Learning Set-to-Set Bregman Digergence with Permutation-Invariant NNs

Takahiro Kawashima (ZOZO Research)

@ RIMS-ML 2026

May 28, 2026

ZOZONEXT

Introduction of Introduction

Who?

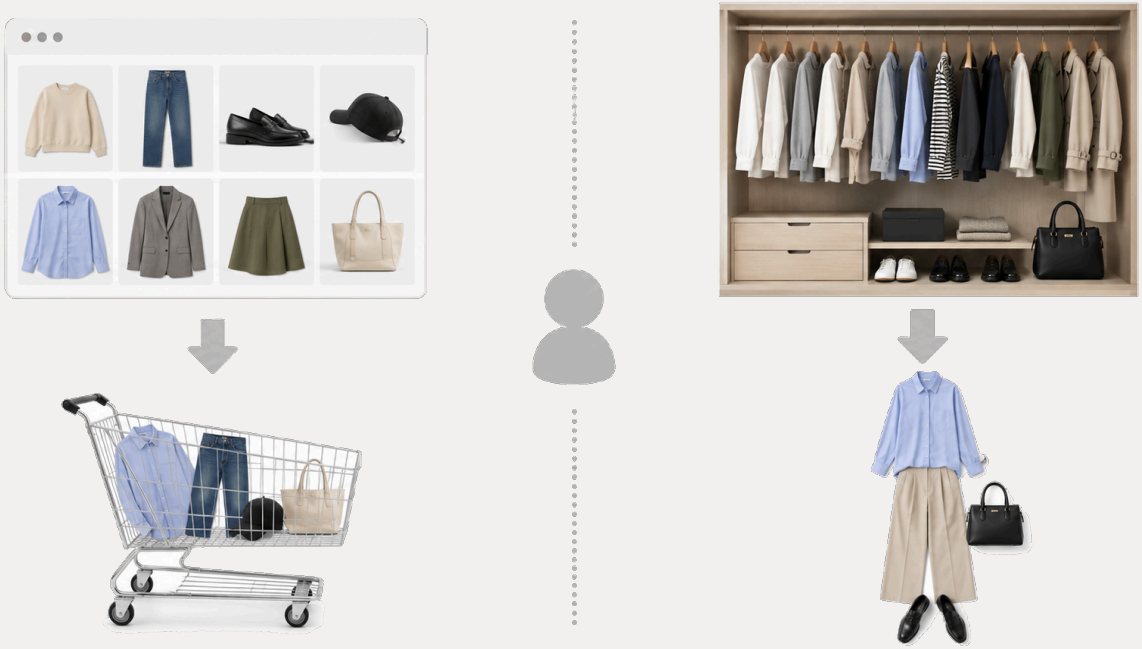
- Name: **Takahiro Kawashima** （川島貴大）
- Interests: Statistical ML, Probabilistic Modeling,
  Bayesian Methods
- 2012-2017: Kobe City College of Tech.（神戸高専）
- 2017-2021: Univ. of Electro-Communications (B.Eng. & M.Eng.)
- 2021-2024: SOKENDAI (Ph.D. in Statistical Science)
- 2019-2024: National Center of Neurology and Psychiatry
- 2024-Present: **ZOZO Research** (ZOZO NEXT, Inc.)

About ZOZO Research

- Mission: “**Quantifying Fashion**”（ファッションを数値化する）
- R&D division of ZOZO
- Focusing on the field of **Fashion Tech**
  - Including CV, NLP, Recommender Systems, HCI, Stats, OR, . . .

Set-valued Data in Fashion Tech

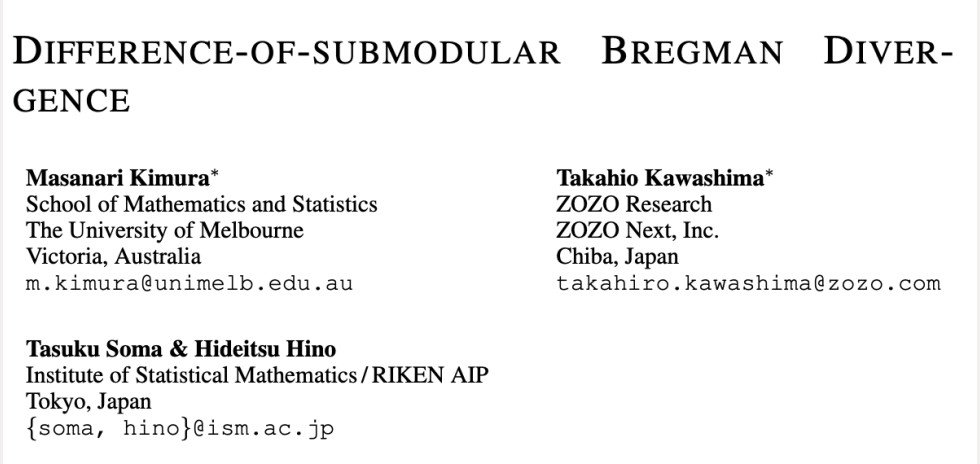
- **Set-valued data** often appear in Fashion Tech
  - Buying some set of items, choosing outfits, . . .



We also focus on
Set-data × AI

Today's Talk

- Novel, flexible, and learnable set-to-set discrepancy metric
  - Based on the idea of **Bregman divergence**
- Published in ICLR 2025
- Worked with Masanari Kimura (Univ. Melbourne), Tasuku Soma (ISM), Hideitsu Hino (ISM & Waseda Univ.)



DIFFERENCE-OF-SUBMODULAR BREGMAN DIVERGENCE

Masanari Kimura*
School of Mathematics and Statistics
The University of Melbourne
Victoria, Australia
m.kimura@unimelb.edu.au

Takahio Kawashima*
ZOZO Research
ZOZO Next, Inc.
Chiba, Japan
takahiro.kawashima@zozo.com

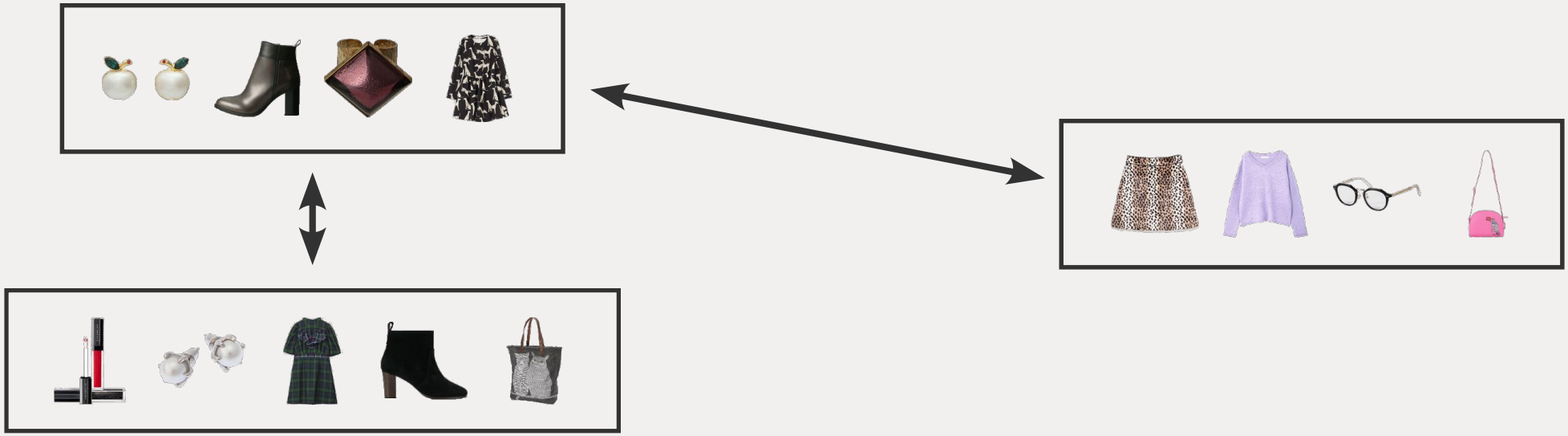
Tasuku Soma & Hideitsu Hino
Institute of Statistical Mathematics / RIKEN AIP
Tokyo, Japan
{soma, hino}@ism.ac.jp

Introduction

Distance Between Finite Sets

Most ML algs implicitly or explicitly assume a distance structure

- Rich distance metrics between sets are desired!
- Applications: outfit recommendation, point cloud proc, . . .



Distance/Similarity Metrics for Sets

Numerous metrics of set-to-set distance/similarity exist:

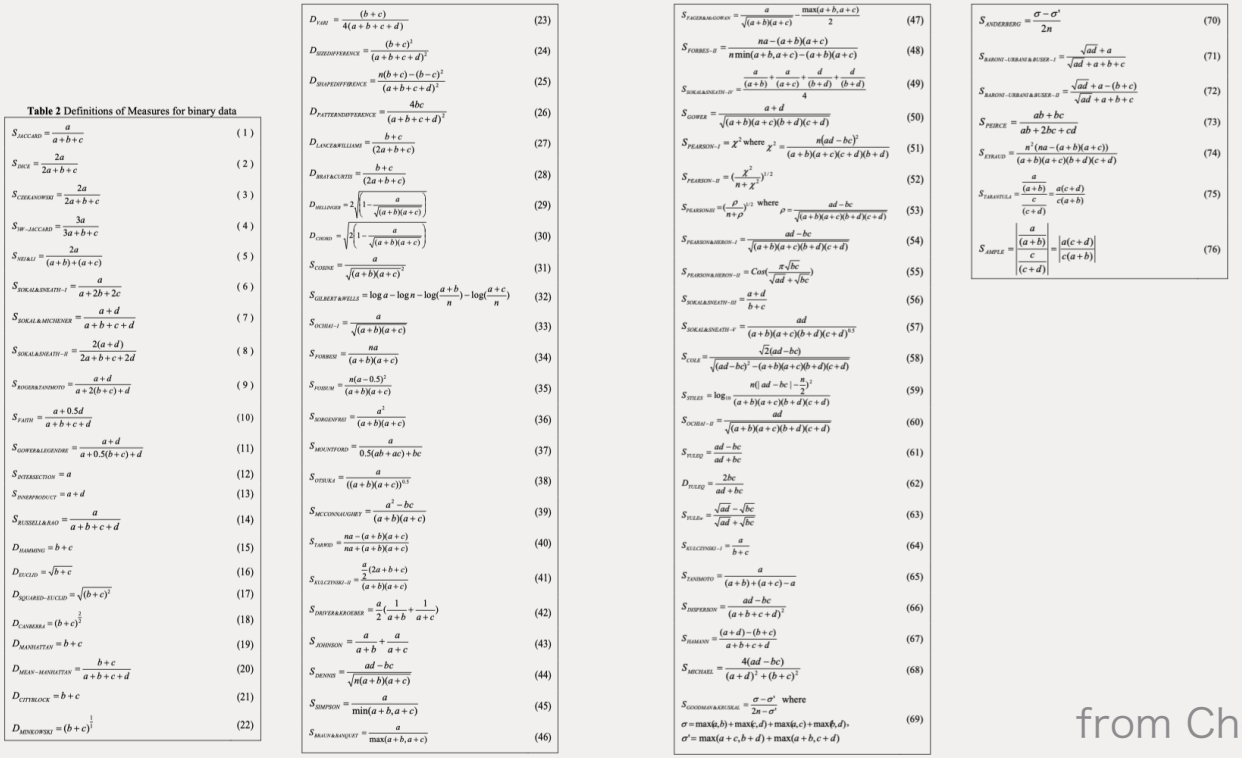


Table 2 Definitions of Measures for binary data

$$S_{JACCARD} = \frac{a}{a+b+c} \quad (1)$$

$$S_{DICE} = \frac{2a}{2a+b+c} \quad (2)$$

$$S_{CHEKANOWSKI} = \frac{2a}{2a+b+c} \quad (3)$$

$$S_{SW-JACCARD} = \frac{3a}{3a+b+c} \quad (4)$$

$$S_{SOKAL\&SNEATH-I} = \frac{2a}{(a+b)+(a+c)} \quad (5)$$

$$S_{SOKAL\&SNEATH-I} = \frac{a}{a+2b+2c} \quad (6)$$

$$S_{SOKAL\&MICHENER} = \frac{a+d}{a+b+c+d} \quad (7)$$

$$S_{SOKAL\&SNEATH-II} = \frac{2(a+d)}{2a+b+c+2d} \quad (8)$$

$$S_{ROGERS\&TANIMOTO} = \frac{a+d}{a+2(b+c)+d} \quad (9)$$

$$S_{FAITH} = \frac{a+0.5d}{a+b+c+d} \quad (10)$$

$$S_{GOWER\&LEGENDRE} = \frac{a+d}{a+0.5(b+c)+d} \quad (11)$$

$$S_{INTERSECTION} = a \quad (12)$$

$$S_{INNERPRODUCT} = a+d \quad (13)$$

$$S_{RUSSELL\&RAO} = \frac{a}{a+b+c+d} \quad (14)$$

$$D_{HAMMING} = b+c \quad (15)$$

$$D_{EUCLID} = \sqrt{b+c} \quad (16)$$

$$D_{SQUARED-EUCLID} = \sqrt{(b+c)^2} \quad (17)$$

$$D_{CANBERRA} = (b+c)^{\frac{1}{2}} \quad (18)$$

$$D_{MANHATTAN} = b+c \quad (19)$$

$$D_{MEAN-MANHATTAN} = \frac{b+c}{a+b+c+d} \quad (20)$$

$$D_{CITYBLOCK} = b+c \quad (21)$$

$$D_{MINKOWSKI} = (b+c)^{\frac{1}{2}} \quad (22)$$

$$D_{VARI} = \frac{(b+c)}{4(a+b+c+d)} \quad (23)$$

$$D_{SIZEDIFFERENCE} = \frac{(b+c)^2}{(a+b+c+d)^2} \quad (24)$$

$$D_{SHAPEDIFFERENCE} = \frac{n(b+c)-(b-c)^2}{(a+b+c+d)^2} \quad (25)$$

$$D_{PATTERNDIFFERENCE} = \frac{4bc}{(a+b+c+d)^2} \quad (26)$$

$$D_{LANCE\&WILLIAMS} = \frac{b+c}{(2a+b+c)} \quad (27)$$

$$D_{BRAY\&CURTIS} = \frac{b+c}{(2a+b+c)} \quad (28)$$

$$D_{HELLINGER} = 2\sqrt{\left(1-\frac{a}{\sqrt{(a+b)(a+c)}}\right)} \quad (29)$$

$$D_{CHORD} = \sqrt{2\left(1-\frac{a}{\sqrt{(a+b)(a+c)}}\right)} \quad (30)$$

$$S_{COSINE} = \frac{a}{\sqrt{(a+b)(a+c)}} \quad (31)$$

$$S_{GILBERT\&WELLS} = \log a - \log n - \log\left(\frac{a+b}{n}\right) - \log\left(\frac{a+c}{n}\right) \quad (32)$$

$$S_{OCHIAI-I} = \frac{a}{\sqrt{(a+b)(a+c)}} \quad (33)$$

$$S_{FORBESI} = \frac{na}{(a+b)(a+c)} \quad (34)$$

$$S_{FOSSUM} = \frac{n(a-0.5)^2}{(a+b)(a+c)} \quad (35)$$

$$S_{SORGENFREI} = \frac{a^2}{(a+b)(a+c)} \quad (36)$$

$$S_{MOUNTFORD} = \frac{a}{0.5(ab+ac)+bc} \quad (37)$$

$$S_{OTSUKA} = \frac{a}{((a+b)(a+c))^{0.5}} \quad (38)$$

$$S_{MCCONNAUGHEY} = \frac{a^2-bc}{(a+b)(a+c)} \quad (39)$$

$$S_{TARWID} = \frac{na-(a+b)(a+c)}{na+(a+b)(a+c)} \quad (40)$$

$$S_{KULCZYNSKI-II} = \frac{\frac{a}{2}(2a+b+c)}{(a+b)(a+c)} \quad (41)$$

$$S_{DRIVER\&KROEBER} = \frac{a}{2}\left(\frac{1}{a+b}+\frac{1}{a+c}\right) \quad (42)$$

$$S_{JOHNSON} = \frac{a}{a+b}+\frac{a}{a+c} \quad (43)$$

$$S_{DENNIS} = \frac{ad-bc}{\sqrt{n(a+b)(a+c)}} \quad (44)$$

$$S_{SIMPSON} = \frac{a}{\min(a+b,a+c)} \quad (45)$$

$$S_{BRAUN\&BANQUET} = \frac{a}{\max(a+b,a+c)} \quad (46)$$

$$S_{FAGE\&MCGOWAN} = \frac{a}{\sqrt{(a+b)(a+c)}} - \frac{\max(a+b,a+c)}{2} \quad (47)$$

$$S_{FORBES-II} = \frac{na-(a+b)(a+c)}{n\min(a+b,a+c)-(a+b)(a+c)} \quad (48)$$

$$S_{SOKAL\&SNEATH-IV} = \frac{\frac{a}{(a+b)}+\frac{a}{(a+c)}+\frac{d}{(b+d)}+\frac{d}{(b+d)}}{4} \quad (49)$$

$$S_{GOWER} = \frac{a+d}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \quad (50)$$

$$S_{PEARSON-I} = \chi^2 \text{ where } \chi^2 = \frac{n(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)} \quad (51)$$

$$S_{PEARSON-II} = \left(\frac{\chi^2}{n+\chi^2}\right)^{1/2} \quad (52)$$

$$S_{PEARSONIII} = \left(\frac{\rho}{n+\rho}\right)^{1/2} \text{ where } \rho = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \quad (53)$$

$$S_{PEARSON\&HERON-I} = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \quad (54)$$

$$S_{PEARSON\&HERON-II} = Cos\left(\frac{\pi\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}\right) \quad (55)$$

$$S_{SOKAL\&SNEATH-III} = \frac{a+d}{b+c} \quad (56)$$

$$S_{SOKAL\&SNEATH-V} = \frac{ad}{(a+b)(a+c)(b+d)(c+d)^{0.5}} \quad (57)$$

$$S_{COLE} = \frac{\sqrt{2}(ad-bc)}{\sqrt{(ad-bc)^2-(a+b)(a+c)(b+d)(c+d)}} \quad (58)$$

$$S_{STILES} = \log_{10}\frac{n(|ad-bc|-\frac{n}{2})^2}{(a+b)(a+c)(b+d)(c+d)} \quad (59)$$

$$S_{OCHIAI-II} = \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \quad (60)$$

$$S_{YULEQ} = \frac{ad-bc}{ad+bc} \quad (61)$$

$$D_{YULEQ} = \frac{2bc}{ad+bc} \quad (62)$$

$$S_{YULEw} = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}} \quad (63)$$

$$S_{KULCZYNSKI-I} = \frac{a}{b+c} \quad (64)$$

$$S_{TANIMOTO} = \frac{a}{(a+b)+(a+c)-a} \quad (65)$$

$$S_{DISPERSION} = \frac{ad-bc}{(a+b+c+d)^2} \quad (66)$$

$$S_{HAMANN} = \frac{(a+d)-(b+c)}{a+b+c+d} \quad (67)$$

$$S_{MICHAEL} = \frac{4(ad-bc)}{(a+d)^2+(b+c)^2} \quad (68)$$

$$S_{GOODMAN\&KRUSKAL} = \frac{\sigma-\sigma'}{2n-\sigma'} \text{ where } \sigma = \max(a,b)+\max(c,d)+\max(a,c)+\max(b,d), \sigma' = \max(a+c,b+d)+\max(a+b,c+d) \quad (69)$$

$$S_{ANDERBERG} = \frac{\sigma-\sigma'}{2n} \quad (70)$$

$$S_{BARONI-URBANI\&BUSER-I} = \frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c} \quad (71)$$

$$S_{BARONI-URBANI\&BUSER-II} = \frac{\sqrt{ad}+a-(b+c)}{\sqrt{ad}+a+b+c} \quad (72)$$

$$S_{PEIRCE} = \frac{ab+bc}{ab+2bc+cd} \quad (73)$$

$$S_{ETRAUD} = \frac{n^2(na-(a+b)(a+c))}{(a+b)(a+c)(b+d)(c+d)} \quad (74)$$

$$S_{TARANTULA} = \frac{\frac{a}{(a+b)}}{\frac{c}{(c+d)}} = \frac{a(c+d)}{c(a+b)} \quad (75)$$

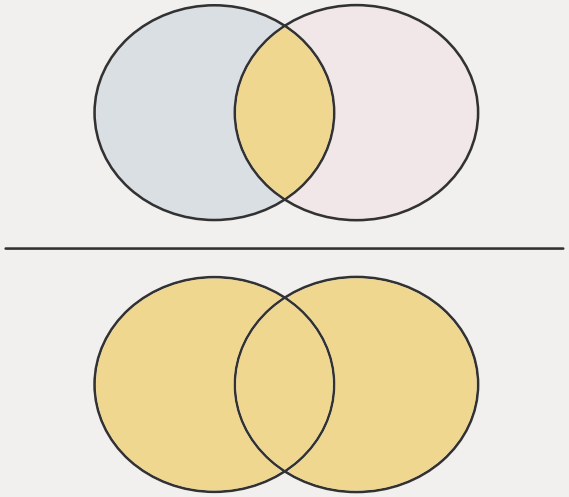
$$S_{AMPLE} = \left|\frac{\frac{a}{(a+b)}}{\frac{c}{(c+d)}}\right| = \left|\frac{a(c+d)}{c(a+b)}\right| \quad (76)$$

from Choi et al. (2010)

Difficulty of Distance Metrics for Sets

Most existing metrics are defined through counting

- such as the sizes: $|X \cap Y|, |X \cup Y|, |X \setminus Y|, \ldots$



Jaccard Index

If the ground set is large and X, Y are small, similarity/distance easily becomes 0/∞ 👎

- **Richer distance metrics** for finite sets are desired!

Our Contributions

We majorly contribute to:

- Develop a **set-to-set divergence** with richer expressiveness
  - Well-definedness and richness are theoretically guaranteed
- Propose a framework to learn the divergence with **permutation-invariant NNs**

This talk covers from necessary background to our latest work

Bregman Divergence

Divergence

Divergence: a core concept in ML&stats, measuring the discrepancy between two objects

Def: Divergence

If $D : \Omega \times \Omega \to \mathbb{R}$ satisfies:

- $D(x, y) \geq 0$ for all $x, y \in \Omega$ (nonnegativity),
- $D(x, y) = 0$ iff $x = y$ for all $x \in \Omega$ (identifiability),

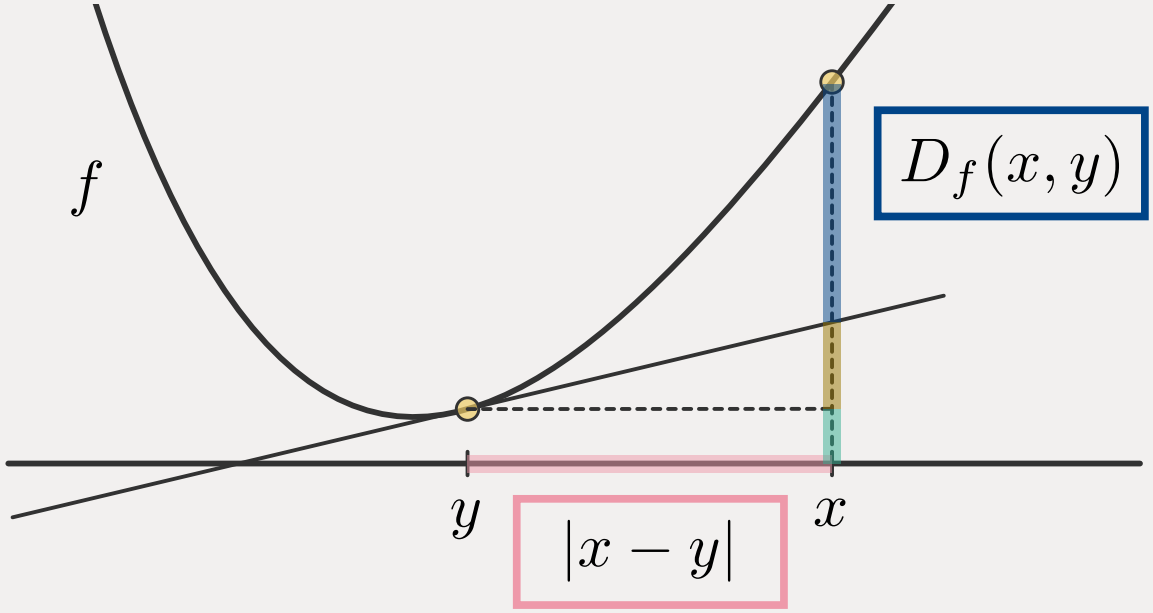
then D is a **divergence** on Ω.

› Generalization of distance; not necessarily symmetric

Bregman Divergence

Bregman divergence constructs a divergence from
a differentiable and strictly convex function f

- Every tangent plane lower-bounds f; nonneg. metric is induced



$$D_f(x, y) := f(x) - f(y) - \langle \nabla f(y), x - y \rangle$$

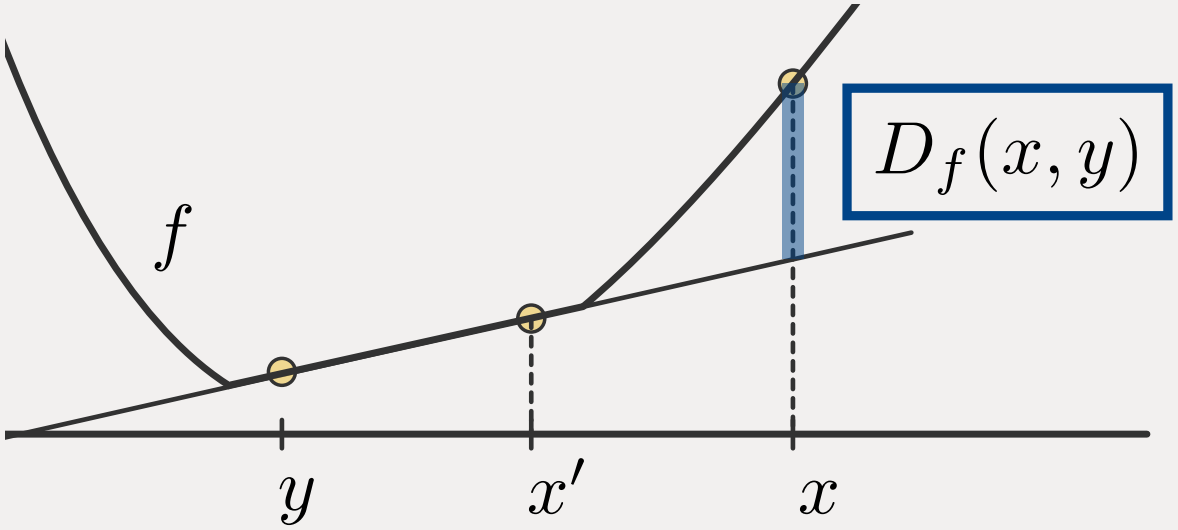
inner prod.

Generalizes KL-div., sq-L2 dist., . . .

Necessity of Strict Convexity

If f is convex but not strictly convex, $D_f(x, y)$ can be 0 for some $x \neq y$

- The identifiability condition is violated



$$D_f(x, y) := f(x) - f(y) - \langle \nabla f(y), x - y \rangle$$

inner prod.

Application of Bregman Divergence

Bregman divergence widely appears in ML&stats

- Clustering (k-means, EM, . . .)
- Bayesian Inference (variational methods, . . .)
- Optimization (mirror descent, optimal tranport, . . .)
- Information Geometry (dually flat manifolds, . . .)
- . . .

Flexible, tractable, and theoretically grounded

› Motivating the development of Bregman divergences for sets!

Existing Work: Submodular-Bregman Divergence

Generalized Bregman Divergence

For defining Bregman-like divergences, differentiability is not necessary

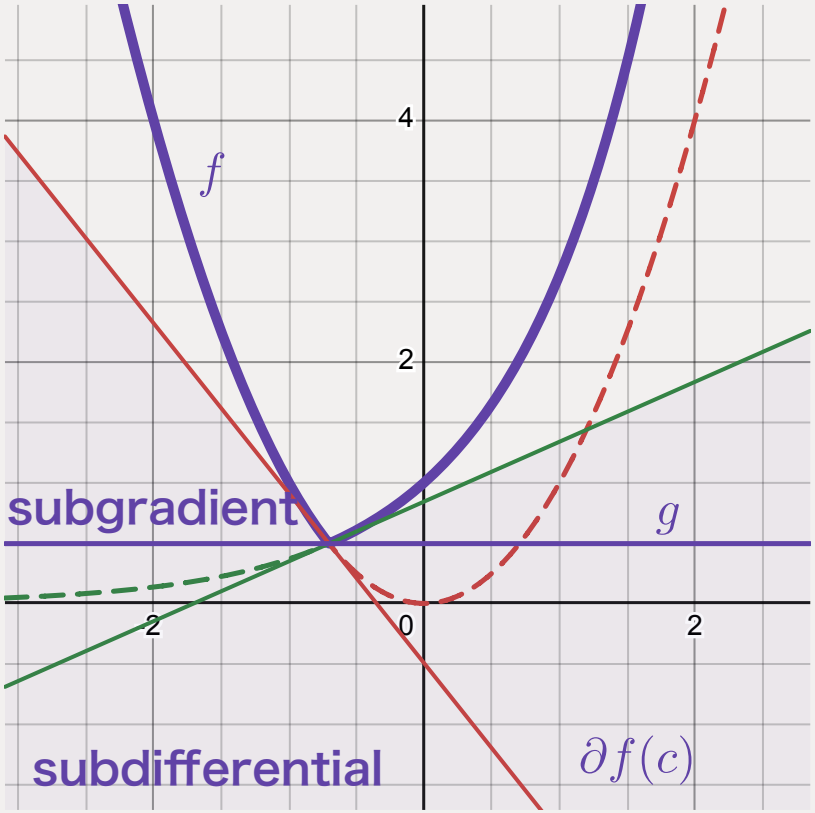
e.g., $f(x) = \max(x^2, e^x)$

- non-differentiable at $c := -0.703...$

By choosing a subgradient $g \in \partial f(c)$, a generalized Bregman divergence

$$D_{f,g}(x, c) := f(x) - f(c) - \langle g, x - c \rangle$$

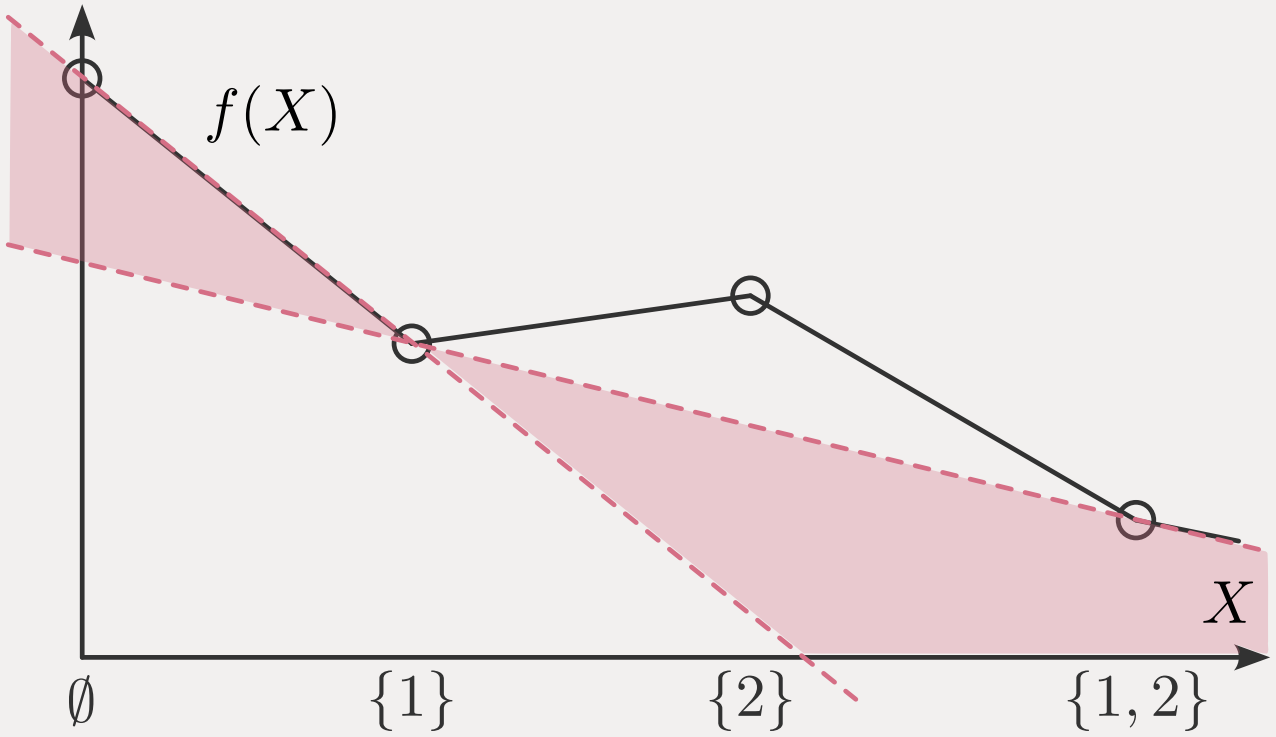
can be defined



Bregman Divergence in Discrete Space

Bregman-like divergence can be defined even in discrete spaces

- as long as a lower-bounding tangent plane (subgradient) exists



Submodular Functions

Our aim: define a Bregman-like divergence between finite sets

What type of set functions can be lower-bounded linearly?

- **Submodular functions**: discrete analog of convex functions!

> **Def: Submodular Function**
> If $f : 2^V \to \mathbb{R}$ satisfies $f(X) + f(Y) \geq f(X \cup Y) + f(X \cap Y)$ for every $X, Y \subseteq V$, then f is a **submodular function** on V.

When the inequality is strict except for $X \subseteq Y$ or $Y \subseteq X$, f is called **strictly submodular**.

Subgradients of Submodular Functions

Linear functions in the discrete space: **modular functions**

- $m : 2^V \to \mathbb{R}$ is modular $\iff$ m and $-m$ are submodular

Every modular function is represented as $m(X) = \sum_{i \in X} m(\{i\})$

- m is equated with the vector $m \in \mathbb{R}^N$, where $N = |V|$

Edmonds (1970): greedy alg. for finding the tightest modular lower-bound of submodular f at $Y \subseteq V$

Subgradients are easily found for submodular functions!

Submodular-Bregman Divergence

Iyer & Bilmes (2012): Bregman-like divergence between finite sets with a submodular function

Submodular-Bregman Divergence

$$D_f(X, Y) = f(X) - f(Y) - \langle h_Y, 1_X - 1_Y \rangle$$

h_Y : Subgradient of f at Y
1_X : Indicator vector of X ($1_X(i) = \mathbb{I}[i \in X]$)

Note: the inner product is the standard one in $\mathbb{R}^N$

Submodular-Bregman Divergence: Example

Concrete discrepancy metrics are constructed with specific f, h_Y

e.g., $f(X) = |X|, h_Y = 2 \times 1_Y$

$$
\begin{aligned}
D_f(X, Y) &= f(X) - f(Y) - \langle h_Y, 1_X - 1_Y \rangle \\
&= |X| - |Y| - 2\langle 1_Y, 1_X - 1_Y \rangle \\
&= |X| - |Y| - 2(|X \cap Y| - |Y|) \\
&= |X| + |Y| - 2|X \cap Y| \\
&= |X \setminus Y| + |Y \setminus X|
\end{aligned}
$$

› Hamming distance between sets!

Submodular-Bregman Divergence: Issues

Submodular-Bregman provides an interesting concept but. . .

- How to choose a submodular function f?
- For non-strictly submodular f, the tightest subgradient does not offer identifiability
  - To construct a divergence, a loose subgradient is needed; $f(X) = |X|, h_Y = 2 \times 1_Y$ gives $f(V) \neq h_Y(V) \Rightarrow$ unregularized

Our work: based on **strictly-submodular functions**

- Tight subgradients can be used, and increases learnability!

Our Work: Difference-of-submodular Bregman Divergence

Strict Submodularity and Divergence

Submodular-Bregman Divergence

$$D_f(X, Y) = f(X) - f(Y) - \langle h_Y, 1_X - 1_Y \rangle$$

h_Y : Subgradient of f at Y

If f is strictly submodular, the tightest h_Y offers identifiability

› properly defines a divergence!

Does this conditioning reduce the expressiveness?

In Fact . . .

Thm: Representation Power (casual)
The representation power of the divergence between finite sets defined as:

$$D_f(X, Y) = f(X) - f(Y) - \langle h_Y, 1_X - 1_Y \rangle$$

increases with the function class of f.

f : Set function
h_Y : Subgradient of f at Y

› Broader function class induces richer divergences

Broadening Representation Power

Make the representation power richer by two remarkable properties of submodular functions:

1. Existence of **Supergradients**
   - › Supergradients also exist for any submodular function
2. **Strong Difference-of-Submodular (DS) Decomposition**
   - › Every set function can be decomposed as the difference of two strictly submodular functions

Both are specific properties in the discrete space

Supergradients

Thm: Supergradients (Iyer & Bilmes, 2012)
For a submodular function f, the following modular functions are supegradients of f at Y:

$$f(j \mid X) \coloneqq f(X \cup \{j\}) - f(X)$$

Name	Notation	$j \in Y$	$j \notin Y$
grow	$\hat{g}_Y(j)$	$f(j \mid V \setminus \{j\})$	$f(j \mid Y)$
shrink	$\check{g}_Y(j)$	$f(j \mid Y \setminus \{j\})$	$f(j \mid \emptyset)$
bar	$\bar{g}_Y(j)$	$f(j \mid V \setminus \{j\})$	$f(j \mid \emptyset)$

If f is strictly submodular, all the above are strict upperbounds

Strong DS Decomposition

Thm: Strong DS Decomposition (Li & Du, 2020)
Any set function f can be decomposed as

$$f = f^1 - f^2,$$

where f^1, f^2 are strictly submodular

Continous analog: Difference-of-Convex (DC) structure

- But the function class is not broad in the continuous space

Note: Finding the decomposition may be NP-hard

Difference-of-submodular Bregman Div.

Bregman-like divergence with any set function f can be defined!

1. Strong DS decomposition: $f = f^1 - f^2$
2. Take the tightest subgradient h^1_Y for f^1 at $Y \subseteq V$
3. Take the grow/shrink/bar supergradient g^2_Y for f^2 at $Y \subseteq V$
4. Define the divergence as:

submod. Bregman of f^1

$$D_f(X,Y) := f^1(X) - f^1(Y) - \langle h^1_Y, 1_X - 1_Y \rangle$$

$$- \left(f^2(X) - f^2(Y) - \langle g^2_Y, 1_X - 1_Y \rangle\right)$$

supmod. Bregman of f^2

$$= f(X) - f(Y) - \langle h^1_Y - g^2_Y, 1_X - 1_Y \rangle$$

DBD construction with NNs

How to construct the divergence in practice?

- Learning the set function from data

Permutation-Invariant NNs

Neural networks whose output is invariant to the permutation of the input sequence; e.g.,

$$f_{\mathrm{NN}}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) = f_{\mathrm{NN}}(\boldsymbol{x}_2, \boldsymbol{x}_1, \boldsymbol{x}_3) = f_{\mathrm{NN}}(\boldsymbol{x}_3, \boldsymbol{x}_2, \boldsymbol{x}_1) = \ldots$$

Prepare two permutation-invariant NNs with strict submodularity a priori, and model f^1, f^2 with them

ε-PointNet

We also propose an architecture of permutation-invariant NNs

ε-PointNet

$$f_{\mathrm{PN},\varepsilon}\left([\boldsymbol{x}_i]_{i\in X}\right) = \gamma\left(\varepsilon\log\sum_{i\in X} e^{\varphi_1(\boldsymbol{x}_i)/\varepsilon}, \ldots, \varepsilon\log\sum_{i\in X} e^{\varphi_K(\boldsymbol{x}_i)/\varepsilon}\right)$$

(K-dim intermediate outputs of φ)

$\gamma,\ \varphi_1, \ldots, \varphi_K$: NN or fixed functions, $\varepsilon > 0$: hypeparameter

- Each input of γ approaches $\max_{i\in X} \varphi_k(\boldsymbol{x}_i)$ as $\varepsilon \to 0$
- Strict submodularity is guaranteed when γ is a weighted sum

Learning the Divergence

f is learnable via some metric learning loss, such as **triplet loss**

Triplet Loss

$$\mathcal{L}(f) := \sum_{i=1}^{n} \max\left(D_f\left(X_A^i, X_P^i\right) - D_f\left(X_A^i, X_N^i\right), 0\right)$$

X_A^i: Anchor set,

X_P^i: Positive set,

X_N^i: Negative set

- Makes the anchor-positive discrepancies smaller, and the anchor-negative discrepancies larger
- Learns f unlike ordinary metric learning, which moves intances

Experiments

Sanity Check : MNIST

Learning to make the distances between sets closer, if the sets contain same-labeled instances

Query set	Reference sets	DBD
		0.173
		0.309
		0.772
		0.237
		0.677
		1.304

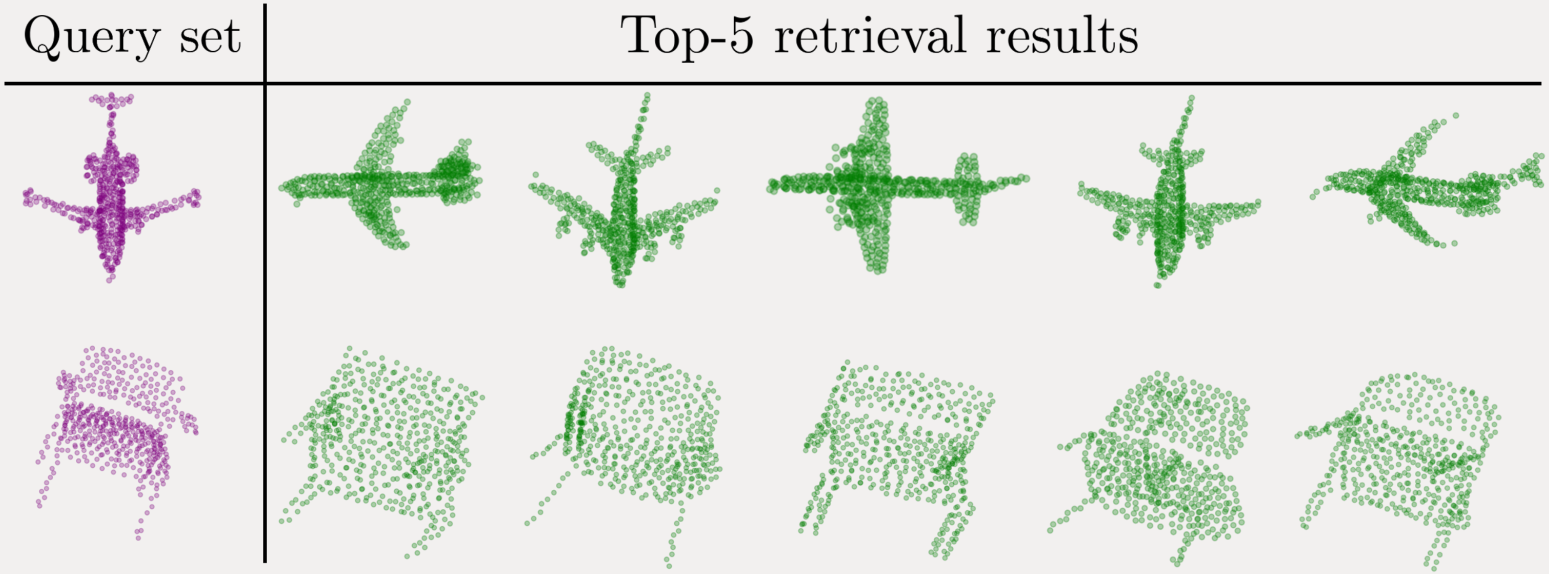
Looks Good 👍

Point Cloud Retrieval (ModelNet40)

Point Cloud : Representative data with set structure

Take top-K closest point clouds to a query cloud

› Assess whether the taken clouds are same-class as the query



Point Cloud Retrieval (ModelNet40)

Near-SoTA performance is achieved with a tiny MLP

- Only uses 64×64 intermediate layers; trainable on CPU!
- Suggests the effectiveness of the proposed framework 👍

Our method w/ 3-types of supergrads

Method	mAP
grow-DBD w/ decomposition	90.13(±0.75)
shrink-DBD w/ decomposition	90.20(±0.77)
bar-DBD w/ decomposition	86.09(±0.85)
grow-DBD w/o decomposition	88.12(±0.80)
shrink-DBD w/o decomposition	88.20(±0.81)
bar-DBD w/o decomposition	83.57(±0.97)
Densepoint (Liu et al., 2019)	89.68(±0.88)
MVTN (Hamdi et al., 2021)	92.9*

Summary

Summary

- **Difference-of-submodular Bregman Divergence** is proposed
  - Extension of submodular Bregman divergence
  - Richer expressiveness is theoretically guaranteed
- Learning the divergence with **permutation-invariant NNs**
  - ε-PointNet: can model strictly submodular functions
- Numerical experiments on MNIST & point cloud data
  - Near-SoTA performance is achieved only with tiny MLPs!