On the relation between the relative earth mover distance and the variation distance (an exposition)

Oded Goldreich Dana Ron

August 21, 2019

Abstract

The "relative earth mover distance" is a technical term introduced by Valiant and Valiant (*43rd STOC*, 2011), and extensively used in their work. They claimed that, for every two distributions, the relative earth mover distance upper-bounds the variation distance up to relabeling, but this claim was not used in their work. The claim appears as a special case of a result proved by Valiant and Valiant in a later work (*48th STOC*, 2016), but we found their proof too terse. The proof presented here is merely an elaboration of (this special case of) their proof.

This is a drastic revision of a text that was posted on the first author's web-site in February 2016.[1]

1 Introduction

The (total) variation distance between (discrete) distributions, defined as half the ℓ_1-norm of their difference, is the most popular notion of distance between distributions (viewed as functions from their domain to $[0,1]$).[2] The natural appeal of this distance measure comes from the fact that it equals the best possible distinguishing probability (i.e., the difference in the verdict of an observer that is given a single sample from one of the two distributions).

In particular, the variation distance is the measure of choice in the context of distribution testing. In that context, given samples from an unknown distribution X the task is to test whether X has some predetermined property (i.e., is in a predetermined set of distributions) or is far from having this property (i.e., is far from any distribution in the class).

Testing label-invariant properties is of special interest, where a property is label-invariant if for every distribution (viewed as a function) $p : D \to [0,1]$ and every bijection $\pi : D \to D$ it holds that p has the property if and only if $p \circ \pi$ has the property. In this case, we implicitly care about the variation distance *up to relabeling*, as defined below (see Definition 2.3).

One of the most striking results in area of distribution testing asserts that *every label-invariant property of distributions over* $[n]$ *can be tested using* $s(n,\epsilon) = O(\epsilon^{-2} \cdot n/\log n)$ *samples.* This celebrated result of Valiant and Valiant is not stated explicitly (in these terms) in their paper [3], since they find it easier to work with an alternative notion of distance that they define, which upper-bounds the variation distance up to relabeling. The latter claim is made and proved in their

[1] See `http://www.wisdom.weizmann.ac.il/~oded/p_remd.html`

[2] Specifically, for two distributions presented by their probability functions $p, q : D \to [0,1]$, their variation distance equals $0.5 \cdot \sum_{i \in D} |p(i) - q(i)|$, which in turn equals $\min_{S \subseteq D}\{p(S) - q(S)\}$, where $p(S) = \sum_{i \in S} p(i)$. The set S may be viewed as the set of samples on which an observer (as discussed next) outputs the verdict 1.

subsequent work [4, 5]; actually, it appears there as a special case of the more general Fact 1 (i.e., the case of $\tau = 0$), but we found their proof too terse.

Given the importance of the results derived by the foregoing claim, which appears below as Theorem 3.1, we believe that it is important to provide and make accessible a more detailed proof of this claim. In fact, we wrote a sketch of the current document when working on [2], where we relied on this claim. Furthermore, this claim is pivotal to the exposition in [1, Sec. 11.4], which highlights the aforementioned result of Valiant and Valiant [3] as well as its optimality. These results appear as [1, Cor. 11.28] and [1, Thm. 11.29], respectively, and both of them rely on Theorem 3.1.

Unfortunately, the rest of this text is quite technical. We start with the definitions that constitute the alternative notion of distance defined by Valiant and Valiant [3], which is called the *relative earth-mover distance*. This definition (i.e., Definition 2.2) actually refers to the "relative histrograms" of the distributions, which are defined first (see Definition 2.1). The (simpler) non-relative notions play a pivotal role in the proof, so they are defined too (in Section 2). (For sake of good order, we also define the variation distance up to relabeling (see Definition 2.3).) With all these definitions in place, the foregoing claim is stated as Theorem 3.1 and proved in Section 3. The proof is quite technical and is merely an elaboration of (the special case of) the proof presented by Valiant and Valiant (in [4, Apdx A] and [5, Apdx B]).

2 Definitions

We start by introducing two label-invariant representations of distributions.

Definition 2.1 (histograms and relative histograms of distributions): *For a distribution $p : [n] \to [0,1]$, the corresponding* histogram*, denoted $h_p : [0,1] \to \mathbb{N}$, counts the number of elements that occur with each possible probability; that is, $h_p(x) \stackrel{\rm def}{=} |\{i \in [n] : p(i) = x\}|$ for each $x \in [0,1]$. The corresponding* relative histogram*, denoted $h_p^R : [0,1] \to \mathbb{R}$, satisfies $h_p^R(x) = h_p(x) \cdot x$ for every $x \in [0,1]$.*

That is, $h_p(x)$ equals the *number* of elements in p that are assigned probability mass x, whereas $h_p^R(x)$ equals the *total probability mass* assigned to these elements. Hence, $h_p(0)$ may be positive, whereas $h_p^R(0)$ is always zero. Indeed, both functions are label-invariant (i.e., for every $p : [n] \to [0,1]$ and every bijection $\pi : D \to D$, it holds that $h_p \equiv h_{p \circ \pi}$ and $h_p^R \equiv h_{p \circ \pi}^R$).

For a non-negative function h, let $S(h) \stackrel{\rm def}{=} \{x : h(x) > 0\}$ denote the support of h. Observe that for any distribution $p : [n] \to [0,1]$ we have that $\sum_{x \in S(h_p)} h_p(x) = n$ and $\sum_{x \in S(h_p^R)} h_p^R(x) = 1$. Also note that $S(h_p^R) = S(h_p) \setminus \{0\}$.

The following definitions interpret the distance between non-negative functions h and h' as the cost of transforming h into h' by moving $m(x,y)$ (fractional) units from x in h to y in h' (for every $x \in S(h)$ and $y \in S(h')$), where the cost of moving a single unit from x to y is either $|x - y|$ or $|\log(x/y)|$ (depending on the distance).[3] We stress that $m(x,y) \geq 0$ need not be an integer, although in the case of EMD it will be shown to be an integer (w.l.o.g.).

[3] Here and in the sequel, the logarithm is to base 2. The proof of Theorem 3.1 as presented in Section 3 remains valid for any base $b \in (1, e]$; our only reference to this base is that it (i.e., b) should satisfy $\log_b z > 1 - (1/z)$ for every $z > 1$. It seems that Valiant and Valiant do mean to take $b = 2$ (although other parts of their text suggest $b = e$). Indeed, both $b = 2$ and $b = e$ seems natural choices.

Definition 2.2 (Earth-Mover Distance and Relative Earth-Mover Distance): *For a pair of non-negative functions h and h' over $[0,1]$ such that $\sum_{x\in S(h)} h(x) = \sum_{x\in S(h')} h'(x)$, the* earth-mover distance *between them, denoted* $\mathtt{EMD}(h,h')$*, is the minimum of*

$$\sum_{x\in S(h)} \sum_{y\in S(h')} m(x,y)\cdot |x-y|\,, \tag{1}$$

taken over all non-negative functions $m\colon S(h)\times S(h')\to\mathbb{R}$ that satisfy:

- *For every $x\in S(h)$, it holds that $\sum_{y\in S(h)} m(x,y) = h(x)$, and*
- *For every $y\in S(h')$, it holds that $\sum_{x\in S(h')} m(x,y) = h'(y)$.*

The relative earth-mover distance *between h and h', denoted* $\mathtt{REMD}(h,h')$*, is the minimum of*

$$\sum_{x\in S(h)} \sum_{y\in S(h')} m(x,y)\cdot |\log(x/y)|\,, \tag{2}$$

subject to the same constraints on m as for $\mathtt{EMD}$*.*

Note that satisfying the constraints may require having $m(x,x)>0$ for some x's, but the contribution of the corresponding terms to Eq. (1) (resp., Eq. (2)) is zero (since $|x-x|=0$ (resp., $\log(x/x)=0$)).

The term *earth-mover* comes from viewing the functions as piles of earth, where for each $x\in S(h)$ there is a pile of size $h(x)$ in location x and similarly for each $y\in S(h')$ there is a pile of size $h'(y)$ in location y. The goal is to transform the piles defined by h so as to obtain the piles defined by h', with minimum "transportation cost". Specifically, $m(x,y)$ captures the (possibly fractional) number of units transferred from pile x in h to pile y in h'. For $\mathtt{EMD}$ the const of transporting a unit from x to y is $|x-y|$, while for $\mathtt{REMD}$ it is $|\log(x/y)|$.

In what follows, for a pair of distributions p and q over $[n]$ we shall apply $\mathtt{EMD}$ to the corresponding pair of histograms h_p and h_q, and apply $\mathtt{REMD}$ to the corresponding relative histograms h_p^R and h_q^R. For example, Lemma 3.2 asserts that $\mathtt{EMD}(h_p,h_q)\le 2\cdot\mathtt{REMD}(h_p^R,h_q^R)$, for every two distributions p and q.

Variation distance up to relabeling. As stated in the introduction, variation distance up to relabeling (defined next) is a natural notion in the context of testing properties of distributions (see [1, Sec. 11.1.3]): It arises naturally when the properties are label-invariant (i.e., properties that are invariant under relabeling of the elements of the distribution).

Definition 2.3 (variation distance up to relabeling): *For two distributions p and q over $[n]$, the* variation distance up to relabeling *between p and q, denoted* $\mathtt{VDR}(p,q)$*, is the minimum over all permutations $\pi:[n]\to[n]$ of*

$$\frac{1}{2}\cdot\sum_{i=1}^{n} |p(i)-q(\pi(i))|\,. \tag{3}$$

In other words, $\mathtt{VDR}(p,q)$ is the variation distance between the distribution p and the set of distributions obtained by a relabeling of q (i.e., the set $\{q\circ\pi : \pi\in\mathrm{Sym}_n\}$). Observing that the latter set equals the set of distributions having histogram h_p, it follows that $\mathtt{VDR}(p,q)$ equals $\min_{q':h_{q'}=h_q}\{0.5\cdot\sum_{i=1}^{n}|p(i)-q'(i)|\}$.

3 The Claim and its Proof

Our goal is to present a proof of the following result.

Theorem 3.1 (special case (i.e., $\tau = 0$) of Fact 1 in [4, 5]): *For every two distributions p and q over $[n]$, it holds that*

$$\texttt{VDR}(p,q) \leq \texttt{REMD}(h_p^R, h_q^R)\,.$$

The proof will consist of two steps (captured by lemmas):

1. $\texttt{VDR}(p,q) = \frac{1}{2} \cdot \texttt{EMD}(h_p, h_q)$.
2. $\texttt{EMD}(h_p, h_q) \leq 2 \cdot \texttt{REMD}(h_p^R, h_q^R)$.

Actually, we start with the second step.

Lemma 3.2 (upper-bounding EMD in terms of REMD): *For every two distributions p and q over $[n]$,*

$$\texttt{EMD}(h_p, h_q) \leq 2 \cdot \texttt{REMD}(h_p^R, h_q^R)\ .$$

The following proof shows how to construct, for every transportation function m' used for the relative histograms (h_p^R and h_q^R) a corresponding transportation function m for the corresponding histograms (h_p and h_q) such that the EMD-cost of m is at most twice the REMD-cost of m'.

Proof: It will be convenient to consider two distributions, $\widetilde{p}$ and $\widetilde{q}$ that are slight variations of p and q, respectively. They are both defined over $[2n]$, where $\widetilde{p}(i) = p(i)$ and $\widetilde{q}(i) = q(i)$ for every $i \in [n]$, and $\widetilde{p}(i) = \widetilde{q}(i) = 0$ for every $i \in [2n] \setminus [n]$. Since $h_{\widetilde{p}}^R = h_p^R$ and $h_{\widetilde{q}}^R = h_q^R$, we have that $\texttt{REMD}(h_{\widetilde{p}}^R, h_{\widetilde{q}}^R) = \texttt{REMD}(h_p^R, h_q^R)$. As for $h_{\widetilde{p}}$ and $h_{\widetilde{q}}$, they agree with h_p and h_q, respectively, everywhere except on 0, where $h_{\widetilde{p}}(0) = h_p(0) + n$ and $h_{\widetilde{q}}(0) = h_q(0) + n$, so $\texttt{EMD}(h_{\widetilde{p}}, h_{\widetilde{q}}) = \texttt{EMD}(h_p, h_q)$ as well. Therefore, it suffices to show that $\texttt{EMD}(h_{\widetilde{p}}, h_{\widetilde{q}}) \leq 2 \cdot \texttt{REMD}(h_{\widetilde{p}}^R, h_{\widetilde{q}}^R)$.

Let m' be a function over $S(h_{\widetilde{p}}^R) \times S(h_{\widetilde{q}}^R)$ that satisfies the constraints stated in Definition 2.2 for the pair of relative histograms $h_{\widetilde{p}}^R$ and $h_{\widetilde{q}}^R$. We next show that there exists a non-negative function m over $S(h_{\widetilde{p}}) \times S(h_{\widetilde{q}})$ that satisfies the constraints stated in Definition 2.2 for the pair of histograms $h_{\widetilde{p}}$ and $h_{\widetilde{q}}$, and also satisfies

$$\sum_{x \in S(h_{\widetilde{p}})} \sum_{y \in S(h_{\widetilde{q}})} m(x,y) \cdot |x-y| \ \leq\ 2 \cdot \sum_{x \in S(h_{\widetilde{p}}^R)} \sum_{y \in S(h_{\widetilde{q}}^R)} m'(x,y) \cdot |\log(x/y)|\,. \tag{4}$$

Note that the range of m' is $[0,1]$, since it is defined over relative histograms, while m is not upper bounded by 1. However, the constraints that Definition 2.2 imposes on the two functions are related, since for every $x \in S(h_{\widetilde{p}}^R) = S(h_{\widetilde{p}}) \setminus \{0\}$ it is required that $\sum_{y \in S(h_{\widetilde{q}}^R)} m'(x,y)/x = h_{\widetilde{p}}(x) = h_{\widetilde{p}}(x) = \sum_{y \in S(h_{\widetilde{q}})} m(x,y)$ and for every $y \in S(h_{\widetilde{q}}^R) = S(h_{\widetilde{q}}) \setminus \{0\}$ it is required that $\sum_{x \in S(h_{\widetilde{p}}^R)} m'(x,y)/y = h_{\widetilde{q}}(y) = \sum_{x \in S(h_{\widetilde{q}})} m(x,y)$. (Indeed, m is also subjected to constraints on $x = 0$ and $y = 0$, whereas m' is not.)

The construction of the transportation function m. Essentially, for every $(x,y) \in S(h_{\widetilde{p}}^R) \times S(h_{\widetilde{q}}^R)$ we set $m(x,y) = m'(x,y)/x$ if $x > y$ and $m(x,y) = m'(x,y)/y$ otherwise, and place the excess (arising from the difference between $m'(x,y)/x$ and $m'(x,y)/y$) in either $m(0,y)$ or $m(x,0)$ according to its sign. This is done by scanning all pairs $(x,y) \in S(h^R_{\widetilde{p}}) \times S(h^R_{\widetilde{q}})$, in arbitrary order, setting $m(x,y)$, and increasing $m(0,y)$ or $m(x,0)$ appropriately. Specifically, for each $x \in S(h^R_{\widetilde{p}})$, we initialize $m(x,0)$ to 0, and similarly we initialize $m(0,y)$ to 0 for each $y \in S(h_{\widetilde{q}})$. For every pair $(x,y) \in S(h^R_{\widetilde{p}}) \times S(h^R_{\widetilde{q}})$, if $m'(x,y) = 0$, then we set $m(x,y) = 0$, and otherwise (i.e., $m'(x,y) > 0$) we proceed as follows:

- If $x > y$, then we set $m(x,y)$ to $m'(x,y)/x$ and increase $m(0,y)$ by $m^x(0,y) \stackrel{\rm def}{=} \frac{m'(x,y)}{y} - \frac{m'(x,y)}{x} > 0$ units.

  Observe that $m(x,y) \cdot (x-y) = m'(x,y) \cdot (x-y)/x = m^x(0,y) \cdot y$. Therefore, the contribution of $m(x,y) \cdot |x-y| + m^x(0,y) \cdot |0-y|$ to the left-hand-side of Eq. (4) is

  $$m(x,y) \cdot (x-y) + m^x(0,y) \cdot y = 2 \cdot m'(x,y) \cdot \frac{x-y}{x} < 2 \cdot m'(x,y) \cdot \log(x/y)\,,$$

  where the last inequality uses $1-(y/x) < \ln(x/y)$, which is due to the fact that $f(z) = \log z + (1/z) - 1 \geq \ln z + (1/z) - 1$ is positive for all $z > 1$.[4]

- If $x < y$, then we set $m(x,y)$ to $m'(x,y)/y$ and increase $m(x,0)$ by $m^y(x,0) \stackrel{\rm def}{=} \frac{m'(x,y)}{x} - \frac{m'(x,y)}{y} > 0$ units.

  Similarly to the previous case, we have $m(x,y) \cdot (y-x) = m'(x,y) \cdot (y-x)/y = m^y(x,0) \cdot x$, and the contribution to the left-hand-side of Eq. (4) is

  $$m(x,y) \cdot (y-x) + m^y(x,0) \cdot x = 2 \cdot m'(x,y) \cdot \frac{y-x}{y} < 2 \cdot m'(x,y) \cdot \log(y/x)\ .$$

- If $x = y$, then we set $m(x,y) = m'(x,y)/x$ (which equals $= m'(x,y)/y$).

  In this case $m(x,y) \cdot |x-y| = 0$, so there is no contribution to the left-hand-side of Eq. (4). Note that in this case $m'(x,y) \cdot |\log(x/y)| = 0$.

Finally, we set $m(0,0) = h_{\widetilde{p}}(0) - \sum_{y \in S(h^R_{\widetilde{q}})} m(0,y)$. To see that $m(0,0) \geq 0$, observe that $h_{\widetilde{p}}(0) \geq n$ (since $\widetilde{p}(i) = 0$ for every $i \in [2n] \setminus [n]$), whereas

$$\begin{aligned}
\sum_{y \in S(h^R_{\widetilde{q}})} m(0,y) &= \sum_{y \in S(h^R_{\widetilde{q}})} \ \sum_{x \in S(h^R_{\widetilde{p}}) \cap (y,1]} m^x(0,y) \\
&\leq \sum_{y \in S(h^R_{\widetilde{q}})} \ \sum_{x \in S(h^R_{\widetilde{p}}) \cap (y,1]} \frac{m'(x,y)}{y} \\
&\leq \sum_{y \in S(h^R_{\widetilde{q}})} h_{\widetilde{q}}(y) \\
&\leq n
\end{aligned}$$

[4] As stated in Footnote 3, we assume that the logarithm is to base $b \in (1,e]$. Indeed, here we use $\log_b z + (1/z) > 1$ for all $z > 1$, and this is the only place in the proof in which the choice of b matters.

(since $\widetilde{q}(i) = 0$ for every $i \in [2n] \setminus [n]$). Lastly, by combining the contribution of all pairs $(x, y) \in S(h_{\widetilde{p}}) \times S(h_{\widetilde{q}})$ to the left-hand-side of Eq. (4), we upper-bound the total contribution by

$$\begin{aligned}
&\sum_{(x,y)\in S(h^R_{\widetilde{p}})\times S(h^R_{\widetilde{q}}):x>y} 2\cdot m'(x,y)\cdot \log(x/y) + \sum_{(x,y)\in S(h^R_{\widetilde{p}})\times S(h^R_{\widetilde{q}}):x<y} 2\cdot m'(x,y)\cdot \log(y/x) \\
&\qquad = \sum_{(x,y)\in S(h^R_{\widetilde{p}})\times S(h^R_{\widetilde{q}})} 2\cdot m'(x,y)\cdot |\log(x/y)|,
\end{aligned}$$

which equals the right-hand-side of Eq. (4).

Verifying that the transportation function m is legal. It remains to verify that m satisfies the constraints in Definition 2.2. For each $x \in S(h_{\widetilde{p}}) \setminus \{0\}$,

$$\begin{aligned}
\sum_{y\in S(h_{\widetilde{q}})} m(x,y) &= m(x,0) + \sum_{y\in S(h_{\widetilde{q}})\cap(0,x]} m(x,y) + \sum_{y\in S(h_{\widetilde{q}})\cap(x,1]} m(x,y) \\
&= \sum_{y\in S(h^R_{\widetilde{q}})\cap(x,1]} m^y(x,0) + \sum_{y\in S(h^R_{\widetilde{q}})\cap(0,x]} m(x,y) + \sum_{y\in S(h^R_{\widetilde{q}})\cap(x,1]} m(x,y) \\
&= \sum_{y\in S(h^R_{\widetilde{q}})\cap(x,1]} \left(\frac{1}{x} - \frac{1}{y}\right)\cdot m'(x,y) + \sum_{y\in S(h^R_{\widetilde{q}})\cap(0,x]} \frac{m'(x,y)}{x} + \sum_{y\in S(h^R_{\widetilde{q}})\cap(x,1]} \frac{m'(x,y)}{y} \\
&= \sum_{y\in S(h^R_{\widetilde{q}})} \frac{m'(x,y)}{x} \\
&= h_{\widetilde{p}}(x)\,.
\end{aligned}$$

Similarly, for each $y \in S(h_{\widetilde{q}}) \setminus \{0\}$,

$$\begin{aligned}
\sum_{x\in S(h_{\widetilde{p}})} m(x,y) &= m(0,y) + \sum_{x\in S(h_{\widetilde{p}})\cap(0,y]} m(x,y) + \sum_{x\in S(h_{\widetilde{p}})\cap(y,1]} m(x,y) \\
&= \sum_{x\in S(h^R_{\widetilde{q}})\cap(y,1]} m^x(0,y) + \sum_{x\in S(h^R_{\widetilde{q}})\cap(0,y]} m(x,y) + \sum_{x\in S(h^R_{\widetilde{q}})\cap(y,1]} m(x,y) \\
&= \sum_{x\in S(h^R_{\widetilde{q}})\cap(y,1]} \left(\frac{1}{y} - \frac{1}{x}\right)\cdot m'(x,y) + \sum_{x\in S(h^R_{\widetilde{q}})\cap(0,y]} \frac{m'(x,y)}{y} + \sum_{x\in S(h^R_{\widetilde{q}})\cap(y,1]} \frac{m'(x,y)}{x} \\
&= \sum_{x\in S(h^R_{\widetilde{p}})} \frac{m'(x,y)}{y} \\
&= h_{\widetilde{q}}(y)\,.
\end{aligned}$$

Recall that we defined $m(0,0)$ such that $m(0,0) + \sum_{y\in S(h^R_{\widetilde{q}})} m(0,y) = h_{\widetilde{p}}(0)$, and it follows that $\sum_{y\in S(h_{\widetilde{q}})} m(0,y) = h_{\widetilde{p}}(0)$. Lastly, we observe that

$$\sum_{x\in S(h_{\widetilde{p}})} m(x,0) \;=\; \sum_{x\in S(h_{\widetilde{p}})} \sum_{y\in S(h_{\widetilde{q}})} m(x,y) - \sum_{x\in S(h_{\widetilde{p}})} \sum_{y\in S(h_{\widetilde{q}})\setminus\{0\}} m(x,y)$$

$$
\begin{aligned}
&= \sum_{x\in S(h_{\widetilde{p}})} h_{\widetilde{p}}(x) - \sum_{y\in S(h_{\widetilde{p}})\setminus\{0\}} \; \sum_{y\in S(h_{\widetilde{q}})} m(x,y) \\
&= 2n - \sum_{y\in S(h_{\widetilde{q}})\setminus\{0\}} h_{\widetilde{q}}(y) \\
&= h_{\widetilde{q}}(0)\,.
\end{aligned}
$$

This completes the verification of all contraints, and the lemma follows. ■

Lemma 3.3 ($\mathtt{VDR}$ equals $\mathtt{EMD}/2$): *For every two distributions p and q over $[n]$,*

$$\mathtt{VDR}(p,q) = \frac{1}{2}\cdot \mathtt{EMD}(h_p,h_q)\ .$$

Intuitively, there is a one-to-one correspondence between the cost of the best *integer-valued* transportation functions m satisfying the conditions of Definition 2.2 (w.r.t h_p and h_q) and the differences (between p and $q\circ\pi$) under the best relabeling permutation π used in Definition 2.3. The core of the following proof is showing that integer-value transportation functions m obtain the minimum for $\mathtt{EMD}$.

Proof: Consider a constrained version of the earth-mover distance in which we also require that $m(x,y)$ is an *integer* for every $x\in S(h_p)$ and $y\in S(h_q)$, and denote this distance measure by $\mathtt{IEMD}$. Using the definition of $\mathtt{VDR}$ and $\mathtt{IEMD}$, one can verify that $\mathtt{VDR}(q,p) = \frac{1}{2}\cdot \mathtt{IEMD}(h_p,h_q)$. To see this consider the best permutation π used in Definition 2.3, and let $q' = q\circ\pi$. In this case, the ℓ_1-norm of $p-q'$ equals $\sum_{i\in[n]}|p(i)-q'(i)|$, which in turn equals $\sum_{x,y\in[0,1]} m(x,y)\cdot|x-y|$, where $m(x,y) = |\{i\in[n]: p(i)=x \,\&\, q'(i)=y\}|$ is the desired transportation function. Likewise, an integer-valued trasportation function m (from h_p to $h_{q'}$) yields an upper bound on the ℓ_1-norm of $p-q'$. Last, noting that $h_{q'} = h_q$ and that the variation distance between distributions equals half the ℓ_1-norm between them, the claim follows.

It remains to prove that $\mathtt{EMD}(h_p,h_q) = \mathtt{IEMD}(h_p,h_q)$; that is, that a function m that obtains the minimum of the $\mathtt{EMD}$ objective function has integer values. To this end, we define a specific integer-valued function m (based on a simple iterative assignment procedure), and show that it is optimal.

Constructing the transportation function m. We use an iterative construction. Initially, $m(x,y)=0$ for every $x\in S(h_p)$ and $y\in S(h_q)$. We also initialize $s(x) = h_p(x)$ for every $x\in S(h_p)$, and $d(y) = h_q(y)$ for every $y\in S(h_q)$. Intuitively, $s(x)$ is the supply of h_p at x, and $d(y)$ is the demand of h_q at y. Note that $\sum_{x\in S(h_p)} s(x) = n = \sum_{y\in S(h_q)} d(y)$, and that the supplies and demands are initially integers (and will remain so throughout the iterations).

In each iteration, we consider the smallest $x\in S(h_p)$ for which $s(x)>0$ and the smallest $y\in S(h_q)$ for which $d(y)>0$, set $m(x,y) = \min\{s(x),d(y)\}$ and reduce both $s(x)$ and $d(y)$ by $m(x,y)$. Since the demands are initially integers, the values of the $m(x,y)$'s that we set are integers, and the supplies and demands remain (non-negative) integers. (Equivalently, we may transport one unit in each iteration such that in the i^{th} iteration we transport the i^{th} smallest p-value to the i^{th} smallest q-value.)[5] By its construction, the function m satisfies the constraints of Definition 2.2.

[5] That is, letting π_p and π_q be permutations over $[n]$ such that $p(\pi_p(j)) \le p(\pi_p(j+1))$ and $q(\pi_q(j)) \le q(\pi_q(j+1))$ for every $j\in[n-1]$, in the i^{th} iteration we transport one unit from location $p(\pi_p(i))$ of h_p to location $q(\pi_q(i))$ of h_q.

Showing that m is optimal for EMD. Recall that the cost of a transportation function is the value assigned to it by Eq. (1), and our aim is to show that the constructed function m has minimal cost among all transportation functions that satisfy the constraints of Definition 2.2. Note that these constraints form a linear system, and so the notions of minimiality referred to here and below are well-defined. Assuming that m is not optimal, consider a non-negative function ℓ over $S(h_p) \times S(h_q)$ that satisfies the constraints of Definition 2.2 and is minimal in the following sense.

1. The function ℓ has the smallest cost among all non-negative functions that satisfy the constraints of Definition 2.2. (Indeed, this means that $\ell \neq m$.)

2. Among the functions satisfying the foregoing, the function ℓ agrees with m on the longest prefix of pairs (x, y) according to the lexicographical order on pairs.

   Let (x^*, y^*) be the first pair on which ℓ and m differ; that is, $\ell(x^*, y^*) \neq m(x^*, y^*)$ whereas $\ell(x, y) = m(x, y)$ for every $(x, y) < (x^*, y^*)$.

3. Among all functions satisfying the foregoing, the function ℓ attains a minimal value for $|\ell(x^*, y^*) - m(x^*, y^*)|$.

We reach a contradiction to the hypothesis regarding ℓ by first proving that $\ell(x^*, y^*) < m(x^*, y^*)$, and then presenting a function ℓ' that violates the minimiality of ℓ (as defined above (w.r.t all functions that satisfy the constraints of Definition 2.2)).

Proving that $\ell(x^*, y^*) < m(x^*, y^*)$. Towards this end, we consider the supply of x^* and the demand of y^* just before $m(x^*, y^*)$ is reset; that is, $s(x^*) = h_p(x^*) - \sum_{y<y^*} m(x^*, y)$ and $d(y^*) = h_q(y^*) - \sum_{x<x^*} m(x, y^*)$, where we rely on the fact that (by construction of m) before $m(x^*, y^*)$ is reset it holds that $m(x^*, y) > 0$ only if $y < y^*$ and $m(x, y^*) > 0$ only if $x < x^*$. Assuming towards the contradiction that $\ell(x^*, y^*) > m(x^*, y^*)$, and recalling that $m(x^*, y^*) = \min(s(x^*), d(y^*))$, we note that if $m(x^*, y^*) = s(x^*)$, then

$$\sum_{y \leq y^*} \ell(x^*, y) = \sum_{y<y^*} m(x^*, y) + \ell(x^*, y^*) > \sum_{y<y^*} m(x^*, y) + m(x^*, y^*) = h_p(x^*)\,,$$

which means that ℓ violates a constraint of Definition 2.2. A similar contradiction is obtained by assuming that $\ell(x^*, y^*) > m(x^*, y^*) = d(y^*)$, when in this case we get $\sum_{x \leq x^*} \ell(x, y^*) > h_q(x^*)$.

Demonstrating that ℓ is not minimal. Having shown that $\ell(x^*, y^*) < m(x^*, y^*)$, we now derive a function ℓ' that violates the minimality of ℓ. First, using the hypothesis that ℓ satisfies the constraints of Definition 2.2, we observe that $\ell(x^*, y^*) < m(x^*, y^*)$ (combined with $\ell(x, y) = m(x, y)$ for every $(x, y) < (x^*, y^*)$) implies that there exists $x' > x^*$ such that $\ell(x', y^*) > m(x', y^*)$ and $y' > y^*$ such that $\ell(x^*, y') > m(x^*, y')$. Letting $c = \min(m(x^*, y^*) - \ell(x^*, y^*), \ell(x', y^*), \ell(x^*, y')) > 0$, define ℓ' as equal to ℓ on all pairs except for the following four pairs that satisfy $\ell'(x^*, y^*) = \ell(x^*, y^*) + c$, $\ell'(x', y^*) = \ell(x', y^*) - c$, $\ell'(x^*, y') = \ell(x^*, y') - c$, and $\ell'(x', y') = \ell(x', y') + c$. Then, ℓ' preserves the constraints of Definition 2.2, but[6] (1) its cost is not higher than that of ℓ,

[6] To see that (1) holds, note that the cost of ℓ' equals the cost of ℓ plus $c \cdot |x^* - y^*| - c \cdot |x' - y^*| - c \cdot |x^* - y'| + c \cdot |x' - y'|$. Hence, we need to verify that the added value is not positive; equivalently, that $|x^* - y^*| - |x^* - y'| \leq |x' - y^*| - |x' - y'|$. Consider the following cases:

1. The diagonal line $y = x$ does not cross the rectangle spanned by (x^*, y^*) (i.e., either $y' \leq x^*$ or $y^* \geq x'$). If $y' \leq x^*$, then $|y^* - x^*| - |y' - x^*| = y' - y^* = |y^* - x'| - |y' - x'|$, and otherwise $|y^* - x^*| - |y' - x^*| = -(y' - y^*) = |y^* - x'| - |y' - x'|$.

(2) its first point of disagreement with m is not before (x^*, y^*), and (3) $|\ell'(x^*, y^*) - m(x^*, y^*)| = |\ell(x^*, y^*) - m(x^*, y^*)| - c$. This contradicts the minimality of ℓ, since $c > 0$.

Having proved that no transportation function that satisfies the constraints of Definition 2.2 has cost that is smaller than the cost of m, the lemma follows. ■

Comments. As noted in [3], there exist distributions p and q for which $\texttt{VDR}(h_p, h_q) \ll \texttt{REMD}(h_p^R, h_q^R)$. The source of this phenomenon is the unbounded cost of transportation under the REMD (i.e., transforming a unit of mass from x to y costs $|\log(x/y)|$). For example, for any $\epsilon \in (0, 0.5)$, consider the pair (p, q) such that p is uniform over $[n]$ (i.e., $p(i) = 1/n$ for every $i \in [n]$) and q is extremely concentrated on a single point in the sense that $q(n) = 1 - \epsilon$ and $q(i) = \epsilon/(n-1)$ for every $i \in [n-1]$. Then, the variation distance between p and q is $\frac{n-1}{n} - \epsilon$, but the REMD is greater than $\frac{n-1}{n} \cdot \log(1/\epsilon)$. This phenomenon is reflected in the proof of Lemma 3.2 at the point we used the inequality $1 - (1/z) < \log z$ for $z > 1$. This inequality becomes more crude when z grows.

References

[1] O. Goldreich. *Introduction to Property Testing.* Cambridge University Press, 2017.

[2] O. Goldreich and D. Ron. On Sample-Based Testers. In *6th Innovations in Theoretical Computer Science*, pages 337–345, 2015.

[3] G. Valiant and P. Valiant. Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *43rd ACM Symposium on the Theory of Computing*, pages 685–694, 2011. See *ECCC* TR10-180 for the algorithm, and TR10-179 for the lower bound.

[4] G. Valiant and P. Valiant. Instance optimal learning. *CoRR*, abs/1504.05321, 2015.

[5] G. Valiant and P. Valiant. Instance optimal learning of discrete distributions. In *48th ACM Symposium on the Theory of Computing*, pages 142–155, 2016.

2. The diagonal line $y = x$ separates one corner-point of the rectangle from the other three corner-points (e.g., $y' > x^*$ but $y < x$ for $(x, y) \in \{(x^*, y^*), (x', y^*), (x', y')\}$). If $y' > x^*$, then $|y^* - x^*| - |y' - x^*| < y' - y^* = |y^* - x'| - |y' - x'|$, and similarly for the case that (x', y^*) is separated.

3. The diagonal line $y = x$ crosses both horizontal lines of the rectangle (i.e., $y^*, y' \in [x^*, x']$). In this case, $|y^* - x^*| - |y' - x^*| = -(y' - y^*)$ and $|y^* - x'| - |y' - x'| = y' - y^*$.

4. The diagonal line $y = x$ crosses both vertical lines of the rectangle (i.e., $x^*, x' \in [y^*, y']$). In this case $|y^* - x^*| - |y' - x^*| < |y^* - x'| - |y' - x'|$, since $|y^* - x^*| < |y^* - x'|$ and $|y' - x^*| > |y' - x'|$.

To see that (2) holds, recall that $\ell'(x, y) = \ell(x, y) = m(x, y)$ for every $(x, y) < (x^*, y^*)$.