# A note on the triangle inequality for the Jaccard distance

*Sven Kosub*
Department of Computer & Information Science, University of Konstanz
Box 67, D-78457 Konstanz, Germany
Sven.Kosub@uni-konstanz.de

December 9, 2016

**Abstract**

Two simple proofs of the triangle inequality for the Jaccard distance in terms of nonnegative, monotone, submodular functions are given and discussed.

The Jaccard index [8] is a classical similarity measure on sets with a lot of practical applications in information retrieval, data mining, machine learning, and many more (cf., e.g., [7]). Measuring the relative size of the overlap of two finite sets $A$ and $B$, the Jaccard index $J$ and the associated Jaccard distance $J_\delta$ are formally defined as:

$$J(A,B) =_{\text{def}} \frac{|A \cap B|}{|A \cup B|}, \qquad J_\delta(A,B) =_{\text{def}} 1 - J(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|} = \frac{|A \triangle B|}{|A \cup B|}$$

where $J(\emptyset,\emptyset) =_{\text{def}} 1$. The Jaccard distance $J_\delta$ is known to fulfill all properties of a metric, most notably, the triangle inequality—a fact that has been observed many times, e.g., via metric transforms [12, 13, 4], embeddings in vector spaces (e.g., [15, 11, 4]), min-wise independent permutations [1], or sometimes cumbersome arithmetics [10, 3]. A very simple, elementary proof of the triangle inequality was given in [5] using an appropriate partitioning of sets.

Here, we give two more simple, direct proofs of the triangle inequality. One proof comes without any set difference or disjointness of sets. It is based only on the fundamental equation $|A \cup B| + |A \cap B| = |A| + |B|$. As such, the proof is generic and leads to (sub)modular versions of the Jaccard distance (as defined below). The second proof unfolds a subtle difference between the two possible versions. Though the original motivation was to give a proof of the triangle inequality as simple as possible, the link with submodular functions is interesting in itself (as also recently suggested in [6]).

Let $X$ be a finite, non-empty ground set. A set function $f : \mathcal{P}(X) \to \mathbb{R}$ is said to be *submodular* on $X$ if $f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$ for all $A, B \subseteq X$. If all inequalities are equations then $f$ is called *modular* on $X$. It is known that $f$ is submodular on $X$ if and only if the following condition holds (cf., e.g., [14]):

$$f(A \cup \{x\}) - f(A) \geq f(B \cup \{x\}) - f(B) \qquad \text{for all } A \subseteq B \subseteq X,\ x \in \overline{B} \tag{1}$$

A set function $f$ is monotone if $f(A) \leq f(B)$ for all $A \subseteq B \subseteq X$; $f$ is nonnegative if $f(A) \geq 0$ for all $A \subseteq X$. Each nonnegative, monotone, modular function $f$ on $X$ can be

written as $f(A) = \gamma + \sum_{i \in A} c_i$ where $\gamma, c_i \geq 0$ for all $i \in X$ (cf., e.g., [14]). Examples are set cardinality or degree sum in graphs. Standard examples of nonnegative, monotone, submodular set functions are matroid rank, network flow to a sink, entropy of sets of random variables, and neighborhood size in bipartite graphs.

Let $f$ be a nonnegative, monotone, submodular set function on $X$. For sets $A, B \subseteq X$, we define two candidates for *submodular Jaccard distances*, $J_{\delta,f}$ and $J^{\Delta}_{\delta,f}$, as follows:

$$J_{\delta,f}(A,B) =_{\text{def}} 1 - \frac{f(A \cap B)}{f(A \cup B)}, \qquad J^{\Delta}_{\delta,f} =_{\text{def}} \frac{f(A \triangle B) - f(\emptyset)}{f(A \cup B)},$$

where $J_{\delta,f}(A,B) = J^{\Delta}_{\delta,f}(A,B) =_{\text{def}} 0$ if $f(A \cup B) = 0$. It is clear that $0 \leq J_{\delta,f}(A,B) \leq J^{\Delta}_{\delta,f}(A,B)$. If $f$ is modular then $J_{\delta,f} = J^{\Delta}_{\delta,f}$. In particular, for $f(A) = |A|$ (i.e., the cardinality of the set $A \subseteq X$), we obtain the standard Jaccard distance $J_\delta = J_{\delta,f} = J^{\Delta}_{\delta,f}$.

First, we give a simple proof of the triangle inequality for $J_{\delta,f}$. Interestingly, this is only possible for modular set functions (see the third remark after Theorem 3).

**Lemma 1.** *Let $f$ be a nonnegative, monotone, submodular set function on $X$. Then, for all sets $A, B, C \subseteq X$, it holds that*

$$f(A \cap C) \cdot f(B \cup C) + f(A \cup C) \cdot f(B \cap C) \;\leq\; f(C) \cdot \big(f(A) + f(B)\big).$$

*Proof.* We easily obtain

$$\begin{aligned} & f(A \cap C) \cdot f(B \cup C) \\ & \quad \leq \; f(A \cap C) \cdot \big(f(B) + f(C) - f(B \cap C)\big) && \text{(submodularity of } f\text{)} \\ & \quad \leq \; f(C) \cdot \big(f(B) - f(B \cap C) + f(A \cap C)\big) && \text{(monotonicity of } f\text{)} \end{aligned}$$

and, by swapping $A$ and $B$, $f(A \cup C) \cdot f(B \cap C) \leq f(C) \cdot \big(f(A) - f(A \cap C) + f(B \cap C)\big)$. Overall,

$$\begin{aligned} & f(A \cap C) \cdot f(B \cup C) + f(A \cup C) \cdot f(B \cap C) \\ & \quad \leq \; f(C) \cdot \big(f(B) - f(B \cap C) + f(A \cap C) + f(A) - f(A \cap C) + f(B \cap C)\big) \\ & \quad = \; f(C) \cdot \big(f(B) + f(A)\big) \end{aligned}$$

This shows the lemma. □

**Corollary 2.** *Let $f$ be a nonnegative, monotone, submodular set function on $X$. Then, for all sets $S, T \subseteq X$, it holds that*

$$f(S \cap T) \cdot f(S \cup T) \leq f(S) \cdot f(T).$$

*Proof.* Apply Lemma 1 to sets $A =_{\text{def}} S$, $B =_{\text{def}} S$ and $C =_{\text{def}} T$. □

**Theorem 3.** *Let $f$ be a nonnegative, monotone, modular set function on $X$. Then, for all sets $A, B, C \subseteq X$, it holds that*

$$J_{\delta,f}(A,B) \leq J_{\delta,f}(A,C) + J_{\delta,f}(C,B).$$

*Proof.* Say that a set $A$ is a null set iff $f(A) = 0$. Observe that if at least one of the sets is a null set then the inequality is satisfied. So, it is enough to show the equivalent inequality

$$\frac{f(A \cap C)}{f(A \cup C)} + \frac{f(B \cap C)}{f(B \cup C)} \;\leq\; 1 + \frac{f(A \cap B)}{f(A \cup B)} = \frac{f(A) + f(B)}{f(A \cup B)} \tag{2}$$

for arbitrary non-null sets $A, B, C \subseteq I$. This is seen as follows:

$$\begin{aligned}
&\frac{f(A \cap C)}{f(A \cup C)} + \frac{f(B \cap C)}{f(B \cup C)} \\
&\quad= \frac{f(A \cap C) \cdot f(B \cup C) + f(A \cup C) \cdot f(B \cap C)}{f(A \cup C) \cdot f(B \cup C)} \\
&\quad\leq \frac{f(C) \cdot \big(f(A) + f(B)\big)}{f(A \cup C) \cdot f(B \cup C)} && \text{(by Lemma 1)} \\
&\quad\leq \frac{f(C) \cdot \big(f(A) + f(B)\big)}{f\big((A \cup C) \cap (B \cup C)\big) \cdot f(A \cup B \cup C)} && \text{(by Corollary 2)} \\
&\quad\leq \frac{f(C)}{f\big((A \cap B) \cup C\big)} \cdot \frac{f(A) + f(B)}{f(A \cup B)} && \text{(monotonicity of } f\text{)} \\
&\quad\leq \frac{f(A) + f(B)}{f(A \cup B)} && \text{(monotonicity of } f\text{)}
\end{aligned}$$

This proves the theorem. □

*Remarks*: We comment on the proof of the triangle inequality for $J_{\delta,f}$:

1. It follows from Theorem 3 that the triangle inequality is valid for the standard Jaccard distance $J_\delta$, the generalized Jaccard distance given for vectors $x, y \in \mathbb{R}^n$ by

$$1 - \frac{\sum_{i=1}^n \min\{x_i, y_i\}}{\sum_{i=1}^n \max\{x_i, y_i\}}$$

(with the subcase that $x_i = \mu_A(z)$ and $y_i = \mu_B(z)$ denote multiplicities of (occurrences of) $z$ in multisets $A$ and $B$; cf. [9]), and the Steinhaus distance [12, 4] (i.e., any set measures, including probability measures). We mention that all these results can equally easily be proven by the arguments in [5]; however, for modular functions satisfying $f(\emptyset) > 0$, these arguments fail.

2. Theorem 3 is true for nonnegative, monotone, modular functions defined over distributive lattices; Lemma 1 and Corollary 2 also hold for nonnegative, monotone, submodular functions defined over distributive lattices. Notice that $J^{\Delta}_{\delta,f}$ is not defined over all distributive lattices (see also the third remark after Theorem 4).

3. In general, Theorem 3 is not true for nonnegative, monotone, submodular functions: Any set function $f$ such that $f(A) = f(B) = f(A \cup B) > f(A \cap B) \geq 0$ for nonempty, incomparable sets $A, B$ refutes $J_{\delta,f}(A, B) \leq J_{\delta,f}(A, A \cup B) + J_{\delta,f}(A \cup B, B)$. Concrete examples include linear cost functions with budget restrictions, i.e., $f(A) = \min\{B, \sum_{i \in A} c_i\}$, or the neighborhood size in a bipartite graph $G = (U \uplus V, E)$, i.e., $f(A) = |\Gamma(A)|$ where $A \subseteq U$ and $\Gamma(A) = \bigcup_{u \in A} \{v \in V | \{u, v\} \in E\}$.

Next we give a simple proof of the triangle inequality for $J^{\Delta}_{\delta,f}$.

**Theorem 4.** *Let $f$ be a nonnegative, monotone, submodular set function on $X$. Then, for all sets $A, B, C \subseteq X$, it holds that*

$$J^{\Delta}_{\delta,f}(A,B) \leq J^{\Delta}_{\delta,f}(A,C) + J^{\Delta}_{\delta,f}(C,B).$$

*Proof.* We split the set $C$ into two disjoint sets $C_0 \subseteq A \cup B$ and $C_1 \subseteq \overline{A \cup B}$, both possibly empty, such that $C = C_0 \cup C_1$. We obtain

$$\begin{aligned}
&\frac{f(A\triangle C) - f(\emptyset)}{f(A \cup C)} + \frac{f(B\triangle C) - f(\emptyset)}{f(B \cup C)} \\
&\quad\geq \frac{f(A\triangle C) + f(B\triangle C) - 2f(\emptyset)}{f(A \cup B \cup C_1)} && \text{(monotonicity of } f\text{)} \\
&\quad\geq \frac{f(A\triangle C \cup B\triangle C) - f(\emptyset)}{f(A \cup B \cup C_1)} && \text{(submodularity, monotonicity of } f\text{)} \\
&\quad\geq \frac{f(A\triangle B \cup C_1) - f(\emptyset)}{f(A \cup B \cup C_1)} && \text{(monotonicity of } f\text{)} \\
&\quad\geq \frac{f(A\triangle B)}{f(A \cup B)} - \frac{f(\emptyset)}{f(A \cup B \cup C_1)} && \text{(submodularity of } f\text{, Cond. (1))} \\
&\quad\geq \frac{f(A\triangle B)}{f(A \cup B)} - \frac{f(\emptyset)}{f(A \cup B)} && \text{(monotonicity of } f\text{)}
\end{aligned}$$

This shows the theorem. □

*Remarks*: We comment on the proof of the triangle inequality for $J^{\Delta}_{\delta,f}$:

1. It follows once more from Theorem 4 that the standard Jaccard distance, the generalized Jaccard distance, and the Steinhaus distance satisfy the triangle inequality. Moreover, $J^{\Delta}_{\delta,f}$ is also a (pseudo)metric for, e.g., linear cost functions with budget restrictions and the neighborhood size in bipartite graphs.

2. Theorem 4 suggests that $J^{\Delta}_{\delta,f}$ is the right definition of a submodular Jaccard distance. As a consequence, one might say that the submodular Jaccard (similarity) index should be defined as the inverse submodular Jaccard distance, i.e.,

$$J^{\Delta}_{f}(A,B) =_{\text{def}} 1 - J^{\Delta}_{\delta,f} = 1 - \frac{f(A\triangle B) - f(\emptyset)}{f(A \cup B)}$$

   Again, if $f(A) = |A|$ then we obtain the standard Jaccard index $J = J^{\Delta}_{f} = 1 - J_{\delta,f}$.

3. Though $J^{\Delta}_{\delta,f}$ might generally not be defined over a given distributive lattice, it can be seen that for each nonnegative, monotone, submodular function $f : \mathcal{F} \to \mathbb{R}$ defined on a family $\mathcal{F} \subseteq \mathcal{P}(X)$ closed under union and intersection, there is a (not necessarily unique) nonnegative, monotone, submodular extension $\overline{f} : \mathcal{P}(X) \to \mathbb{R}$ on $X$ such that $\overline{f}(A) = f(A)$ for all $A \in \mathcal{F}$ (e.g., [16]), so that $J^{\Delta}_{\delta,\overline{f}}$ can be used instead.

**Acknowledgments:** I am grateful to Ulrik Brandes (Konstanz) and Julian Müller (Konstanz) for helpful discussions.

# References


[1] M. S. Charikar. Similarity Estimation Techniques from Rounding Algorithms. In: *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC'2002)*, pp. 380–388. ACM Press, New York, NY, 2002.

[2] M. M. Deza, E. Deza. *Encyclopedia of Distances.* Springer, Berlin, 2009.

[3] O. Fujita. Metrics based on average distance between sets. *Japan Journal of Industrial and Applied Mathematics*, 30(1):1–19, 2013.

[4] A. Gardner, J. Kanno, C. A. Duncan, R. Selmic. Measuring Distance Between Unordered Sets of Different Sizes. In: *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2014)*, pp. 137–143. IEEE, New Jersey, NJ, 2014.

[5] G. Gilbert. Distance between sets. Letters to *Nature*, 239(5368):174, 1972.

[6] J. Gillenwater, R. Iyer, B. Lusch, R. Kidambi, J. A. Bilmes. Submodular Hamming Metrics. In: *Advances in Neural Information Processing Systems 28*, 3141-3149. NIPS Proceedings, December 2015.

[7] J. C. Gower. Similarity, Dissimilarity and Distance, Measures of. In: S. Kotz, C. B. Read, N. Balakrishnan, B. Vidakovic (eds.), *Encyclopedia of Statistical Sciences*, vol. 12., pp. 7730–7738. 2nd edition, John Wiley, New York, NY, 2008.

[8] P. Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37(142):547–579, 1901.

[9] W. A. Kosters, J. F. J. Laros. Metrics for Mining Multisets. In: M. Bramer, F. Coenen, M. Petridis (eds.), *Research and Development in Intelligent Systems XXIV, Proceedings of the Twenty-seventh SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence (AI'2007)*, pp. 293–303. Springer, Berlin, 2007.

[10] M. Levandowsky, D. Winter. Distance between sets. Letters to *Nature*, 234(5323):34–35, 1971.

[11] A. H. Lipkus. A proof of the triangle inequality for the Tanimoto distance. *Journal of Mathematical Chemistry*, 26:263–265, 1999.

[12] E. Marczewski, H. Steinhaus. On a certain distance of sets and the corresponding distance of functions. *Colloquium Mathematicum*, 6:319–327, 1958.

[13] D. A. Simovici, C. Djeraba. *Mathematical Tools for Data Mining.* Springer, London, 2008.

[14] A. Schrijver. *Combinatorial Optimization*, vol. B. Springer, Berlin, 2003.

[15] T. T. Tanimoto. An elementary mathematical theory of classification and prediction. IBM Report, November 1958.

[16] D. M. Topkis. Minimizing a submodular function on a lattice. *Operations Research*, 26(2):305–321, 1978.