Conditional Hardness of Earth Mover Distance

Dhruv Rohatgi
MIT, Cambridge, Massachusetts, USA
drohatgi@mit.edu

Abstract

The Earth Mover Distance (EMD) between two sets of points $A, B \subseteq \mathbb{R}^d$ with $|A| = |B|$ is the minimum total Euclidean distance of any perfect matching between A and B. One of its generalizations is asymmetric EMD, which is the minimum total Euclidean distance of any matching of size $|A|$ between sets of points $A, B \subseteq \mathbb{R}^d$ with $|A| \le |B|$. The problems of computing EMD and asymmetric EMD are well-studied and have many applications in computer science, some of which also ask for the EMD-optimal matching itself. Unfortunately, all known algorithms require at least quadratic time to compute EMD exactly. Approximation algorithms with nearly linear time complexity in n are known (even for finding approximately optimal matchings), but suffer from exponential dependence on the dimension.

In this paper we show that significant improvements in exact and approximate algorithms for EMD would contradict conjectures in fine-grained complexity. In particular, we prove the following results:

- Under the Orthogonal Vectors Conjecture, there is some $c > 0$ such that EMD in $\Omega(c^{\log^* n})$ dimensions cannot be computed in truly subquadratic time.
- Under the Hitting Set Conjecture, for every $\delta > 0$, no truly subquadratic time algorithm can find a $(1 + 1/n^\delta)$-approximate EMD matching in $\omega(\log n)$ dimensions.
- Under the Hitting Set Conjecture, for every $\eta = 1/\omega(\log n)$, no truly subquadratic time algorithm can find a $(1 + \eta)$-approximate asymmetric EMD matching in $\omega(\log n)$ dimensions.

1 Introduction

In the *Earth Mover Distance (EMD) problem*, we are given two sets A and B each with n vectors in $\mathbb{R}^d$, and want to find the minimum cost of any perfect matching between A and B, where an edge between $a \in A$ and $b \in B$ has cost $\|a - b\|_2$.

In a harder variant of the problem ("EMD matching"), we want to actually *find* a perfect matching with the optimal cost. This is a special case of the *geometric transportation problem*, in which each vector of A has a positive supply and each vector of B has a positive demand, and the goal is to find an optimal "transportation map", i.e., match each unit of supply with a unit of demand while minimizing the total distance, summed over all units of supply.

A more general variant of the EMD problem (with an analogous extension to arbitrary supplies/demands) allows for the possibility that $|A| < |B|$, and requires the map from A to B to be an injection. We refer to this variant as the *asymmetric* EMD problem.

Earth Mover Distance is a discrete analogue of the Monge-Kantorovich metric for probability measures, which has connections to various areas of mathematics [26]. Furthermore, computing distance between probability measures is an important problem in machine learning [23, 20, 7, 13] and computer vision [22, 10, 25], to which Earth Mover Distance is often applied. To provide a few specific examples, computing geometric transportation cost has applications in image retrieval [22], where asymmetric EMD allows the distance to deal with occlusions and clutter. In computer graphics, computing the actual transportation map is useful for interpolation between distributions, though the metric may be non-Euclidean [10].

For the exact geometric transportation problem, the best known algorithm simply formulates the problem in terms of minimum cost flow, yielding a runtime of $O(n^{2.5} \cdot \text{polylog}(U))$ where U is the total supply (assuming that d is subpolynomial in n) [18, 19]. Even for EMD, the best known algorithm follows directly from the general graph algorithms for maximum matching in $O(m\sqrt{n})$ time [14].

The situation is better for approximation algorithms. There has been considerable work on both estimating the transportation cost [15, 6] and computing the actual map [24, 3, 5] in time nearly linear in n but exponential in dimension d. Most recently, it was shown [17] that there is an $O(n\epsilon^{-O(d)} \log(U)^{O(d)} \log^2 n)$ time algorithm which outputs a transportation map with cost at most $(1 + \epsilon)$ times the optimum. This algorithm is very efficient when the dimension d is constant or nearly constant, and when ϵ is not too small – say, constant or $O(1/\text{polylog}(n))$. However, when $d = \omega(\log n)$, the algorithm is not guaranteed to find even a constant-factor approximation in quadratic time.

Despite considerable progress on improving the *algorithms* for geometric matching problems over the last two decades, little is known about *lower bounds* on their computational complexity. In particular, we do not have any evidence that a running time of the form $O(n \cdot \text{poly}(d, \log n, 1/\epsilon))$ is not achievable. This is the question we address in this paper.

1.1 Our Results

In this paper we provide evidence that geometric transportation problems in high-dimensional spaces cannot be solved in (truly) subquadratic time. This applies to both exact and approximate variants of the problem, and even in the special case of unit supplies. In particular we show a conditional quadratic hardness for the exact EMD problem, as well as the approximate variant of EMD when the (approximately) optimal matching must be reported.

Our hardness results are based on two well-studied conjectures in fine-grained complexity: Orthogonal Vectors Conjecture and Hitting Set Conjecture (see [29] for a comprehensive survey).

1.1.1 Exact EMD and Orthogonal Vectors Conjecture

The *Orthogonal Vectors (OV) problem* takes as input two sets $A, B \subseteq \{0, 1\}^{d(n)}$ where $|A| = |B| = n$ and asks whether there are some vectors $a \in A$ and $b \in B$ such that $a \cdot b = 0$. The popular *Orthogonal Vectors Conjecture* hypothesizes that in sufficiently large dimensions, the obvious quadratic time algorithm for OV is nearly optimal:

▶ **Orthogonal Vectors Conjecture.** *Let $d(n) = \omega(\log n)$. For every constant $\epsilon > 0$, no randomized algorithm can solve $d(n)$-dimensional OV in $O(n^{2-\epsilon})$ time.*

A plethora of problems have been shown to have nontrivial lower bounds under the Orthogonal Vectors Conjecture; often these lower bounds are essentially tight (e.g. [1, 2, 8, 11, 28]; see [29] for a comprehensive survey). It is known that if the conjecture fails, then the Strong Exponential Time Hypothesis (SETH) fails as well [27], providing evidence for hardness of OV, and by extension of these problems to which OV can be reduced.

Our first result shows that EMD in "nearly constant" dimension is hard to compute exactly in truly subquadratic time, under the Orthogonal Vectors Conjecture:

▶ **Theorem 1.** *There is a constant $c > 0$ under which the following holds. If there exists $\epsilon > 0$ and $d(n) = \Omega(c^{\log^* n})$ such that EMD on $O(\log n)$-bit vectors in $d(n)$ dimensions can be computed in $O(n^{2-\epsilon})$ time, then the Orthogonal Vectors Conjecture is false.*

Using techniques similar to those for the above theorem, we also address a question raised in [9] about the complexity of the maximum/minimum weighted assignment problem when the weight matrix has low rank. The minimum weighted assignment problem is defined as follows: given an $n \times n$ weight matrix which determines a complete bipartite graph, find the cost of the minimum weight perfect matching. Motivated by the observation that the problem can be solved in $O(n \log n)$ time if the weight matrix is rank-1, it is asked whether there is an $O(nr^2 \log n)$ time algorithm for rank-r matrices [9]. We can answer this question in the negative, under the Orthogonal Vectors Conjecture. In fact, we can show something stronger (see Appendix A for the proof):

▶ **Theorem 2.** *There is a constant $c > 0$ under which the following holds. If there exists $\epsilon > 0$ and $r(n) = \Omega(c^{\log^* n})$ such that the minimum assignment problem with rank-r weight matrices can be solved in $O(n^{2-\epsilon})$ time, then the Orthogonal Vectors Conjecture is false.*

1.1.2 Approximate EMD and the Hitting Set Conjecture

The second conjecture on which we base some of our results is hardness of the *Hitting Set (HS) problem*. This problem, similar to OV, takes two sets of vectors $A, B \subseteq \{0,1\}^d$ as input, and asks whether there exists some $a \in A$ such that $a \cdot b \neq 0$ for every $b \in B$.

▶ **Hitting Set Conjecture.** *Let $d(n) = \omega(\log n)$. For every constant $\epsilon > 0$, no randomized algorithm can solve $d(n)$-dimensional HS in $O(n^{2-\epsilon})$ time.*

It is known that HS reduces to OV, but the reverse reduction is unknown, so the Hitting Set Conjecture is "stronger" than the Orthogonal Vectors Conjecture [2]. The Hitting Set Conjecture has been used to prove conditional hardness of the Radius problem in sparse graphs [2]. The utility of the Hitting Set problem in conditional hardness results comes from the difference between its "∃∀" logical structure and the "∃∃" logical structure of the Orthogonal Vectors problem, which makes it more natural for some types of problems.

Under the Hitting Set Conjecture, we prove hardness of approximation for the EMD matching problem (in which we want to find the optimal or nearly-optimal matching). Simultaneously we obtain stronger hardness of approximation for asymmetric EMD matching.

▶ **Theorem 3.** *For any $\delta > 0$ and $d(n) = \omega(\log n)$, if $(1 + 1/n^\delta)$-approximate EMD matching can be solved in $d(n)$ dimensions in truly subquadratic time, then the Hitting Set conjecture is false.*

▶ **Theorem 4.** *For any $d(n) = \omega(\log n)$ and $\eta = 1/\omega(\log n)$, if $(1 + \eta)$-approximate asymmetric EMD matching can be solved in $d(n)$ dimensions in truly subquadratic time, then the Hitting Set Conjecture is false.*

Finally, motivated by the question of how hard Hitting Set really is, compared to Orthogonal Vectors, we generalize the result that Hitting Set reduces to Orthogonal Vectors by finding a set of approximation problems that lie between Orthogonal Vectors and Hitting Set in difficulty. For a positive integer function $k(n) \leq n/2$, we define the $(k, 2k)$-*Find-OV*
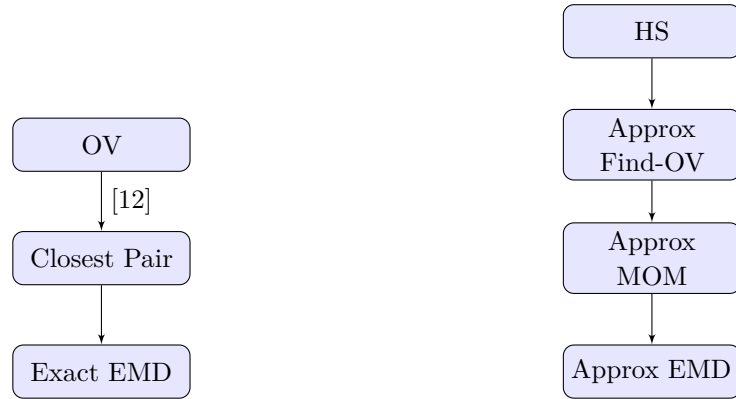
(a) Structure of Theorem 1.

(b) Structure of Theorem 3.

■ **Figure 1** Summary of reductions.

problem: given two sets $A, B \subseteq \{0, 1\}^{d(n)}$ with $|A| = |B| = n$ and the guarantee that there exist at least $2k$ orthogonal pairs between A and B, find k pairs $\{(a_i, b_i)\}_{i=1}^{k}$ such that $a_i \cdot b_i = 0$ for every i.

We prove the following theorem in Appendix C.

▶ **Theorem 5.** *Let $k(n) \leq n/2$. If $(k, 2k)$-Find-OV can be solved in truly subquadratic time, then the Hitting Set conjecture is false.*

See Figure 1 for an overview of the structure of our main results (Theorems 1 and 3 respectively; the proof of Theorem 4 has the same structure as the latter). We provide the remaining definitions of the relevant problems in the next section.

2   Preliminaries

Before diving into the reductions, we formally define the remainder of the problems which we're studying. Each problem we study takes sets of vectors as input, so one parameter of a problem is the dimension d, which is a function of the input size n. That is, every function $d : \mathbb{N} \to \mathbb{N}$ defines a $d(n)$-dimensional EMD problem, and a $d(n)$-dimensional OV problem, and so forth. We gloss over this choice of d in the subsequent definitions.

2.1   Earth Mover Distance

The *Earth Mover Distance (EMD)* problem is defined as follows: given two sets $A, B \subseteq \mathbb{R}^{d(n)}$ with $|A| = |B|$, find

$$\min_{\pi : A \to B} \sum_{a \in A} \|a - \pi(a)\|_2$$

where π is a bijection. We'll restrict our attention to the special cases where $A, B \subseteq \mathbb{Z}^{d(n)}$ with polynomially bounded entries (for hardness of exact EMD) and $A, B \subseteq \{0, 1\}^{d(n)}$ (for hardness of approximate EMD).

We can define the *asymmetric EMD* problem as above, except we relax the constraint $|A| = |B| = n$ to $|A| \leq |B| = n$, and require π to be a injection rather than a bijection.

The *EMD matching* problem is the variant of the EMD problem in which the desired output is the optimal matching π. Similarly we can define the *asymmetric EMD matching* problem. An algorithm "solves" EMD matching (or its asymmetric variant) up to a certain additive or multiplicative factor if the cost of the bijection it outputs differs from the optimal cost by at most that additive or multiplicative factor.

2.2 Variants of Orthogonal Vectors

The reduction from Hitting Set to approximate EMD matching will go through the variants of OV defined next.

The *Maximum Orthogonal Matching (MOM)* problem is defined as follows: given two sets $A, B \subseteq \{0, 1\}^{d(n)}$, with $|A| \leq |B| = n$, find an injection $\pi : A \to B$ which maximizes

$$|\{a \in A \mid a \cdot \pi(a) = 0\}|.$$

And the *Find-OV* problem is defined as follows: given two sets $A, B \subseteq \{0, 1\}^{d(n)}$ with $|A| = |B| = n$, find the set $S \subseteq A$ of vectors $a \in A$ such that there exists some $b \in B$ with $a \cdot b = 0$. An algorithm solves Find-OV up to an additive error of t if it returns a set $S' \subseteq S$ for which $|S'| \geq |S| - t$.

2.3 Relevant prior work

We will apply the following theorem from [12] to our low-dimensional hardness result of exact EMD:

▶ **Theorem 6** ([12])**.** *Assuming OVC, there is a constant $c > 0$ such that Bichromatic ℓ_2-Closest Pair in $c^{\log^* n}$ dimensions requires $n^{2-o(1)}$ time, with vectors of $O(\log n)$ bit entries.*

3 Exact EMD in low dimensions

To prove hardness of the exact EMD problem under the Orthogonal Vectors Conjecture, we reduce to the bichromatic closest pairs problem, and then apply Theorem 6 due to [12]. The intuition for the reduction is as follows: given two sets A and B of n vectors, we'd like to augment set A with $n - 1$ copies of a vector that is equidistant from all of B, and much closer to B than A is. Similarly, we'd like to augment set B with $n - 1$ copies of a vector that is equidistant from all of A, and much closer to A than B is. If this were possible, then the minimum cost matching between the augmented sets would only match one pair of the original sets: the desired closest pair.

Unfortunately, it is in general impossible to find a vector equidistant from n vectors in $d \ll n$ dimensions. But this can be circumvented by embedding the vectors in a slightly higher-dimensional space, and adjusting coordinates in the "free" dimensions to ensure that an equidistant vector exists. So long as the free dimensions used to adjust set A are disjoint from the free dimensions used to adjust set B, the inner products between A and B are unaffected, and the distances change in an accountable way.

Since we are working in the ℓ_2 norm, we will need the following simple lemma which shows that any integer can be efficiently decomposed as a sum of a constant number of perfect squares.

▶ **Lemma 7.** *For any $\rho > 0$ and any positive integer m, there is an $O(m^\rho)$ time algorithm to decompose m as a sum of $O(\log 1/\rho)$ perfect squares.*

Proof. Here is the algorithm: repeatedly find the largest square which does not push the total above m, until the remainder does not exceed $O(m^{\rho/2})$. Then compute the minimal square decomposition for the remainder by dynamic programming.

The first, greedy phase takes $O(\mathrm{polylog}(m))$ time and finds $O(\log 1/\rho)$ perfect squares which sum to some m' with $m - m^{\rho/2} \le m' \le m$. The second, dynamic programming phase takes $O(m^\rho)$ time (even naively). By Lagrange's four-square theorem, a decomposition of $m - m'$ into at most four perfect squares is found. ◀

Now we describe the main reduction of this section. We'll use a shorthand notation to define vectors more concisely: for example, $a^x b^y c^z$ refers to an $(x + y + z)$-dimensional vector with value a in the first x dimensions, b in the next y dimensions, and c in the next z dimensions.

▶ **Theorem 8.** *Let $d = d(n) \le n$ be a dimension, and let $k > 0$ be a constant. There is a constant $c = c(k)$ for which the following holds. Suppose that there is an algorithm which computes the ℓ_2 earth mover distance between sets $A', B' \subseteq [1, n^{16k}]^{2d+2c+2}$ of size n in $O(n^{2-\epsilon})$ time. Then bichromatic closest pair between sets $A, B \subseteq [1, n^k]^d$ of size n can be computed in $O(n^{2-\epsilon})$ time as well.*

Proof. Set $\rho = 1/(16k)$, and let $c = O(\log 1/\rho)$ be the constant in Lemma 7 for the number of perfect squares in a decomposition. Let A and B be two sets of vectors from $\{1, \ldots, n^k\}^d$. Let $N = n^{16k}$. Our goal is to compute

$$\min_{a \in A, b \in B} \|a - b\|_2 \, .$$

We can assume without loss of generality that $\|a\|_2^2$ and $\|b\|_2^2$ are odd for all $a \in A$ and $b \in B$: for instance, we can replace each vector $z = (z_1, \ldots, z_d)$ by $(2z_1, \ldots, 2z_d, 1)$.

We construct sets A' and B' of $(2d + 2c + 2)$-dimensional vectors as follows. Let $u = 0^d (10^c) 0^{c+1} 0^d$ (parentheses for clarity). Let $v = N^d 0^{c+1} (10^c) 0^d$. Add $n - 1$ copies of u to B' and add $n - 1$ copies of v to A'. For each $a \in A$, add the following vector to A', where we'll define vector $\mathrm{adj}_a \in \mathbb{Z}^{c+1}$ later:

$$a' = f(a) = 0^d (\mathrm{adj}_a) 0^{c+1} a.$$

Similarly, for each $b \in B$, add the following vector to B', where we'll define $\mathrm{adj}_b \in \mathbb{Z}^{c+1}$ later:

$$b' = g(b) = N^d 0^{c+1} (\mathrm{adj}_b) b.$$

Now pick any $a \in A$. We'll construct adj_a so that the following equalities are both satisfied:

$$\|a' - u\|_2^2 = n^{4k} d^2 = \|\mathrm{adj}_a\|_2^2 \, .$$

Define the first element $\mathrm{adj}_a(0) = (\|a\|_2^2 + 1)/2$. Since $\|a\|_2^2 \le n^{2k} d$, we can then use Lemma 7 to find c integers $\mathrm{adj}_a(1), \ldots, \mathrm{adj}_a(c)$ so that $\|\mathrm{adj}_a\|_2^2 = n^{4k} d^2$. Furthermore,

$$\begin{aligned}
\|a' - u\|_2^2 &= \|\mathrm{adj}_a - 10^c\|_2^2 + \|a\|_2^2 \\
&= \|\mathrm{adj}_a\|_2^2 - 2 \cdot \mathrm{adj}_a(0) + 1 + \|a\|_2^2 \\
&= n^{4k} d^2.
\end{aligned}$$

For each $b \in B$, we can similarly construct adj_b so that $\|b' - v\|_2^2 = \|\mathrm{adj}_b\|_2^2 = n^{4k} d^2$.

We claim that

$$\text{EMD}(A', B') = 2(n-1)n^{2k}d + \min_{a \in A, b \in B} \sqrt{N^2 d + 2n^{4k}d^2 + \|a - b\|_2^2}.$$

To prove this claim, notice that $\|u - v\|_2 \geq N\sqrt{d}$ and $\|a' - b'\|_2 \geq N\sqrt{d}$ for every $a' \in A' \setminus \{v\}$ and $b' \in B' \setminus \{u\}$, whereas $\|a' - u\|_2 \ll N\sqrt{d}/n$ and $\|b' - v\| \ll N\sqrt{d}/n$. This means that the optimal matching between A' and B' will minimize the number of (u, v) and (a', b') edges. Hence, exactly one element of $A' \setminus \{v\}$ will be matched to an element in $B' \setminus \{u\}$. So if M denotes this optimal matching, and $x' = f(x) \in A'$ is matched with $y' = g(y) \in B'$, then the cost of M is

$$\text{cost}(M) = \left(\sum_{a' \in A' \setminus \{v, x'\}} \|a' - u\|_2 + \sum_{b' \in B' \setminus \{u, y'\}} \|b' - v\|_2 \right) + \|x' - y'\|_2$$

$$= 2(n-1)n^{2k}d + \sqrt{N^2 d + \|\text{adj}_x\|_2^2 + \|\text{adj}_y\|_2^2 + \|x - y\|_2^2}$$

$$= 2(n-1)n^{2k}d + \sqrt{N^2 d + 2n^{4k}d^2 + \|x - y\|_2^2}.$$

The claim follows. So the algorithm is simply: run the EMD algorithm on (A', B') and use the computed matching cost to find the closest pair distance, according to the above formula.

The time complexity of constructing A', B' is $O(n^{5/4}d^{1/8})$, dominated by computing a square decomposition for each vector. Since A' and B' are sets of $O(n)$ vectors in $\mathbb{Z}^{2d+2c+2}$ with entries bounded by $\max(N, n^{2k}d) \leq n^{16k}$, the EMD between A' and B' can be computed in $O(n^{2-\epsilon})$ time. Thus, the overall algorithm takes $O(n^{2-\epsilon})$ time. ◀

Theorem 1 follows from the above reduction and Theorem 6.

4 Approximate EMD under the Hitting Set Conjecture

In this section we prove hardness of approximation for the EMD matching problem when the approximately optimal matching must be reported. Note that the techniques from the previous section do not immediately generalize to this scenario, since the reduction in Theorem 8 is not approximation-preserving. A multiplicative error of $1 + \epsilon$ in the EMD algorithm would induce an additive error of $\tilde{O}(\epsilon n^{16k})$ in the closest pair algorithm, due to the large integers constructed in the reduction. A bucketing scheme, to ensure that the diameter of the input point set is within a constant factor of the closest pair, could eliminate the dependence on the values of the input coordinates, yielding a multiplicative error of only $1 + \tilde{O}(\epsilon n)$.

However, $(1 + \epsilon)$-approximate closest pair is only quadratically hard for $\epsilon = o(1)$ [21]; for any constant $\epsilon > 0$, there is a subquadratic $(1 + \epsilon)$-approximation algorithm [16, 4]. Thus, the above arguments would only yield $(1 + \tilde{O}(1/n))$-approximate hardness. Furthermore, the factor of n loss intuitively feels intrinsic to the approach of reducing from closest pair, since the EMD is the sum of n distances. Thus, a different approach seems necessary if we are to achieve hardness for $\epsilon = \omega(1/n)$.

Our method broadly consists of two steps. First, we show that EMD can encode orthogonality, by reducing approximate Maximum Orthogonal Matching (the problem of reporting a maximum matching in the implicit graph with an edge for each orthogonal pair) to approximate EMD matching. Second, we show that approximate Maximum Orthogonal matching can solve an instance (A, B) of Hitting Set by finding an orthogonal pair (a, b) for every $a \in A$ if possible, even if the set of orthogonal pairs does not constitute a matching.

We start by proving that asymmetric EMD matching reduces to EMD matching for the appropriate choices of error bounds. The reduction pads the smaller set of vectors A with a vector that is equidistant from the opposite set B, so that its contribution to the earth mover distance can be accounted for. Of course, it is first necessary to transform the vectors so that an equidistant vector exists.

▶ **Lemma 9.** *Suppose that $(1+\epsilon)$-approximate EMD matching in D dimensions can be solved in $T(n, D)$ time. Then $(1 + \epsilon)$-approximate asymmetric EMD matching in d dimensions can be solved with an additional additive factor of $n\epsilon\sqrt{d}$ in $T(n, 2d)$ time.*

Proof. Let $A, B \subseteq \{0, 1\}^d$ with $|A| \leq |B|$. Define sets $A', B' \subseteq \{0, 1\}^{2d}$ by mapping $a \in A$ to the vector

$$(a_1, \ldots, a_d, 1 - a_1, \ldots, 1 - a_d)$$

and similarly mapping $b \in B$ to

$$(b_1, \ldots, b_d, 1 - b_1, \ldots, 1 - b_d).$$

Then add $|B| - |A|$ copies of the zero vector to A'.

Now $|A'| = |B'|$, so we can run the approximate EMD algorithm on A' and B' to find some bijection $\pi : A' \to B'$ such that

$$\sum_{a' \in A'} \|a' - \pi(a')\|_2 \leq (1 + \epsilon) EMD(A', B').$$

Each vector $b' \in B'$ has $\|b'\|_2^2 = d$, so the distance from the zero vector to each match is exactly $\sqrt{d}$. And for any $a \in A$ and $b \in B$ which map to $a' \in A'$ and $b' \in B'$,

$$\|a' - b'\|_2^2 = 2 \|a - b\|_2^2.$$

Hence, the cost of π is

$$\sum_{a' \in A'} \|a' - \pi(a')\|_2 = (|B| - |A|)\sqrt{d} + \sqrt{2} \cdot \sum_{a \in A} \|a - \pi(a)\|_2$$

and the optimal cost is

$$EMD(A', B') = (|B| - |A|)\sqrt{d} + \sqrt{2} \cdot EMD(A, B).$$

It follows that

$$\sum_{a \in A} \|a - \pi(a)\|_2 \leq \frac{\epsilon}{\sqrt{2}}(|B| - |A|)\sqrt{d} + (1 + \epsilon)EMD(A, B),$$

which is the stated error bound. ◀

Next, we reduce approximate Maximum Orthogonal Matching to approximate asymmetric EMD matching. The general idea, given input sets (A, B), is to deform A and B so that orthogonal pairs (a, b) are mapped to pairs (a'', b'') with distance d_0, and all other pairs are mapped to pairs with distance at least $d_1 > d_0$. Then add $|A|$ auxiliary vectors to B, each with distance exactly d_1 from all vectors in A. Thus, in an optimal matching, each vector of A is either matched with an orthogonal vector at distance d_0, or some vector with distance exactly d_1. This introduces a nonlinearity, ensuring that in the additive matching cost, an

orthogonal pair's contribution is not "cancelled out" by the contribution of a pair with dot product 2, for instance. A similar trick was used by [8] in the context of edit distance, another "additive" metric.

The following simple lemma will be useful:

▶ **Lemma 10.** *There are maps* $\phi_1, \phi_2 : \{0,1\}^d \to \{0,1\}^{3d}$ *such that for any* $a, b \in \{0,1\}^d$,

$$\phi_1(a) \cdot \phi_2(b) = d - (a \cdot b).$$

Furthermore, the maps can be evaluated in $O(d)$ *time.*

Proof. Each dimension expands into three dimensions as follows:

$$a_i \mapsto (\phi_1(a)_{3i}, \phi_1(a)_{3i+1}, \phi_1(a)_{3i+2}) = (a_i, 1 - a_i, 1 - a_i)$$

$$b_i \mapsto (\phi_2(b)_{3i}, \phi_2(b)_{3i+1}, \phi_2(b)_{3i+2}) = (1 - b_i, b_i, 1 - b_i).$$

Then for each i,

$$\sum_{j=3i}^{3i+2} \phi_1(a)_j \phi_2(b)_j = a_i(1 - b_i) + (1 - a_i)b_i + (1 - a_i)(1 - b_i) = 1 - a_i b_i.$$

Summing over $i = 1, \ldots, d$ we get $\phi_1(a) \cdot \phi_2(b) = d - (a \cdot b)$ as desired. ◀

▶ **Lemma 11.** *Suppose that* $(1 + \epsilon)$-*approximate asymmetric EMD in* D *dimensions can be solved with an additional additive factor of* $n\epsilon\sqrt{D}$ *in* $T(n, D)$ *time. Then the Maximum Orthogonal Matching problem in* d *dimensions can be solved up to an additive factor of* $O(n\epsilon d)$ *in* $T(2n, 12d + 1)$ *time.*

Proof. Let $A, B \subseteq \{0,1\}^d$ with $|A| \leq |B| = n$. Define $A', B' \subseteq \{0,1\}^{3d}$ by $A' = \phi_1(A)$ and $B' = \phi_2(B)$, where ϕ_1, ϕ_2 are as defined in Lemma 10.

Let $d' = 3d$ for convenience. Now we construct sets $A'', B'' \subseteq \{0,1\}^{4d'+1}$ as follows, starting from sets A' and B'. We add $2d'$ dimensions to ensure that $\|a''\|_2^2 = \|b''\|_2^2 = d'$ for every $a'' \in A''$ and $b'' \in B''$ without changing the inner products. Add another $d' + 1$ dimensions, extending each $a'' \in A''$ so that $a''_{3d'+1} = 1$ and $a''_i = 0$ otherwise; and extend each $b'' \in B''$ so that $b''_{3d'+2} = 1$ and $b''_i = 0$ otherwise. Finally augment B'' with $|A|$ copies of the vector $v \in \{0,1\}^{4d'+1}$ with $3d'$ zeros followed by $d' + 1$ ones.

Notice that for every $a \in A$ and $b \in B$ corresponding to some $a'' \in A''$ and $b'' \in B''$,

$$\|a'' - b''\|_2^2 = \|a''\|_2^2 + \|b''\|_2^2 - 2a'' \cdot b'' = 2(d' + 1) - 2a'' \cdot b'' = 2a \cdot b + 4d + 2,$$

and

$$\|a'' - v\|_2^2 = 2(d' + 1) - 2a'' \cdot v = 4d + 4.$$

Now we run the approximate asymmetric EMD matching algorithm on A'' and B'', yielding an injection $\pi : A'' \to B''$ such that

$$\sum_{a'' \in A''} \|a'' - \pi(a'')\|_2 \leq |B''|\epsilon\sqrt{4d' + 1} + (1 + \epsilon)EMD(A'', B'').$$

For each $a'' \in A''$, if $\|a'' - \pi(a'')\|_2^2 > 4d + 4$, then we can set $\pi(a'') = v$, preserving injectivity and decreasing the cost of the matching. Therefore every edge has cost either $\sqrt{4d + 2}$ or $\sqrt{4d + 4}$. In particular, if there are m orthogonal pairs in the matching, the total cost is

$$\sum_{a'' \in A''} \|a'' - \pi(a'')\|_2 = m\sqrt{4d + 2} + (|A| - m)\sqrt{4d + 4}.$$

By the same argument as above, the minimum cost matching is obtained by maximizing the number of orthogonal pairs. If the maximum possible number of orthogonal pairs in a matching is m_{OPT}, then

$$EMD(A'', B'') = m_{\mathrm{OPT}}\sqrt{4d+2} + (|A| - m_{\mathrm{OPT}})\sqrt{4d+4}.$$

Substituting these expressions into the approximation guarantee and solving, we get that $m \geq m_{\mathrm{OPT}} - O(\epsilon n d)$ as desired.   ◄

In the above lemma we assumed that we are given an algorithm for asymmetric EMD matching which has both a multiplicative error of $1 + \epsilon$ and an additive error of $n\epsilon\sqrt{d}$, since this is the error introduced by the reduction to (symmetric) EMD. However, we are also interested in the hardness of $(1 + \epsilon)$-approximate asymmetric EMD matching in its own right. Removing the additive error from the hypothesized algorithm in Lemma 11 directly translates to an improved Maximum Orthogonal Matching algorithm, with an additive error of $O(\epsilon|A|d)$ instead of $O(\epsilon n d)$, where $n = |A| + |B|$:

▶ **Lemma 12.** *Suppose that there is an algorithm which solves $(1+\epsilon)$-approximate asymmetric EMD matching in $T(|A| + |B|, d)$ time, where the input is $A, B \subseteq \{0,1\}^d$. Then the Maximum Orthogonal Matching problem can be solved up to an additive error of $O(\epsilon|A|d)$ in $T(2n, 12d + 1)$ time.*

Now we could reduce OV to approximate Maximum Orthogonal Matching. The proof of the following theorem is given in Appendix B for completeness.

▶ **Theorem 13.** *Let $d = \omega(\log n)$. Under the Orthogonal Vectors Conjecture, for any $\epsilon > 0$ and $\delta \in (0,1)$, $(1 + 1/n^\delta)$-approximate EMD matching in $\{0,1\}^d$ cannot be solved in $O(n^{2\delta - \epsilon})$ time.*

However, Theorem 13 does not prove quadratic hardness for any approximation factor larger than $(1 + 1/n)$, and in fact breaks down completely for $(1 + 1/\sqrt{n})$-approximate EMD matching.

Instead, we reduce Hitting Set to approximate Maximum Orthogonal Matching, through approximate Find-OV. These two problems are structurally similar; the technical difficulty is that Find-OV may require finding many orthogonal pairs even when the largest orthogonal matching may be small, in which case applying the Maximum Orthogonal Matching algorithm would result in little progress. We resolve this with the following insight: if many vectors in set A are orthogonal to at least one vector in set B but there is not a large orthogonal matching, then some vector in set B is orthogonal to *many* vectors in A. But these vectors can be found efficiently by sampling.

In the proof of the following theorem we formalize the above idea.

▶ **Theorem 14.** *Let $d = d(n)$ be a dimension. Suppose that the Maximum Orthogonal Matching problem can be solved up to an additive error of $E(|A|, |B|)$ in $O(n^{2-\epsilon}poly(d))$ time, where the input is $A, B \subseteq \{0,1\}^d$. Then for any (sufficiently small) $\alpha > 0$ there is some $\gamma > 0$ such that Find-OV can be solved with high probability up to an additive error of $E(|A|, 2|B|^{1+\alpha})$ in $O(n^{2-\gamma}poly(d))$ time.*

Proof. Let $A, B \subseteq \{0,1\}^d$ with $|A| = |B| = n$. Let $\alpha > 0$ be a constant we choose later. We may safely assume that $\alpha < 1$. Let the degree of a vector $a \in A$, denoted $d(a)$, be the number of $b \in B$ which are orthogonal to a. The algorithm for Find-OV consists of three steps:

1. For every $a \in A$, sample $n^{1-\alpha/4}$ vectors from B to get an estimate $\hat{d}(a)$ of $d(a)$. Mark and remove the vectors for which $\hat{d}(a) \geq n^{\alpha/2}$.
2. Next, for every $b \in B$, sample $n^{1-\alpha/2}$ vectors from A to get an estimate $\hat{d}(b)$ of $d(b)$. Let $B_{\text{large}} \subseteq B$ be the set of vectors for which $\hat{d}(b) \geq n^{\alpha}$. For each $b \in B_{\text{large}}$, iterate over A and mark and remove each $a \in A$ for which $a \cdot b = 0$. Now remove B_{large} from B.
3. Run the Maximum Orthogonal Matching algorithm on the remaining set A, and the multiset consisting of $2n^{\alpha}$ copies of each remaining $b \in B$. This produces a set of pairs (a_i, b_i) where $a_i \cdot b_i = 0$. Output the union of $\{a_i\}_i$ and the set of all vectors marked and removed from A in the previous steps.

In the first step, a Chernoff bound shows that with high probability, every vector for which $d(a) \geq 2n^{\alpha/2}$ is marked and removed. Now summing over the remaining vectors,

$$\sum_{a \in A} d(a) = \sum_{b \in B} d(b) \leq 2n^{1+\alpha/2}.$$

In the second step, with high probability B_{large} contains no $b \in B$ for which $d(b) \leq \frac{1}{2}n^{\alpha}$, by a Chernoff bound on each such $b \in B$. Therefore $|B_{\text{large}}| \leq 4n^{1-\alpha/2}$. Furthermore, with high probability B_{large} contains every $b \in B$ for which $d(b) \geq 2n^{\alpha}$.

So after the first two steps, every remaining vector $b \in B$ has degree at most $2n^{\alpha}$. Suppose there are t vectors $a \in A$ with positive degree, and t' of these are found in the first two steps. Then by the degree bound, the remaining $t - t'$ vectors inject into $2n^{\alpha}$ copies of B. Therefore there is an orthogonal matching of size at least $t - t'$. By the approximation guarantee of the Maximum Orthogonal Matching algorithm, we find an orthogonal matching of size at least $t - t' - 2n^{(1+\alpha)\delta}$ in step 3. Overall, we find at least $t - 2n^{(1+\alpha)\delta}$ vectors with positive degree, which gives the desired approximation guarantee.

The time complexity is $O((n^{2-\alpha/4} + n^{(2-\epsilon)(1+\alpha)})\text{poly}(d))$. This is subquadratic in n for sufficiently small α.   ◀

As the final step of the reduction, we show that approximate Find-OV can solve Hitting Set. Note that exact Find-OV obviously solves Hitting Set. It's also clear that Find-OV with an additive error of $n^{1-\epsilon}$ solves Hitting Set: simply run Find-OV, and then exhaustively check the remaining unpaired vectors of A – unless there are more than $n^{1-\epsilon}$ unpaired vectors, in which case there must be a hitting vector.

To reduce Hitting Set to Find-OV with additive error of $\Theta(n)$, the essential idea is simply to repeatedly run Find-OV on the remaining unpaired vectors. If the Find-OV algorithm has an additive error of $n/2$, then given an input A, B with no hitting vector, the algorithm will find orthogonal pairs for at least $n/2$ vectors of A. Naively, we'd like to recurse on the remaining half of A. Unfortunately, the set B cannot similarly be halved, so the error bound in the next step would not be halved. Thus, the algorithm might make no further progress.

The workaround is to duplicate every unpaired vector of A before recursing. If $n/2$ orthogonal pairs are found but every vector of A has been duplicated once, then matches are found for at least $n/4$ distinct vectors. This suffices to terminate the recursion in $O(\log n)$ steps.

▶ **Theorem 15.** *Suppose that Find-OV in d dimensions can be solved up to additive error of $n/2$ in $T(n, d)$ time. Then Hitting Set in d dimensions can be solved in $O((T(n, d) + nd) \log n)$ time.*

Proof. Let $A, B \subseteq \{0, 1\}^d$ with $|A| = |B| = n$. Our hitting set algorithm consists of $t = \lceil \log n \rceil + 1$ phases. Initialize $R_1 = A$.

In phase $i \geq 1$, run Find-OV on $(2^{i-1}R_i, B)$, where $2^i R_i$ is the multiset with 2^i copies of each vector in R_i. Let $P \subseteq A$ be the output multiset and let P' be the corresponding set (removing duplicates). Set $R_{i+1} = R_i \setminus P'$. If $|R_{i+1}| > n/2^i$, report failure (i.e. there is a hitting vector). Otherwise, proceed to the next phase. If phase t is complete, report success (i.e. no hitting vector).

Suppose that the algorithm reports success. Then after phase t, we have $R_{t+1} \leq n/2^t < 1$. Then for every $a \in A$ there was some phase i in which a was removed from R_i, and therefore was orthogonal to some $b \in B$. So there is no hitting vector.

Suppose that the algorithm reports failure in phase i. Then $|R_i| \leq n/2^{i-1}$ and $|R_{i+1}| > n/2^i$, so $|P'| < n/2^i$. Therefore $|P| \leq 2^{i-1}|P'| < n/2$. By the Find-OV approximation guarantee, not every element of R_i is orthogonal to an element of B. So there is a hitting vector.

The time complexity is dominated by $O(\log n)$ applications of Find-OV on inputs of size $O(n)$, along with $O(nd)$ extra processing in each phase. Thus, the time complexity is $O((T(n,d) + nd)\log n)$.   ◄

The next theorem shows that hardness for approximate EMD matching (conditioned on the Hitting Set Conjecture) follows from chaining together the above reductions.

▶ **Theorem 16.** *If there are any $\epsilon, \delta > 0$ such that $(1+1/n^\delta)$-approximate EMD matching can be solved in $O(n^{2-\epsilon})$ time for some dimension $d = \omega(\log n)$, then the Hitting Set Conjecture is false.*

Proof. Fix $d = \omega(\log n)$, and assume without loss of generality that $d(n)$ is polylogarithmic. Let $\epsilon, \delta > 0$ and suppose that $(1+1/n^\delta)$-approximate EMD matching can be solved in $O(n^{2-\epsilon})$ time. Then $(1 + 1/n^\delta)$-approximate asymmetric EMD can be solved with an additional additive error of $n^{1-\delta}\sqrt{d}$ with the same time complexity, by Lemma 9. Hence, the Maximum Orthogonal Matching problem can be solved with an additive error of $n^{1-\delta}d$ in the same time, by Lemma 11.

Applying Theorem 14 with parameter $\alpha = \delta$, we get a randomized algorithm for Find-OV with an additive error of $O(n^{1-\delta^2}d^{1+\delta})$ and time complexity $O(n^{2-\gamma})$ for some $\gamma > 0$. For sufficiently large n, the error is at most $n/2$. Thus, we can apply Theorem 15 to get a randomized algorithm for Hitting Set with time complexity $\tilde{O}(n^{2-\gamma})$, which contradicts the Hitting Set Conjecture.   ◄

Furthermore, we obtain stronger hardness of approximation for asymmetric EMD matching:

▶ **Theorem 17.** *Let $d = \omega(\log n)$ and $\eta = 1/\omega(\log n)$. If there is a truly subquadratic $(1 + \eta)$-approximation algorithm for asymmetric EMD matching in d dimensions, then the Hitting Set Conjecture is false.*

Proof. Fix $d' = \omega(\log n)$ and $\eta = 1/\omega(\log n)$ and $\epsilon > 0$. Suppose that there is an $O(n^{2-\epsilon})$ time algorithm which achieves a $(1 + \eta)$ approximation for asymmetric EMD matching in d' dimensions. Set $d = \min(d', \sqrt{(\log n)/\eta})$. Since $\mathbb{R}^d$ embeds isometrically in $\mathbb{R}^{d'}$, the algorithm also achieves a $(1 + \eta)$ approximation for asymmetric EMD in d dimensions.

By Lemma 12, the Maximum Orthogonal Matching problem can be solved up to an additive error of $O(\eta|A|d)$ in $O(d)$ dimensions and $O(n^{2-\epsilon})$ time. By Theorem 14 there is some $\gamma > 0$ such that Find-OV can be solved up to an additive error of $O(\eta nd)$ in $O(d)$ dimensions and $O(n^{2-\gamma})$ time. By choice of d we have $\eta nd = o(n)$, so for sufficiently large n the algorithm achieves additive error of at most $n/2$. Therefore by Theorem 15, Hitting Set can be solved in $O(d)$ dimensions and $\tilde{O}(n^{2-\epsilon})$ time. Since $d = \omega(\log n)$, this contradicts the Hitting Set Conjecture.   ◄

References

1 Amir Abboud, Arturs Backurs, and Virginia Vassilevska Williams. Tight Hardness Results for LCS and Other Sequence Similarity Measures. In *Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, FOCS '15, pages 59–78, Washington, DC, USA, 2015. IEEE Computer Society. `doi:10.1109/FOCS.2015.14`.

2 Amir Abboud, Virginia Vassilevska Williams, and Joshua Wang. Approximation and Fixed Parameter Subquadratic Algorithms for Radius and Diameter in Sparse Graphs. In *Proceedings of the Twenty-seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '16, pages 377–391, Philadelphia, PA, USA, 2016. Society for Industrial and Applied Mathematics. URL: `http://dl.acm.org/citation.cfm?id=2884435.2884463`.

3 Pankaj K. Agarwal, Kyle Fox, Debmalya Panigrahi, Kasturi R. Varadarajan, and Allen Xiao. Faster Algorithms for the Geometric Transportation Problem. In Boris Aronov and Matthew J. Katz, editors, *33rd International Symposium on Computational Geometry (SoCG 2017)*, volume 77 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 7:1–7:16, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. `doi:10.4230/LIPIcs.SoCG.2017.7`.

4 Josh Alman, Timothy M. Chan, and Ryan Williams. Polynomial Representations of Threshold Functions and Algorithmic Applications. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 467–476, October 2016. `doi:10.1109/FOCS.2016.57`.

5 Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear Time Approximation Algorithms for Optimal Transport via Sinkhorn Iteration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 1961–1971, USA, 2017. Curran Associates Inc. URL: `http://dl.acm.org/citation.cfm?id=3294771.3294958`.

6 Alexandr Andoni, Aleksandar Nikolov, Krzysztof Onak, and Grigory Yaroslavtsev. Parallel Algorithms for Geometric Graph Problems. In *Proceedings of the Forty-sixth Annual ACM Symposium on Theory of Computing*, STOC '14, pages 574–583, New York, NY, USA, 2014. ACM. `doi:10.1145/2591796.2591805`.

7 Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint*, 2017.

8 Arturs Backurs and Piotr Indyk. Edit Distance Cannot Be Computed in Strongly Subquadratic Time (Unless SETH is False). In *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*, STOC '15, pages 51–58, New York, NY, USA, 2015. ACM. `doi:10.1145/2746539.2746612`.

9 Amitabh Basu. Open Problem: Maximum weighted assignment problem. In *Workshop: Combinatorial Optimization*, Oberwolfach Report 50/2018, page 44, 2018. `doi:10.4171/OWR/2018/50`.

10 Nicolas Bonneel, Michiel van de Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement Interpolation Using Lagrangian Mass Transport. *ACM Transactions on Graphics*, 30(6):158:1–158:12, December 2011. `doi:10.1145/2070781.2024192`.

11 Karl Bringmann and Marvin Kunnemann. Quadratic Conditional Lower Bounds for String Problems and Dynamic Time Warping. In *Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, FOCS '15, pages 79–97, Washington, DC, USA, 2015. IEEE Computer Society. `doi:10.1109/FOCS.2015.15`.

12 Lijie Chen. On the Hardness of Approximate and Exact (Bichromatic) Maximum Inner Product. In *Proceedings of the 33rd Computational Complexity Conference*, CCC '18, pages 14:1–14:45, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. `doi:10.4230/LIPIcs.CCC.2018.14`.

13 Rémi Flamary, Marco Cuturi, Nicolas Courty, and Alain Rakotomamonjy. Wasserstein Discriminant Analysis. *Machine Learning*, 107(12):1923–1945, December 2018. `doi:10.1007/s10994-018-5717-1`.

14 John Hopcroft and Richard Karp. An $n^{5/2}$ Algorithm for Maximum Matchings in Bipartite Graphs. *SIAM Journal on Computing*, 2(4):225–231, 1973. `doi:10.1137/0202019`.

15   Piotr Indyk. A Near Linear Time Constant Factor Approximation for Euclidean Bichromatic Matching (Cost). In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 39–42, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics. URL: `http://dl.acm.org/citation.cfm?id=1283383.1283388`.

16   Piotr Indyk and Rajeev Motwani. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, pages 604–613, New York, NY, USA, 1998. ACM. `doi:10.1145/276698.276876`.

17   Andrey Boris Khesin, Aleksandar Nikolov, and Dmitry Paramonov. Preconditioning for the Geometric Transportation Problem. In Gill Barequet and Yusu Wang, editors, *35th International Symposium on Computational Geometry (SoCG 2019)*, volume 129 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 15:1–15:14, Dagstuhl, Germany, 2019. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. `doi:10.4230/LIPIcs.SoCG.2019.15`.

18   Yin Tat Lee and Aaron Sidford. Path Finding II: An $\tilde{O}(m\sqrt{n})$ Algorithm for the Minimum Cost Flow Problem. *arXiv preprint*, 2013. `arXiv:1312.6713`.

19   Yin Tat Lee and Aaron Sidford. Path Finding Methods for Linear Programming: Solving Linear Programs in Õ(Vrank) Iterations and Faster Algorithms for Maximum Flow. In *Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, FOCS '14, pages 424–433, Washington, DC, USA, 2014. IEEE Computer Society. `doi:10.1109/FOCS.2014.52`.

20   Jonas Mueller and Tommi Jaakkola. Principal Differences Analysis: Interpretable Characterization of Differences Between Distributions. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 1702–1710, Cambridge, MA, USA, 2015. MIT Press. URL: `http://dl.acm.org/citation.cfm?id=2969239.2969429`.

21   Aviad Rubinstein. Hardness of Approximate Nearest Neighbor Search. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, pages 1260–1268, New York, NY, USA, 2018. ACM. `doi:10.1145/3188745.3188916`.

22   Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121, November 2000. `doi:10.1023/A:1026543900054`.

23   Roman Sandler and Michael Lindenbaum. Nonnegative Matrix Factorization with Earth Mover's Distance Metric for Image Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1590–1602, August 2011. `doi:10.1109/TPAMI.2011.18`.

24   R. Sharathkumar and Pankaj K. Agarwal. A Near-linear Time ϵ-approximation Algorithm for Geometric Bipartite Matching. In *Proceedings of the Forty-fourth Annual ACM Symposium on Theory of Computing*, STOC '12, pages 385–394, New York, NY, USA, 2012. ACM. `doi:10.1145/2213977.2214014`.

25   Justin Solomon, Fernando de Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains. *ACM Transactions on Graphics*, 34(4):66:1–66:11, July 2015. `doi:10.1145/2766963`.

26   Cédric Villani. *Topics in optimal transportation*. Number 58 in Graduate Studies in Mathematics. American Mathematical Society, 2003.

27   Ryan Williams. A new algorithm for optimal 2-constraint satisfaction and its implications. *Theoretical Computer Science*, 348(2):357–365, 2005. Automata, Languages and Programming: Algorithms and Complexity (ICALP-A 2004). `doi:10.1016/j.tcs.2005.09.023`.

28   Ryan Williams. On the Difference Between Closest, Furthest, and Orthogonal Pairs: Nearly-linear vs Barely-subquadratic Complexity. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, pages 1207–1215, Philadelphia, PA, USA, 2018. Society for Industrial and Applied Mathematics. URL: `http://dl.acm.org/citation.cfm?id=3174304.3175348`.

29   Virginia Vassilevska Williams. On some fine-grained questions in algorithms and complexity. In *Proceedings of the ICM*, 2018.

A   Hardness of Low-Rank Minimum Weighted Assignment

The methods we used to prove hardness of exact EMD in low dimensions can be adapted to prove hardness of minimum weighted assignment with low-rank weight matrices, under the Orthogonal Vectors Conjecture. In particular, we show in the following theorem that bichromatic closest pair in d dimensions can be reduced to minimum weighted assignment with a rank-$O(d)$ weight matrix. The reduction algorithm uses the same input transformation as Theorem 8, and then solves minimum weighted assignment on the matrix M with entries $M_{ij} = \left\| A'_i - B'_j \right\|_2^2$, where A' and B' are the transformed input sets. The key is that M has rank $O(d)$, and its minimum weight assignment encodes the squared closest pair distance of the input – just as the EMD of the transformed input in Theorem 8 encoded the closest pair distance of the input.

▶ **Theorem 18.** *Fix a dimension $d = d(n) \leq n$, and let $\epsilon > 0$. Suppose that there is an algorithm which solves minimum weighted assignment in $O(n^{2-\epsilon})$ time, if the weight matrix has rank at most $O(d)$. Then bichromatic closest pair in d dimensions can be solved in $O(n^{2-\epsilon})$ time.*

Proof. Let A and B be two sets of n vectors in d dimensions, with entries in $\{1, \ldots, n^k\}$ for some constant $k > 0$. Apply the transformation described in Theorem 8 to construct sets $A', B' \in \{0, \ldots, n^{16k}\}^{2d+2c+2}$ where c is as defined in the proof of the theorem. Define

$$\mathrm{SQEMD}(A', B') = \min_{\sigma : A' \to B'} \sum_{a' \in A'} \|a' - \sigma(a')\|_2^2$$

where σ ranges over all bijections from A' to B'. Since $\|u - v\|_2^2 \geq N^2 d$ and $\|a' - b'\|_2^2 \geq N^2 d$ for every $a' \in A' \backslash \{v\}$ and $b' \in B' \backslash \{u\}$, whereas $\|a' - u\|_2^2 \ll N^2 d/n$ and $\|b' - v\|_2^2 \ll N^2 d/n$, the optimal matching σ minimizes the number of (u, v) and (a', b') edges. In particular, exactly one element of $A' \backslash \{v\}$ is matched to an element of $B' \backslash \{u\}$. Thus, paralleling the proof of Theorem 8, we get

$$\mathrm{SQEMD}(A', B') = 2(n-1)n^{4k}d^2 + \left(N^2 d + 2n^{4k}d^2 + \min_{a \in A, b \in B} \|a - b\|_2^2 \right).$$

Hence, to compute the bichromatic closest pair distance between A and B, it suffices to compute $\mathrm{SQEMD}(A', B')$. Representing A' and B' as $n \times (2d + 2c + 2)$ matrices, let M be the $n \times n$ matrix defined by $M_{ij} = \left\| A'_i - B'_j \right\|_2^2$. Then observing that

$$M_{ij} = \sum_{k=1}^{2d+2c+2} (A'_{ik} - B'_{jk})^2 = \sum_{k=1}^{2d+2c+2} (A'_{ik})^2 + \sum_{k=1}^{2d+2c+2} (B'_{jk})^2 - 2 \sum_{k=1}^{2d+2c+2} A'_{ik} B'_{jk},$$

we can write M as the sum of $2d + 2c + 4$ rank-1 matrices, so $\mathrm{rank}(M) \leq 2d + 2c + 4$. So by assumption, the minimum weight perfect matching in the complete bipartite graph determined by M can be found in $O(n^{2-\epsilon}\mathrm{poly}(d))$ time. But the cost of the optimal matching is precisely $\mathrm{SQEMD}(A', B')$.   ◀

Applying Theorem 6 completes the proof of Theorem 2.

B   Proof of Theorem 13

The theorem follows immediately from the reduction from Maximum Orthogonal Matching to EMD matching shown in section 4, and this next proposition.

▶ **Proposition 19.** *Suppose the Maximum Orthogonal Matching problem can be solved up to an additive factor of n^δ in $O(n^\gamma)$ time where $\delta < 1/2$. Then OV can be solved in $O(n^{\gamma/(1-\delta)})$ time.*

Proof. Let $A, B \subseteq \{0,1\}^d$ with $|A| = |B| = n$. We construct multisets A' and B' which consist of $2n^{\delta/(1-\delta)}$ copies of each $a \in A$, and $2n^{\delta/(1-\delta)}$ copies of each $b \in B$, respectively. We then run our approximate Maximum Orthogonal Matching algorithm on A' and B'. If any orthogonal pair is found, we return it; otherwise we return that there is no orthogonal pair.

Since $|A'| = |B'| = 2n^{1/(1-\delta)}$, the time complexity of this algorithm is $O(n^{\gamma/(1-\delta)})$. It is clear that if A and B have no orthogonal pair, then A' and B' have no orthogonal pair, so the algorithm correctly returns "no pair".

Suppose that there are $a \in A$ and $b \in B$ with $a \cdot b = 0$ but the algorithm returns "no pair". Then the matching found by the algorithm had no orthogonal pairs. However, there is a matching consisting of $2n^{\delta/(1-\delta)}$ pairs. Since $|B'|^\delta < 2n^{\delta/(1-\delta)}$, this contradicts the approximation guarantee of the Maximum Orthogonal Matching algorithm. ◀

C Hardness of $(k, 2k)$-Find-OV

The $(k, 2k)$-Find-OV problem provides some sense of the relative "powers" of the Orthogonal Vectors Conjecture and the Hitting Set Conjecture, as well as another example of how the Hitting Set Conjecture can be used to explain hardness of approximation problems. Reducing from OV, we get the following hardness result, and it is not clear how to make any improvement. Note that this proof extends to the $(1, 2k)$-Find-OV problem, for which this lower bound is tight, due to a random sampling algorithm.

▶ **Proposition 20.** *Fix $\delta \in (0,1)$. Assuming OVC, any algorithm for $(n^\delta, 2n^\delta)$-Find-OV requires $\Omega(n^{2-\delta-o(1)})$ time.*

Proof. Suppose that there exists an $O(n^{2-\delta-\epsilon})$ time algorithm FIND for $(n^\delta, 2n^\delta)$-Find-OV. Here is an algorithm for OV: given sets $A, B \subseteq \{0,1\}^d$ with $|A| = |B| = n$, duplicate each $a \in A$ and each $b \in B$ exactly $2n^{\delta/(2-\delta)}$ times. If the original number of orthogonal pairs was r, then the new number is $4rn^{2\delta/(2-\delta)}$. For $r \geq 1$, this exceeds $2(n \cdot 2n^{\delta/(2-\delta)})^\delta$, so applying FIND yields a positive number of orthogonal vectors if and only if $r > 0$. It's easy to check that the time complexity is subquadratic. ◀

On the other hand, under the Hitting Set Conjecture, we can obtain quadratic hardness. When $k = n/2$, hardness follows from Theorem 15, but it holds in greater generality. In particular, we provide a proof of conditional hardness for $k = \sqrt{n}$, and it extends naturally to any $k = n^\gamma$ for $\gamma \in (0,1)$. The proof takes inspiration from the reduction from Hitting Set to OV [2], with a few extra twists.

▶ **Theorem 21.** *If the $(\sqrt{n}, 2\sqrt{n})$-Find-OV problem can be solved in $O(n^{2-\epsilon})$ time for some $\epsilon > 0$, then Hitting Set can be solved in $O(n^{2-\delta})$ time for some $\delta > 0$.*

Proof. Let FIND be the presupposed algorithm for $(\sqrt{n}, 2\sqrt{n})$-Find-OV. Set $\alpha = \epsilon/7$. Let $A, B \subseteq \{0,1\}^d$ with $|A| = |B| = n$. Without loss of generality, assume that no vector is all-zeroes. Here is an algorithm:
1. For each $a \in A$, randomly sample $n^{1-\alpha}$ vectors from B. If any of these is orthogonal to a, mark a and remove it from A, replacing it with an all-ones vector.
2. Set $k = n^{1/3-\alpha}$. Partition A into sets $A_1, \ldots, A_k$ of approximately equal size, and similarly partition B into sets $B_1, \ldots, B_k$. For each pair (A_i, B_j):

a. Apply FIND to (A_i, B_j).

b. If the output is not $\sqrt{n/k}$ orthogonal pairs, then continue to the next pair (A_i, B_j).

c. Otherwise, suppose that the output is $\{(a_m, b_m)\}_{m=1}^{\sqrt{n/k}}$. For each vector $a \in \{a_m\}_{m=1}^{\sqrt{n/k}}$, mark a and remove it from A_i (and from A), replacing it with an all-ones vector.

d. Go to (a).

3. If the number of unmarked input vectors exceeds $2n^{1-3\alpha/2}$, return "NO" and exit.

4. For each $a \in A$, if a is not the all-ones vector, iterate over all $b \in B$, and mark a if any $b \in B$ is orthogonal.

5. Return "YES" if every vector originally in A is now marked, and "NO" otherwise.

We claim that this algorithm solves Hitting Set in strongly subquadratic time. Correctness is relatively simple: a vector $a \in A$ is only marked by the above algorithm if some $b \in B$ is found for which $a \cdot b = 0$. Thus, if some $a \in A$ is a hitting vector for B, then it is never marked, so the algorithm returns "NO".

Conversely, suppose that every $a \in A$ is orthogonal to some $b \in B$. Then the number of unmarked input vectors in Step 3 is at most the number of remaining orthogonal pairs. But each (A_i, B_j) contains at most $2\sqrt{n/k}$ orthogonal pairs after Step 2 finishes, so the number of remaining orthogonal pairs in Step 3 is at most $k^2(2\sqrt{n/k}) = 2n^{1-3\alpha/2}$. Thus, the algorithm continues to Step 4. Every $a \in A$ which has not been marked by the end of Step 2 is tested against every $b \in B$ in Step 4. Therefore every vector is marked, so the algorithm returns "YES".

Turning to time complexity, Step 1 takes $O(n^{2-\alpha})$ time. The complexity of Step 2 is dominated by the calls to FIND. For each pair (A_i, B_j) there is at most one call to FIND for which the output is not $\sqrt{n/k}$ orthogonal pairs. Hence, there are $k^2 = n^{2/3-2\alpha}$ such "failed" calls. To bound the number of "successful" calls to FIND, for which the output *is* $\sqrt{n/k}$ orthogonal pairs, note that after Step 1, with high probability each $a \in A$ is orthogonal to at most $n^{2\alpha}$ vectors $b \in B$, so the total number of orthogonal pairs is at most $n^{1+2\alpha}$. Each successful call eliminates $\sqrt{n/k} = n^{1/3+\alpha/2}$ orthogonal pairs, so there are at most $n^{2/3+3\alpha/2}$ successful calls. This bound dominates the bound on failed calls. Each call takes time $O((n/k)^{2-\epsilon})$, so the time complexity of Step 2 is asymptotically

$$n^{\left(\frac{2}{3}+\alpha\right)(2-\epsilon)}n^{\frac{2}{3}+\frac{3\alpha}{2}} = n^{2-\frac{\epsilon}{6}-\frac{\epsilon^2}{7}}.$$

Step 3 takes negligible time. Finally, in Step 4, there are at most $2n^{1-3\alpha/2}$ vectors $a \in A$ which are not the all-ones vector (since each of these is unmarked), so the complexity is $O(n^{2-3\alpha/2})$.

Hence, the overall time complexity is bounded by $O(n^{2-\epsilon/7})$. ◀