The Power of Bounds: Answering Approximate Earth Mover's Distance with Parametric Bounds (Extended abstract)

Tsz Nam Chan*†, Man Lung Yiu†, Leong Hou U‡
*Department of Computer Science, The University of Hong Kong
tnchan@cs.hku.hk
†Department of Computing, Hong Kong Polytechnic University
{cstnchan,csmlyiu}@comp.polyu.edu.hk
‡State Key Laboratory of Internet of Things for Smart City
‡Department of Computer and Information Science, University of Macau
ryanlhu@umac.mo

***Abstract*—The *Earth Mover's Distance* (EMD) is a robust similarity measure between two histograms (e.g., probability distributions). It has been extensively used in a wide range of applications, e.g., multimedia, data mining, computer vision, etc. As EMD is a computationally intensive operation, many efficient lower and upper bound functions of EMD have been developed. However, they provide no guarantee on the error. In this work, we study how to compute approximate EMD value with bounded error, using these bound functions. First, we propose an approximation framework that leverages on lower and upper bound functions to compute approximate EMD with error guarantee. Then, we present three solutions to solve our problem. Experimental results on real data demonstrate the efficiency of our proposed solutions.**

I. Introduction

The *Earth Mover's Distance* (EMD) is a robust similarity measure between two d-dimensional histograms $\mathbf{q}$ and $\mathbf{p}$, which has been extensively used in many applications, including multimedia databases [7], computer vision [5] etc. Formally, we define $emd_c(\mathbf{q}, \mathbf{p})$ as the following linear programming problem, given the $d \times d$ cost matrix c.

$$\begin{aligned} emd_c(\mathbf{q}, \mathbf{p}) = \quad & \underset{f}{\text{minimize}} \sum_{i=1}^{d}\sum_{j=1}^{d} c_{i,j} f_{i,j} \\ \text{such that} \quad & \forall i,j \in [1..d] : f_{i,j} \geq 0 \\ & \forall i \in [1..d] : \sum_{j=1}^{d} f_{i,j} = q_i \\ & \forall j \in [1..d] : \sum_{i=1}^{d} f_{i,j} = p_j \end{aligned}$$

However, EMD is the computationally expensive operation, which takes $O(d^3 \log d)$ time to obtain the exact value, even with the state-of-the-art solution [4]. On the other hand, many applications [7], e.g., image retrieval, require EMD computations on a massive amount of objects. Therefore, it motivates us for developing the rapid solutions to EMD operation.

In our work [2], our objective is to develop efficient algorithms for obtaining the approximate EMD value R with bounded error (cf. Problem 1).

Problem 1 (Error-Bounded EMD)**.** *Given an error threshold ϵ, this problem returns a value R such that $\mathbb{E}_{\mathbf{q},\mathbf{p}}(R) \leq \epsilon$, where the relative error of R is defined as:*

$$\mathbb{E}_{\mathbf{q},\mathbf{p}}(R) = \frac{|R - emd_c(\mathbf{q}, \mathbf{p})|}{emd_c(\mathbf{q}, \mathbf{p})} \tag{1}$$

Even though many studies have developed different lower and upper bound functions (cf. Table I) for EMD, simply using the bound value as R cannot fulfill the bounded error guarantee. As a remark, both LB_{Red,d_r}, $LB_{skew,\lambda}$ and $UB_{skew,\lambda}$ are the parametric bound functions. Each of these bound functions accepts the additional parameter (e.g., d_r and λ) to control its running time and tightness.

TABLE I: Summary of lower and upper bound functions

Name	Type	Time Complexity	Reference	Parametric
LB_{IM}	lower	$O(d^2)$	[1]	no
LB_{Proj}	lower	$O(d)$	[6]	no
LB_{Red,d_r}	lower	$O(d^2 + {d_r}^3 \log d_r)$	[8]	yes
UB_G	upper	$O(d^2)$	[7]	no
UB_H	upper	$O(d)$	[3]	no
$LB_{skew,\lambda}$	lower	$O((d-\lambda)d + \lambda^3 \log \lambda)$	[2]	yes
$LB_{skew,\lambda}$	upper	$O((d-\lambda)d + \lambda^3 \log \lambda)$	[2]	yes

II. Approximation Framework

In this work, we develop the approximation framework (cf. Figure 1), which is composed of two components, controller and validator.

- The *controller* selects a lower bound function and an upper bound function. Then it computes a lower bound ℓ, an upper bound u, and an approximate result R which is the value between ℓ and u.

This work was supported by grant GRF152201/14E from the Hong Kong RGC. Leong Hou U was funded by the science and technology development fund, Macau SAR (SKL-IOTSC-2018-2020) and the University of Macau (MYRG2019-00119-FST).

- The *validator* receives information (e.g., ℓ, u, R) from the controller, and then checks whether the relative error definitely satisfies $\mathbb{E}_{\mathbf{q},\mathbf{p}}(R) \leq \epsilon$.

If the validator returns true, then the controller reports R to the user. Otherwise, the controller needs to obtain tighter bounds for ℓ and u, and repeats the above procedure.

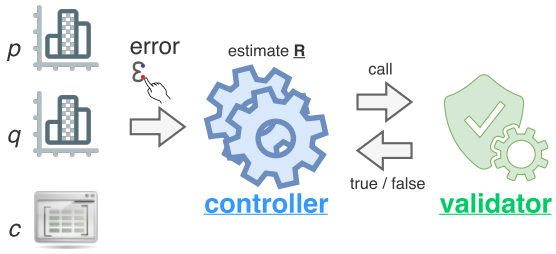

Fig. 1: Framework

A. Validator

In order to secure the correctness of our framework, we specify the following requirements for the validator:

- If it returns true, then it guarantees that the approximate result R must satisfy $\mathbb{E}_{\mathbf{q},\mathbf{p}}(R) \leq \epsilon$.
- Otherwise, it does not provide any guarantee for R.

Theorem 1 illustrates how the lower and upper bound values, ℓ and u respectively, can be used to construct R such that it satisfies $\mathbb{E}_{\mathbf{q},\mathbf{p}}(R) \leq \epsilon$.

Theorem 1. *Given the lower and upper bound values, ℓ and u respectively, if $\frac{u-\ell}{u+\ell} \leq \epsilon$, then $\mathbb{E}_{\mathbf{q},\mathbf{p}}(R) \leq \epsilon$, where $R = \frac{2\ell u}{\ell+u}$.*

B. Adaptive Controller (ADA)

In our work [2], we have developed parametric dual bound functions, $LB_{skew,\lambda}$ and $UB_{skew,\lambda}$ (cf. Table I). ADA gradually applies tighter bounds, by setting larger λ, until passing the validation test (cf. Theorem 1). Theoretically, we show that ADA can be worse than the ADA-Opt (which knows the optimal λ value in advance for each $(\mathbf{q}, \mathbf{p})$ pair) by only a constant factor 5.18 if we select the suitable sequence of λ [2].

C. Lightweight Adaptive Controller (ADA-L)

ADA may examine several λ and compute exact EMD operations multiple times (in the adaptive phase). Thus, ADA can be expensive when ϵ is small. To avoid such overhead, we propose a lightweight version of the adaptive method, called ADA-L, such that it computes $emd_c(\mathbf{q}', \mathbf{p}')$ exactly once (cf. Figure 2). Even though ADA-L does not have theoretical performance guarantee as ADA [2], the practical efficiency performance is better than ADA.

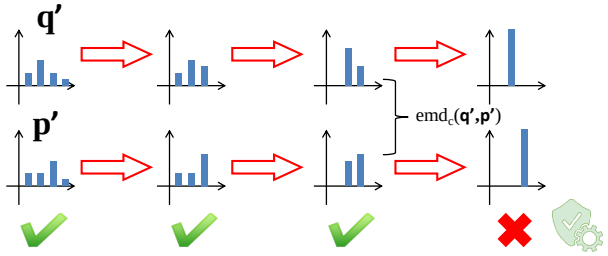

Fig. 2: Lightweight adaptive approach

D. Training-based Controller (ADA-H)

Some applications, e.g., image retrieval and image classification, might have huge historical workload data Γ. Such rich information can help to pick the bounds such that the framework can find a good approximate result R at lower cost, compared to ADA-L and ADA. In this controller, we propose the greedy stretagy [2], based on the statistics of historical workload data, to pick the sequence of bounds in the offline stage, as shown in Figure 3. Then, we adopt this sequence of bounds in the online stage.

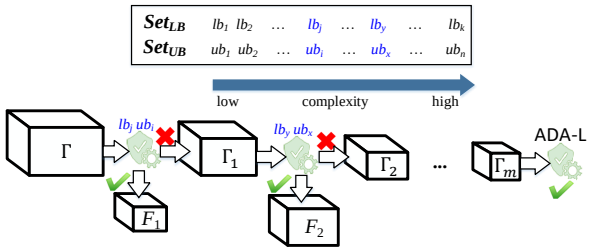

Fig. 3: Picking a sequence of bounds in the offline stage

III. Experiment

In the following, we test the efficiency performance with four methods, EXACT, ADA-L, ADA-H and Oracle, where EXACT is the exact EMD method and Oracle is the omniscient method, which pre-knows the fastest pair of (ℓ, u) which fulfills the validation condition $\frac{u-\ell}{u+\ell} \leq \epsilon$. As such, it acts as the most efficient solution for all control methods in our approximation framework. We report the result in Caltech dataset (30,609 images) with RGB and Lab histogram extraction methods. Observe from Figure 4, our best method ADA-H outperforms EXACT by at least one-order-of-magnitude.

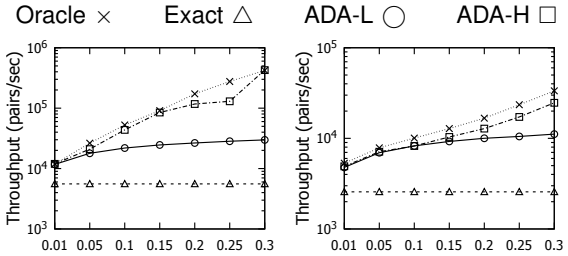

Fig. 4: Effect of the error threshold ϵ on Caltech dataset

References

[1] I. Assent, A. Wenning, and T. Seidl. Approximation techniques for indexing the earth mover's distance in multimedia databases. In *ICDE*, page 11, 2006.

[2] T. N. Chan, M. L. Yiu, and L. H. U. The power of bounds: Answering approximate earth mover's distance with parametric bounds. *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[3] M. Jang, S. Kim, C. Faloutsos, and S. Park. A linear-time approximation of the earth mover's distance. In *CIKM*, pages 505–514, 2011.

[4] J. B. Orlin. A faster strongly polynominal minimum cost flow algorithm. In *STOC*, pages 377–387, 1988.

[5] O. Pele and M. Werman. Fast and robust earth mover's distances. In *ICCV*, pages 460–467, 2009.

[6] B. E. Ruttenberg and A. K. Singh. Indexing the earth mover's distance using normal distributions. *PVLDB*, 5(3):205–216, 2011.

[7] Y. Tang, L. H. U, Y. Cai, N. Mamoulis, and R. Cheng. Earth mover's distance based similarity search at scale. *PVLDB*, 7(4):313–324, 2013.

[8] M. Wichterich, I. Assent, P. Kranen, and T. Seidl. Efficient emd-based similarity search in multimedia databases via flexible dimensionality reduction. In *SIGMOD*, pages 199–212, 2008.