SUBMODULAR HAMMING METRICS

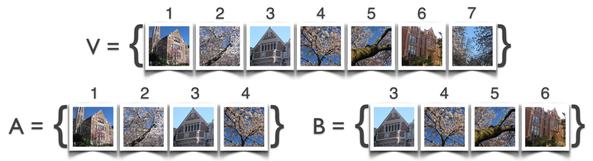
UNIVERSITY of WASHINGTON

MELODI MachinE Learning, Optimization, & Data Interpretation @ UW
http://melodi.ee.washington.edu

JENNIFER GILLENWATER RISHABH IYER BETHANY LUSCH RAHUL KIDAMBI JEFF BILMES

{jengi, rkiyer, herwaldt, rkidambi, bilmes}@uw.edu

DISTANCE METRIC BASICS



Example distance metric: Hamming

$$d_H(A,B) = |A \triangle B| = |(A \setminus B) \cup (B \setminus A)| = |\{1,2\} \cup \{5,6\}| = 4$$

General distance metrics:

Set function $f : 2^V \to \mathbb{R}$ defines distortion $d_f(A,B) = f(A \triangle B)$.

Distortion is a **metric** if it satisfies 4 properties:

non-negativity	$d(A,B) \geq 0$
symmetry	$d(A,B) = d(B,A)$
identity of indiscernables	$d(A,B) = 0 \Leftrightarrow A = B$
triangle inequality	$d(A,B) \leq d(A,C) + d(C,B)$

SUBMODULAR HAMMING (SH) DISTANCES

Set function $f : 2^V \to \mathbb{R}$ is:

1. **positive** if $f(A) > 0$ whenever $A \neq \emptyset$
2. **normalized** if $f(\emptyset) = 0$
3. **monotone** if $f(A) \leq f(B)$ for all $A \subseteq B \subseteq V$
4. **subadditive** if $f(A) + f(B) \geq f(A \cup B)$ for all $A, B \subseteq V$
5. **submodular** if $f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$

Any f with properties 1, 2, 3, and 4 is a **metric**.

Property 5 is needed to make optimization tractable.

For f with all 5 properties, we call d_f an SH distance.

Main focus of our work: exploring SH distance optimization.

Example SH distance: facility location

$$f(A) = \sum_{i \in V} \max_{j \in A} \text{sim}(i,j)$$

$$d_H(A,B) = d_H(A,C) \qquad d_f(A,B) > d_f(A,C)$$

C differs from A only in the "trees", while B differs from A in both "trees" and "buildings".

OPTIMIZATION PROBLEMS

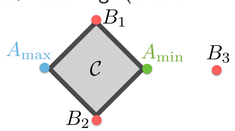
Objective: $F(A) = \sum_{i=1}^{m} f_i(A \triangle B_i)$

Unconstrained: $\mathcal{C} = 2^V$ Cardinality: $|A| \geq k, |A| \leq k$

Other $\mathcal{C}$: knapsack, matroid, cuts, matchings (future work)

SH-min: $\min_{A \in \mathcal{C}} F(A)$

SH-max: $\max_{A \in \mathcal{C}} F(A)$



Centroid-finding problem: find A similar to the B_i.
Diversification problem: find A distinct from the B_i.

EXAMPLE APPLICATIONS

Document clustering: centroid-finding step in k-means

Toy example: cluster has 3 documents—

- Doc 1: Submodularity for Speech Compression
- Doc 2: Summarization via Submodular Mixtures
- Doc 3: Submodularity in Data Selection

After stemming, stop word removal, etc.—

- B_1 = {submodular, speech, compress}
- B_2 = {summarize, submodular, mixture}
- B_3 = {submodular, data, select}

Constrain |A| > 1, find centroid via SH-min— (random second word from B_1, B_2, or B_3)

- with Hamming: centroid = {submodular, speech}
- with $f(A) = \sum_{W \in \mathcal{W}} \sqrt{|A \cap W|}$, given word classes $\mathcal{W}$:

centroid = {submodular, summarize}

Take-away: SH distances allow incorporation of side-information that can yield superior centroids.

Diverse k-best: finding k distinct sets

- **Goal:** Summarize a photo collection (V = set of all photos).
- **Process:** Present summary to user. If user rejects, present new summary that is substantially different. Iterate.
- **Where SH-max comes in:** Find k-th summary via—

$$A_k = \operatorname*{argmax}_{A \subseteq V, |A| = \ell} f(A) + \sum_{i=1}^{k-1} f(A \triangle A_i)$$

- First term: Ensures k-th summary is *internally* diverse.
- Second term: Ensures k-th summary is unlike *previous k-1*.

SH-MIN AND SH-MAX ALGORITHMS

Note: $f(A \triangle B)$ is submodular in $A \triangle B$, but not necessarily in A. Thus, we cannot trivially use standard submodular optimization algorithms for SH-min and SH-max.

Algorithm 1 UNION-SPLIT

Define $f'_i(A) = f_i(A \setminus B_i) + f_i(B_i \setminus A)$
Define $F'(A) = \sum_{i=1}^{m} f'_i(A)$
Output: STANDARD-SUBMOD-OPT (F')

Algorithm 2 BEST-B

$A \leftarrow B_1$
for $i = 2, \ldots, m$ **do**
  if $F(B_i) < F(A)$: $A \leftarrow B_i$
Output: A

Alg 1—key insight: $A \triangle B = (A \setminus B) \cup (B \setminus A)$ and both $f(A \setminus B)$ and $f(B \setminus A)$ are submodular in A itself.

Alg 2—main idea: If $f_i = f_j \; \forall i, j$ (call this the "homogeneous" case), then one of the B_i is a reasonable A.

Algorithm 3 MAJOR-MIN

$A \leftarrow \emptyset$
repeat
  $c \leftarrow F(A)$
  Define $S_j = V \setminus j, T_i = A \triangle B_i$
  Set $\boldsymbol{w}_{\hat{F}}(j) = \sum_{i=1}^{m} \begin{cases} f_i(j \mid S_j) \text{ if } j \in A \triangle B_i \\ f_i(j \mid T_i) \text{ otherwise} \end{cases}$
  $A \leftarrow$ MODULAR-MIN ($\boldsymbol{w}_{\hat{F}}, \mathcal{C}$)
until $F(A) = c$
Output: A

Main idea: Construct $\hat{F}$, a modular upper bound for F at the current solution A, then (exactly) minimize $\hat{F}$ to get a new A. Iterate until convergence.

THEORETICAL RESULTS

n = size of ground set V, m = # of f_i,
κ_f = curvature (larger values imply f is close to modular)

Table 1: Hardness for SH-min and SH-max. UC stands for unconstrained, and Card stands for cardinality-constrained. The entry "open" implies that the problem is potentially poly-time solvable.

	SH-min		SH-max	
	homogeneous	heterogeneous	homogeneous	heterogeneous
UC	Open	4/3	3/4	3/4
Card	$\Omega\left(\frac{\sqrt{n}}{1+(\sqrt{n}-1)(1-\kappa_f)}\right)$	$\Omega\left(\frac{\sqrt{n}}{1+(\sqrt{n}-1)(1-\kappa_f)}\right)$	$1-1/e$	$1-1/e$

Table 2: Approximation guarantees of algorithms for SH-min and SH-max. '-' implies that no guarantee holds for the corresponding pair. BEST-B only works for the homogeneous case, while all other algorithms work in both cases.

	UNION-SPLIT		BEST-B	MAJOR-MIN	RAND-SET
	UC	Card	UC	Card	UC
SH-min	2	-	$2 - 2/m$	$\frac{n}{1+(n-1)(1-\kappa_f)}$	-
SH-max	1/4	$1/2e$	-	-	1/8

SH-MIN EXPERIMENT

Synthetic document clustering: In a setting with 1000 "words" and 100 "documents" assigned to 10 "true" clusters, if each document is a random sampling of 10 words from the 10 word classes $\mathcal{W}$ associated with its cluster, then k-means clustering accuracy with kmeans++ initialization is—

Hamming: 57.0% (±6.8), versus SH: **88.5%** (±8.4)

SH-MAX EXPERIMENT

Diverse k-best image collection summarization: Given 14 image collections, each containing 100 photos, we seek k = 15 summaries of size 10 for each. We then use V-ROUGE to measure summary quality.

Table 4: mV-ROUGE (avg over 14 datasets, ± std dev).

HM	SP	TP
0.38 ± 0.14	0.43 ± 0.20	$\mathbf{0.50 \pm 0.26}$

Table 5: # of wins (out of 14 datasets).

HM	SP	TP
3	1	**10**

HM = Hamming optimized via greedy, SP = SH optimized via greedy, TP = SH optimized via UNION-SPLIT

Summaries of the 6th image collection. HM approach (left) and TP approach (right). Images in the green rectangle tend to be more redundant with images from the previous summaries in the HM case than in the TP case; the HM solution contains many images with a "sky" theme, while TP contains more images with other themes. Quality of the individual summaries also tends to become poorer for the later HM sets; considering the images in the red rectangles overlaid on the montage, the HM sets contain many images of tree branches. By contrast, the TP summary quality remains reasonable even for the last few summaries.