*CIMPA Research School : Data Science for Engineering and Technology*
*Tunis 2019*

# Bregman divergences
# a basic tool for pseudo-metrics building for data structured by physics

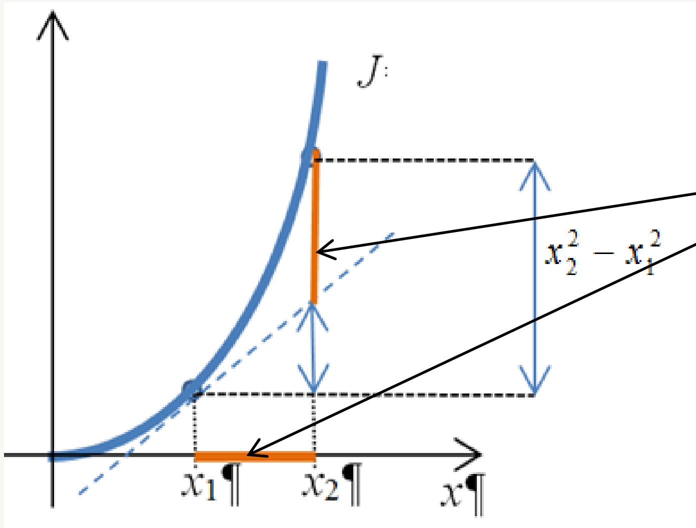
## 2- The Bregman divergence

Stéphane ANDRIEUX

*ONERA - France*

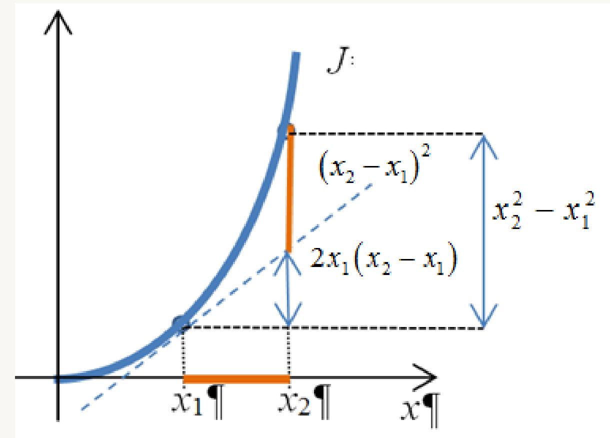
*Member of the National Academy of Technologies of France*

# The basic idea



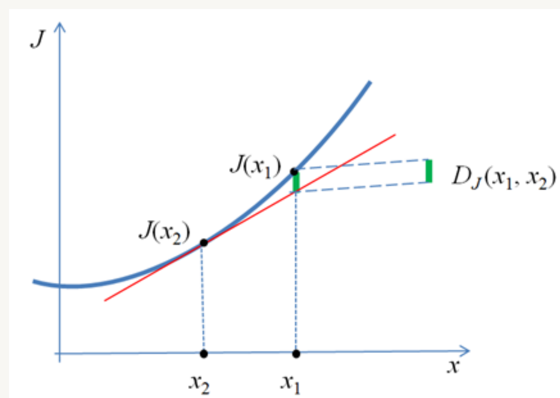

Take $J(x)=x^2$

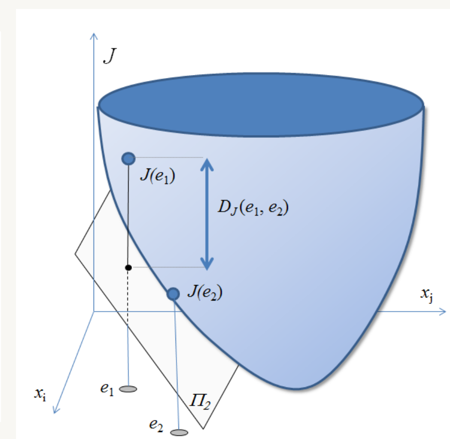
Calculate




**Definition**: Bregman divergence

*Let J be a convex differentiable function, the Bregman divergence generated by J between $e_1$ and $e_2$ ($\in$ dom J), is the non-negative quantity:*




$$D_J(e_1, e_2) = J(e_1) - J(e_2) - \langle \nabla J(e_2), e_1 - e_2 \rangle$$

Not symmetric

No triangle inequality

# First properties of the Bregman divergence

## Why is it a positive quantity ?

By definition of convexity and differentiability , $J$ lies above its tangents

$$J(y) \geq J(x) + \langle \nabla J(x), x - y \rangle$$

Definition of subdifferential $\partial J(e) = \{ p,\ J(d) \geq J(e) + \langle p, d - e \rangle\ \forall d \in dom(J) \}$

## What if $J$ is affine ?

$$D_{ax+b}(e_1,e_2)=0$$

## What if $D_J(e_1,e_2)=0$ and $J$ strictly convex?

By contradiction, suppose $e_1 \neq e_2$ , for any $0<\lambda<1$

$$\begin{aligned} D_J(e,e_2) &= D_J(\lambda e_1 + (1-\lambda)e_2, e_2) \\ &< \lambda D_J(e_1,e_2) + (1-\lambda) D_J(e_2,e_2) = 0 \end{aligned}$$

## Is $D_J(e_1,e_2)$ separately convex ?

$D_J(x,.)$ is $J(x)$+ affine function, hence is convex

$D_J(.,x)$ is not always convex

Counter example $J(x)=x^3$ on $\mathrm{IR}^+$

# First properties of the Bregman divergence (*cont.*)

What if $J$ is quadratic (in IR$^n$) with associated matrix $A$ ?

$$J(x) = x^t A x$$ A symmetric positive

$$DJ(x_1, x_2) = (x_1 - x_2)^t A (x_1 - x_2)$$

Mahalanobis distance

What is $D_{\lambda J + \mu F}$ ?
$(J, F)$ convex functions
$(\lambda, \mu)$ positive scalars

$$D_{\lambda J + \mu F}(e_1, e_2) = \lambda D_J(e_1, e_2) + \mu D_F(e_1, e_2)$$

How is related $D_J$ to $D_{\tilde{J}}$ ?

$$\tilde{J}(e) = J(e) - J(0) - \langle \nabla J(0), e \rangle$$

What is $D_{\tilde{J}}(e,0)$

$$D_{\tilde{J}} = D_J$$ Generating function differing by an affine function

$$D_{\tilde{J}}(e,0) = \tilde{J}(e)$$

# Examples of Bregman divergences

| Domain | Generating function $J(x)$ | Bregman divergence $D_J(x, y)$ | Name |
| --- | --- | --- | --- |
| $IR^n$ | $\|x\|^2$ | $\|x-y\|^2$ | Euclidian Distance |
| $IR^n$ | $J(x) = x^T Ax \quad A$ *symmetric positive* | $(x-y)^T A(x-y)$ | Mahalanobis distance |
| $IR^{+*n}$ | $\sum x_i \log x_i - x_i$ | $\sum x_i \log \frac{x_i}{y_i} - x_i + y_i$ | Kullback–Leibler divergence or Relative Entropy |
| $IR^{+*n}$ | $\sum -\log x_i$ | $\sum \frac{x_i}{y_i} - \log \frac{x_i}{y_i} - 1$ | Itakura-Saito discrete distance |
| $[0,1]$ | $x \log x + (1-x)\log(1-x)$ | $x \log \frac{x}{y} + (1-x)\log \frac{1-x}{1-y}$ | Logistic loss |

Used in learning (speech recognition, image classification, stochastic clustering, …)

# Extensions of Bregman divergences

## Non differentiable generating functions

When $J$ is not differentiable at point $e_2$, the definition would lead to a multivoque function, since the subdifferential of $J$ in $e_2$ is not reduced to a singleton

**Definition:** Extended Bregman Divergences

*Let J be a convex, not necessarily differentiable function, the extended Bregman divergences and generated by J between $e_1$ and $e_2$ ($\in$ dom J), are the non-negative quantities:*

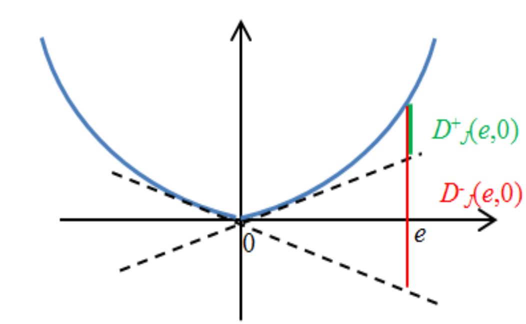
$$D^{+}{}_{J}(e_1,e_2)=\min_{p\,\in\,\partial J(e_2)} J(e_1)-J(e_2)-\langle p,e_1-e_2\rangle \equiv J(e_1)-J(e_2)-\langle \overline{p}_2,e_1-e_2\rangle$$

$$D^{-}{}_{J}(e_1,e_2)=\max_{p\,\in\,\partial J(e_2)} J(e_1)-J(e_2)-\langle p,e_1-e_2\rangle \equiv J(e_1)-J(e_2)-\langle \underline{p}_2,e_1-e_2\rangle$$

with

$$\overline{p}_2=\operatorname*{arg\,min}_{p_2\,\in\,\partial J(e_2)} J(e_1)-J(e_2)-\langle p_2,e_1-e_2\rangle=\operatorname*{arg\,max}_{p_2\,\in\,\partial J(e_2)}\langle p_2,e_1-e_2\rangle$$

$$\underline{p}_2=\operatorname*{arg\,max}_{p_2\,\in\,\partial J(e_2)} J(e_1)-J(e_2)-\langle p_2,e_1-e_2\rangle=\operatorname*{arg\,min}_{p_2\,\in\,\partial J(e_2)}\langle p_2,e_1-e_2\rangle$$




Extended Bregman Divergences for $J(x)=\alpha x^2+|x|$

The subdifferential is a closed convex set
the minimum and maximum exist
argmin and argmax belong to its boundary

$$0\le D^{+}{}_{J}(e_1,e_2)\le D^{-}{}_{J}(e_1,e_2)$$

# Symmetrized Bregman divergences (I)

## Characterization of Symmetric Bregman Divergences

The Bregman Divergences are generally not symmetric

$$D_J(e_1,e_2) = J(e_1) - J(e_2) - \langle \nabla J(e_2), e_1 - e_2 \rangle \neq D_J(e_2,e_1) = J(e_2) - J(e_1) - \langle \nabla J(e_1), e_2 - e_1 \rangle$$

Only Bregman Divergences generated by a quadratic function $J$ are symmetric and they also enjoy the triangle inequality (sub-additivity). They reduce then to Mahalanobis distances

**Property:** Characterization of symmetrical Bregman divergences

*Let J be a strictly convex function, third differentiable on* $IR^n$*, the Bregman divergence generated by J is symmetrical* $D_J(e_1, e_2) = D_J(e_1, e_2)$*, if and only if J is the sum of a quadratic Q(e) and a linear function L(e). Furthermore* $D_J \equiv D_Q$*, and* $D_Q$ *satisfies the triangle inequality*

Using $2\big(J(e_1) - J(e_2)\big) = \langle \nabla J(e_1) + \nabla J(e_2), e_1 - e_2 \rangle$ for any $e_1$=$e$ and $e_2$=0 and $J(0)$=0 $\quad 2J(e) = \langle \nabla J(0) + \nabla J(e), e \rangle \ \forall e$

Deriving $\nabla J(e) = \nabla J(0) + \langle \nabla\nabla J(e), e \rangle$

Replacing in to the symmetry condition $\quad J(e) = \langle \nabla J(0), e \rangle + \frac{1}{2} \langle \nabla\nabla J(e).e, e \rangle \ \forall e$

Deriving again $\langle \nabla\nabla\nabla J(e).e.e, e \rangle = 0 \ \forall e \rightarrow J(e) = L(e) + Q(e)$ , $L(e) = \langle \nabla J(0), e \rangle$ , $Q(e) = \frac{1}{2} \langle \nabla\nabla J(0).e, e \rangle$

# Symmetrized Bregman divergences (II)

## Two notions of Symmetrized Bregman Divergences

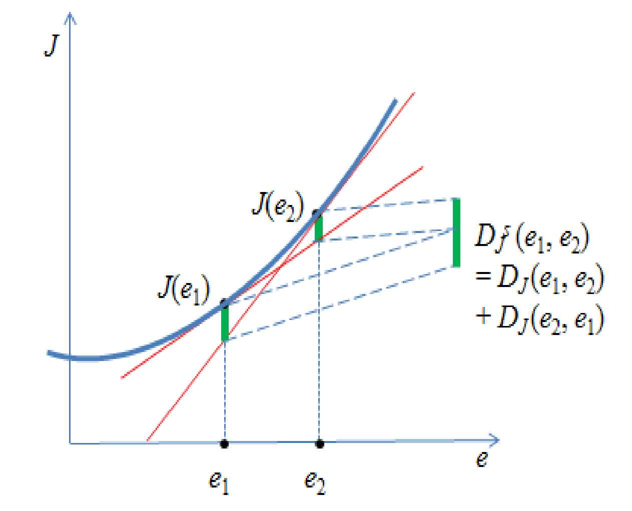
The more intuitive symmetrization is to define the symmetrized Bregman Divergences as

$$D_J^s(e_1,e_2) = D_J(e_1,e_2) + D_J(e_2,e_1)$$

**Definition:** Symmetrized Bregman divergence

*Let J be a convex differentiable function, the symmetrized Bregman divergence generated by J between $e_1$ and $e_2$ ($\in$ dom J), is the non-negative quantity:*

$$D_J^s(e_1,e_2) = \langle \nabla J(e_1) - \nabla J(e_2), e_1 - e_2 \rangle$$



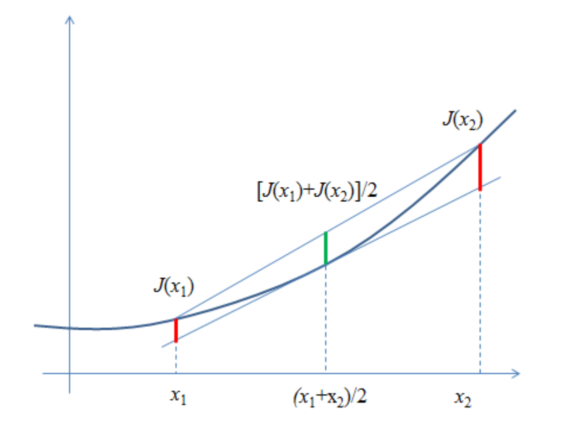

But other definitions exist

**Definition:** Jensen-Bregman divergence

*The Jensen-Bregman divergence generated by the strictly convex function J, is:*

$$JB_J(x,y) = D_J\left(x, \frac{x+y}{2}\right) + D_J\left(y, \frac{x+y}{2}\right)$$

$$\frac{1}{2} JB_J(x,y) = \frac{J(x)+J(y)}{2} - J\left(\frac{x+y}{2}\right)$$



# Symmetrized Bregman divergences (III)

## Natural notion of Symmetrized Bregman Divergences

Calculate the following symmetrized Bregman Divergences

| Domain | Generating function $J(x)$ | Name | Symmetrized Bregman Divergence $D^s_J(x,y)$ |
|---|---|---|---|
| $IR^{+*}$ | $\sum x_i \log x_i - x_i$ | Symmetric Kullback–Leibler | $\sum (\log x_i - \log y_i, x_i - y_i)$ |
| $IR^{+*}$ | $\sum -\log x_i$ | Symmetric Itakura-Saito | $\sum \frac{(x_i - y_i)^2}{x_i y_i}$ |
| $[0,1]$ | $x \log x + (1-x)\log(1-x)$ | Symmetric loss function | $(x-y)\log \frac{x(1-y)}{y(1-x)}$ |

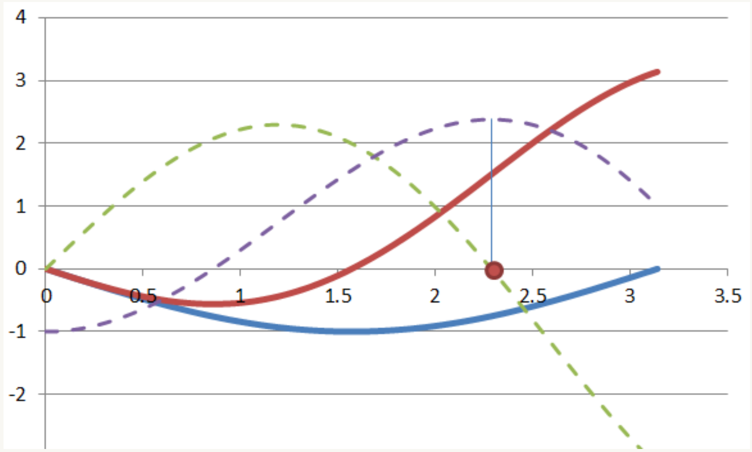
But, the symmetrized Bregman divergence, as a function of $(e_1,e_2)$ is generally **not** separately convex

C. Ex. $J(x) = -\sin x$ Convex on $[0,\pi]$

$$D_J(x,0) = \nabla J(x).x = -x\cos x$$

Convex only on $[0,\beta\pi]$ with

$$2\sin\beta\pi + \beta\pi\cos\beta\pi = 0$$



# Bregman Gaps

## Divergences for pairs of dual variables

When manipulating data from physics, one can have to deal with data pairs constituted by dual variables $(e,p)$, such that the duality product $\langle p, e \rangle$ is for example a work or a power.

Ex : Stress and strain $(\underline{\underline{\sigma}}, \underline{\underline{\varepsilon}}) \rightarrow \langle \underline{\underline{\sigma}}, \underline{\underline{\varepsilon}} \rangle = \underline{\underline{\sigma}} : \underline{\underline{\varepsilon}}$

Flux and Temperature $(\underline{q}, \underline{\nabla T}) \rightarrow \langle (\underline{q}, \underline{\nabla T}) \rangle = \underline{q} . \underline{\nabla T}$

**Definition:** Bregman gap

*Let $J$ be a convex, not necessarily differentiable function, the Bregman gap $BG_J$ generated by $J$ between $e_1$ and the pair of dual quantities $(e_2, p_2)$, $p_2 \in \partial J(e_2)$, is the non-negative quantity:*

$$BG_J\left(e_1,[e_2,p_2]\right) = J(e_1) - J(e_2) - \langle p_2, e_1 - e_2 \rangle$$

**Definition:** Symmetrized Bregman gap

*The Symmetrized Bregman gap generated by the convex function $J$ between the two pairs of dual quantities $(e_1, p_1)$ and $(e_2, p_2)$, , is the nonnegative scalar :*

$$BG_J^s\left([e_1,p_1],[e_2,p_2]\right) = BG_J\left(e_1,[e_2,p_2]\right) + BG_J\left(e_2,[e_1,p_1]\right)$$

# Properties of Bregman Gaps

**1-** Separate convexity of the symmetrized Bregman gap

$$\forall\left([e_1,p_1],[e_2,p_2],[e_0,p_0]\right)$$
$$BG_J^s\left(\lambda[e_1,p_1]+(1-\lambda)[e_2,p_2],[e_0,p_0]\right)\leq\lambda BG_J^s\left([e_1,p_1],[e_0,p_0]\right)+(1-\lambda)BG_J^s\left([e_2,p_2],[e_0,p_0]\right)??$$

Consider the two functions of $\lambda$:

$$F(\lambda)=\left\langle \lambda e_1+(1-\lambda)e_2-e_0,\lambda p_1+(1-\lambda)p_2-p_0\right\rangle$$
$$G(\lambda)=\lambda\left\langle e_1-e_0,p_1-p_0\right\rangle+(1-\lambda)\left\langle e_2-e_0,p_2-p_0\right\rangle$$

Show that the function *f*(*l*)=*F*(*l*)-*G*(*l*) is negative along the segment [0,1] and notice that *f*(0)=0

The derivative of *f* is $f'(\lambda)=(2\lambda-1)\left\langle e_1-e_2,p_1-p_2\right\rangle=C(2\lambda-1)\quad C\geq 0$

And *f* can be calculated as $f(\lambda)=C\lambda(\lambda-1)\left\langle e_1-e_2,p_1-p_2\right\rangle\leq 0\quad for\ \lambda\in[0,1]$

**2-** If *J* is differentiable, symmetrized Bregman gap $\equiv$ symmetrized Bregman divergence:

$$BG_J^s\left([e_1,\nabla J(e_1)],[e_2,\nabla J(e_2)]\right)\equiv D_J^s\left(e_1,e_2\right)$$

**3-** Alternative form of $BG_J^s$ $\quad BG_J^s\left([e_1,p_1],[e_2,p_2]\right)=\left\langle p_1-p_2,e_1-e_2\right\rangle$

**4-** If in addition J is quadratic then: $BG_J^s\left([e_1,p_1],[e_2,p_2]\right)=2J(e_1-e_2)$

# Symmetrized Bregman divergences & Bregman Gaps

## Non differentiable generating functions - Regularization

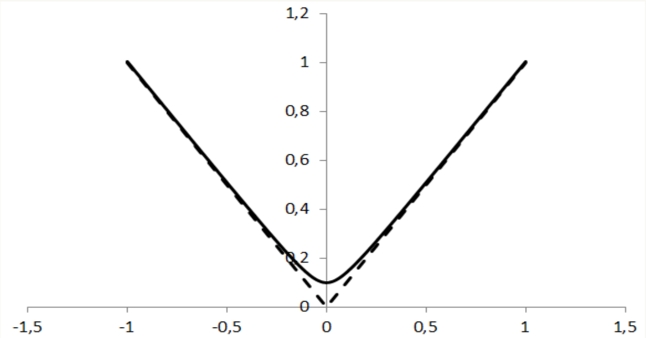
Consider the loss function used in robust statistic $J(x) = |x|$ as the generating function
(as is given rise to better robustness to outliers, *cf.* Linear Regression !)

Calculate the symmetrized Bregman divergence and the symmetrized Bregman gap generated

$$\begin{cases} BG^s_{|.|}([x, sign(x)],[y, sign(y)]) = \begin{vmatrix} 2|x-y| & if\ sign(x) \neq sign(y) \\ 0 & if\ sign(x) = sign(y) \end{vmatrix} \quad if\ |x||y| \neq 0 \\ BG^s_{|.|}([x, sign(x)],[0, p]) = (sign(x) - p)x \qquad p \in [-1,1] \end{cases}$$

$$\begin{cases} D^s_{|.|}(x,y) = \begin{vmatrix} 2|x-y| & if\ sign(x) \neq sign(y) \\ 0 & if\ sign(x) = sign(y) \end{vmatrix} \quad for |x||y| \neq 0 \\ D^s_{|.|}(x,0) = 0 \end{cases}$$

What if one use the regularized version of the loss function $J_\varepsilon(x) = \sqrt{x^2 + \varepsilon^2}$, limit when $\varepsilon \to 0$ ?




Hinge loss and regularized hinge loss ($\varepsilon$=0.1)

$$D^s_{J_\varepsilon} = BG^s_{J_\varepsilon} = \left( \frac{x}{\sqrt{x^2 + \varepsilon^2}} - \frac{y}{\sqrt{y^2 + \varepsilon^2}}, x - y \right)$$

$$D^s_{J_0}(x,0) = \frac{x^2}{\sqrt{x^2}} = sign(x)x \qquad (p_\varepsilon(0) = 0\ \forall \varepsilon)$$

# Thanks for your attention