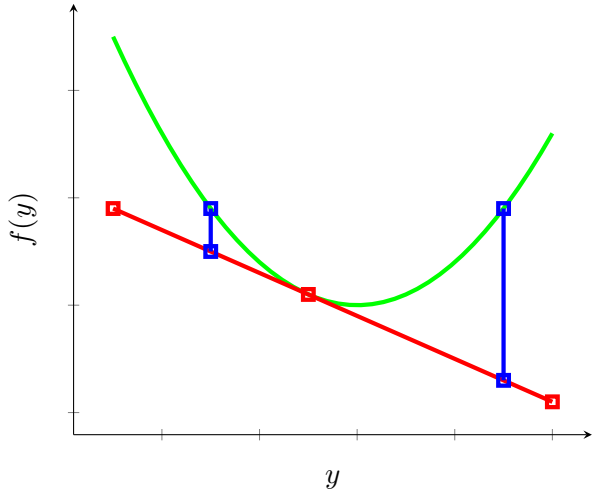
Bregman divergence

1st-order Taylor series of a convex function is an global under-estimator:

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x).$$

$$\underbrace{f(y) - f(x) - \nabla f(x)^\top (y - x)}_{=:B(y)}$$

$B(y) =$ Taylor series error



Lev .M. Bregman, The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming, USSR Computational Mathematics and Mathematical Physics Volume 7, Issue 3, 1967, Pages 200 - 217

Bregman divergence B

$f : \mathbb{R}^n \to \mathbb{R}$ **is convex and differentiable**

- The Bregman divergence at a point $\boldsymbol{p}$ of f anchored at $\boldsymbol{q}$, denoted as $B_f(\boldsymbol{p}, \boldsymbol{q})$, is

$$B_f(\boldsymbol{p}; \boldsymbol{q}) \;\coloneqq\; f(\boldsymbol{p}) - f(\boldsymbol{q}) - \langle \nabla f(\boldsymbol{q}), \boldsymbol{p} - \boldsymbol{q} \rangle.$$

  - $\boldsymbol{p} \in \mathrm{dom} f$ and $\boldsymbol{q} \in \mathrm{dom} f$ have different meaning here
  - $\boldsymbol{q}$ is the "anchor" (where B_f touch f)
  - $\boldsymbol{p}$ is the variable of B

- Why study B_f: it is closely related to convexity of f

- Bregman divergence is used in the design of Bregman method for convex optimization.

Contents

Example of Bregman divergence: ℓ_2-squared norm

Let $f(\boldsymbol{z}) = \|\boldsymbol{z}\|_2^2$.

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

Hence $\|\boldsymbol{x}-\boldsymbol{y}\|_2^2$ is the Bregman divergence of $\|\boldsymbol{x}\|_2^2$.

Example of Bregman divergence: negative Shannon entropy

- Let $f(\boldsymbol{x}) = \sum_i x_i \log x_i$ Shannon entropy is $-f(\boldsymbol{x})$
- We focus on $\boldsymbol{x} \in \Delta := \{\boldsymbol{v} \mid \langle \boldsymbol{v}, \mathbf{1} \rangle = 1\}$ probability simplex because entropy works on probability
- Fact: $\dfrac{\partial}{\partial x} x \log x = \log x + 1$
- Showing Bregman divergence of negative Shannon entropy is KL divergence

$$
\begin{aligned}
B_f(\boldsymbol{x}; \boldsymbol{y}) &= f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y}) \\
&= \sum x_i \log x_i - \sum y_i \log y_i - \sum (\log y_i + 1)(x_i - y_i) \\
&= \sum x_i \log x_i - \sum (\log y_i + 1)(x_i) \\
&= \sum x_i \log x_i - \sum x_i \log y_i + \underbrace{\sum x_i}_{=1} \\
&= \sum x_i (\log x_i - \log y_i) \\
&= \sum x_i \log \frac{x_i}{y_i}
\end{aligned}
$$

Hence $\sum_i x_i \log \dfrac{x_i}{y_i}$ is the Bregman divergence of $\sum_i x_i \log x_i$.

Example of Bregman divergence: matrix entropy

- Let $f(\boldsymbol{X}) = \mathrm{Tr}(\boldsymbol{X} \log \boldsymbol{X})$, then $B_F(\boldsymbol{X}; \boldsymbol{Y}) = \mathrm{Tr}(\boldsymbol{X} \log \boldsymbol{X} - \boldsymbol{X} \log \boldsymbol{Y} - \boldsymbol{X} + \boldsymbol{Y})$
  - This is basically eigen-value version of the vector Shannon entropy
  - This is called von Neumann divergence

- Let $f(\boldsymbol{X}) = -\log \det(\boldsymbol{X})$, then $B_F(\boldsymbol{X}; \boldsymbol{Y}) = \log \det(\boldsymbol{Y}) - \log \det(\boldsymbol{X}) + \mathrm{Tr}(\boldsymbol{Y}^{-1}(\boldsymbol{X} - \boldsymbol{Y}))$
  - This is called logdet divergence

Contents

Facts about $B_f(\boldsymbol{p}; \boldsymbol{q}) \coloneqq f(\boldsymbol{p}) - f(\boldsymbol{q}) - \langle \nabla f(\boldsymbol{q}), \boldsymbol{p} - \boldsymbol{q} \rangle$ ... 1/2

- B_f does not satisfy the triangle inequality
- B_f is asymmetric: $B_f(\boldsymbol{p}; \boldsymbol{q}) \neq B_f(\boldsymbol{q}; \boldsymbol{p})$
- $B_f \geq 0$ because of convexity of f
- $B_f(\boldsymbol{p}; \boldsymbol{q})$ is convex in $\boldsymbol{p}$ not $\boldsymbol{q}$
- B is conic: $B_{\alpha f + \beta g} = \alpha B_f + \beta B_g$ for positive α, β NOT linear!
- B is affine invariant: $B_{f+g} = B_f$ if g is affine (e.g., $g(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{c}$)
- If f is not differentiable, replace $\nabla f(\boldsymbol{q})$ by any subgradient of g at $\boldsymbol{q}$.
- $\nabla_{\boldsymbol{p}} B_f(\boldsymbol{p}; \boldsymbol{q}) = \nabla f(\boldsymbol{p}) - \nabla f(\boldsymbol{q})$
- For $\min f$, if $\boldsymbol{q} = \operatorname{argmin} f$, then $B_f(\boldsymbol{p}; \boldsymbol{q}) = f(\boldsymbol{p}) - f^*$

Facts about $B_f(\boldsymbol{p}; \boldsymbol{q}) \coloneqq f(\boldsymbol{p}) - f(\boldsymbol{q}) - \langle \nabla f(\boldsymbol{q}), \boldsymbol{p} - \boldsymbol{q} \rangle$ ... 2/2

- **Conjugate duality**
  Let f^* be the convex conjugate of f. Then $B_f(\boldsymbol{p}; \boldsymbol{q}) = B_{f^*}\big(\nabla f(\boldsymbol{p}); \nabla f(\boldsymbol{q})\big)$
- (Banerjee et al. 2005)[1] given a random vector $\boldsymbol{z}$, the mean vector minimizes the expected distance from $\boldsymbol{z}$ is a Bregman divergence.
- Bregman projection
  - Given a convex set $\mathcal{C}$, the Bregman projection of a point $\boldsymbol{x}$ onto $\mathcal{C}$ under the Bregman divergence with anchor $\boldsymbol{x}_0$ is
    $$\boldsymbol{x}^* = \operatorname*{argmin}_{\boldsymbol{x} \in \mathcal{C}} B_f(\boldsymbol{x}; \boldsymbol{x}_0)$$
  - **Generalised Pythagorean theorem**
    $$B_f(\boldsymbol{y}; \boldsymbol{x}_0) \geq B_f(\boldsymbol{y}; \boldsymbol{x}^*) + B_f(\boldsymbol{x}^*; \boldsymbol{x}_0)$$
  - Bregman projection is the key of mirror descent.

[1] Banerjee, Gou, and Wang, On the optimality of conditional expectation as a Bregman predictor, 2005

Contents

Last page - summary

- Convexity: $f\big(\lambda \boldsymbol{y} + (1-\lambda)\boldsymbol{x}\big) \leq \lambda f(\boldsymbol{y}) + (1-\lambda) f(\boldsymbol{x}) \iff f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$
- Bregman divergence: $B_f(\boldsymbol{y}; \boldsymbol{x}) \coloneqq f(\boldsymbol{y}) - f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$
  - $\boldsymbol{y}$ is variable
  - $\boldsymbol{x}$ is anchor
- Bregman Divergence = the gap between 1st-order Taylor series and the original function
- $\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$ is the Bregman divergence of $\|\boldsymbol{x}\|_2^2$
- Some basic properties of Bregman divergence (proof to be added in future version)

End of document