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

Oded Goldreich† Dana Ron‡

February 9, 2016

Summary. In this note we present a proof that the variation distance up to relabeling is upper-bounded by the "relative earth mover distance" (to be defined below). The relative earth mover distance was introduced by Valiant and Valiant [VV11], and was extensively used in their work. The foregoing claim was made in [VV11], but was not used there. The claim appears a special case of [VV15, Fact 1] (i.e., the case of $\tau = 0$). The proof we present is merely an elaboration of (this special case of) the proof presented by Valiant and Valiant in [VV15, Apdx A].

1 Definitions

We start by introducing some definitions and notations.

Definition 1 (Histograms and relative histograms for distributions) *For a distribution* $p : [n] \to [0,1]$*, the corresponding* histogram*, denoted* $h_p\colon [0,1] \to \mathbb{N}$*, such that* $h_p(x) \stackrel{\rm def}{=} |\{i \in [n] : p(i) = x\}|$ *for each* $x \in [0,1]$*. The corresponding* relative histogram*, denoted* $h_p^R\colon [0,1] \to \mathbb{R}$*, 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.

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 definition interprets the distance between non-negative functions h and h' as the cost of transforming h into h' by moving $m(x,y)$ 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).

---

*Partially supported by the Israel Science Foundation (grant No. 671/13).

†Department of Computer Science, Weizmann Institute of Science, Rehovot, Israel. E-mail: oded.goldreich@weizmann.ac.il

‡Department of EE–Systems, Tel-Aviv University, Ramat-Aviv, Israel. E-mail: danar@eng.tau.ac.il

Definition 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* $\mathrm{EMD}(h,h')$*, is the minimum of*

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

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* $\mathrm{REMD}(h,h')$*, is the minimum of*

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

subject to the same constraints on m as for EMD*.*

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 EMD the transportation cost of a unit from x to y is $|x-y|$ while for REMD it is $|\log(x/y)|$. In what follows, for a pair of distributions p and q over $[n]$ we shall apply EMD to the corresponding pair of histograms h_p and h_q, and apply REMD to the corresponding relative histograms h_p^R and h_q^R.

Variation distance up to relabeling, as defined next, is a natural notion in the context of testing properties of symmetric distributions (i.e., properties that are invariant under relabeling of the elements of the distribution).

Definition 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* $\mathrm{VDR}(p,q)$*, is the minimum over all permutations σ over $[n]$ of*

$$\frac{1}{2} \sum_{i=1}^{n} |p(i) - q(\sigma(i))| \ .$$

2 Proofs

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

Theorem 4 (special case of Fact 1 in [VV15]) *For every two distributions p and q over $[n]$, it holds that*

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

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

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

Actually, we start with the second step.

Lemma 5 *For every two distributions p and q over $[n]$,*

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

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 $\mathrm{REMD}(h_{\widetilde{p}}^R, h_{\widetilde{q}}^R) = \mathrm{REMD}(h_p^R, h_q^R)$. 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 $\mathrm{EMD}(h_{\widetilde{p}}, h_{\widetilde{q}}) = \mathrm{EMD}(h_p, h_q)$ as well. Therefore, it suffices to show that $\mathrm{EMD}(h_{\widetilde{p}}, h_{\widetilde{q}}) \le 2 \cdot \mathrm{REMD}(h_{\widetilde{p}}^R, h_{\widetilde{q}}^R)$.

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 for the pair of 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 for the pair of histograms $h_{\widetilde{p}}$ and $h_{\widetilde{q}}$, and also satisfies

$$\sum_{x \in S(h_{\widetilde{p}})} \sum_{y \in S(h_{\widetilde{q}})} m(x,y) \cdot |x-y| \le 2 \cdot \sum_{x \in S(h_{\widetilde{p}}^R)} \sum_{y \in S(h_{\widetilde{q}}^R)} m'(x,y) \cdot |\log(x/y)| . \tag{1}$$

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 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 $\sum_{y \in S(h_{\widetilde{q}}^R)} m'(x,y)/x = h_{\widetilde{p}}(x) = \sum_{y \in S(h_{\widetilde{q}})} m(x,y)$ 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.)

We now define the function m. For each $x \in S(h_{\widetilde{p}}^R)$, initialize $m(x,0)$ to 0 and similarly for each $y \in S(h_{\widetilde{q}})$, initialize $m(0,y)$ to 0. For every pair $(x,y) \in S(h_{\widetilde{p}}^R) \times S(h_{\widetilde{q}}^R)$, if $m'(x,y) = 0$, then $m(x,y) = 0$, and otherwise we do the following.

- If $x > y$, let $m(x,y)$ be set to $m'(x,y)/x$ and increase $m(0,y)$ by $m^x(0,y) \stackrel{\text{def}}{=} m'(x,y)/y - m'(x,y)/x > 0$. Observe that $m(x,y) \cdot (x-y) = m'(x,y) \cdot (1 - y/x) = m^x(0,y) \cdot y$. Therefore, the contribution to the left-hand-side of Equation (1) is

  $$m(x,y) \cdot (x-y) + m^x(0,y) \cdot (y-0) = 2m'(x,y) \cdot (1-y/x) < 2m'(x,y) \cdot \log(x/y) ,$$

  where the last inequality is due to the fact that $f(z) = \log z + (1/z) - 1 > \ln z + (1/z) - 1$ is positive for all $z > 1$.

- If $x < y$, let $m(x,y)$ be set to $m'(x,y)/y$ and increase $m(x,0)$ by $m^y(x,0) \stackrel{\text{def}}{=} m'(x,y)/x - m'(x,y)/y > 0$. Similarly to the previous case, $m(x,y) \cdot (y-x) = m^y(x,0) \cdot x$, and the contribution to the left-hand-side of Equation (1) is

$$m(x,y) \cdot (y-x) + m^y(x,0) \cdot (x-0) = 2m'(x,y) \cdot (1 - x/y) < 2m'(x,y) \cdot \log(y/x) \ .$$

- If $x = y$, let $m(x,y) = m'(x,y)/x$ $(= m'(x,y)/y)$. In this case both $m(x,y) \cdot |x-y| = 0$ and $m'(x,y) \cdot |\log(x/y)| = 0$.

Finally, we set $m(0,0) = h_{\widetilde{p}}(0) - \sum_{y \in S(h^R_{\widetilde{q}})} m(0,y)$. To see that $m(0,0) \geq 0$, note that since $h_{\widetilde{p}}(0) \geq n$ while

$$\sum_{y \in S(h^R_{\widetilde{q}})} m(0,y) = \sum_{y \in S(h^R_{\widetilde{q}})} \sum_{x \in S(h^R_{\widetilde{p}}) \cap (y,1]} m^x(0,y) = \sum_{y \in S(h^R_{\widetilde{q}})} \sum_{x \in S(h^R_{\widetilde{p}}) \cap (y,1]} m'(x,y)/y \leq n.$$

By combining the contribution of all pairs x, y as defined above, Equation (1) holds.

It remains to verify that m satisfies the constraints in Definition 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 $y \in S(h_{\widetilde{q}}) \setminus \{0\}$,

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

We defined $m(0,0)$ such that $\sum_{y \in S(h_{\widetilde{q}})} m(0,y) = m(0,0) + \sum_{y \in S(h^R_{\widetilde{q}})} m(0,y) = h_{\widetilde{p}}(0)$, and

$$\sum_{x \in S(h_{\widetilde{p}})} m(x,0) = \sum_{x \in S(h_{\widetilde{p}})} \sum_{y \in \widetilde{q}} m(x,y) - \sum_{x \in S(h_{\widetilde{p}})} \sum_{y \in S(h_{\widetilde{q}}) \setminus \{0\}} m(x,y) = 2n - \sum_{y \in S(h_{\widetilde{q}})} h_{\widetilde{q}}(y) = h_{\widetilde{q}}(0) \ ,$$

and the proof is completed. ■

Lemma 6 *For every two distributions p and q over $[n]$,*

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

Intuitively, there is a one-to-one correspondence between integer-valued transportation functions m as in Definition 2 and the relabeling permutations σ used in Definition 3. The core of the following proof is showing that integer-value transportation functions m obtain the minimum for 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 IEMD. Using the definition of VDR and IEMD, one can verify that $\mathrm{VDR}(q,p) = \frac{1}{2} \cdot \mathrm{IEMD}(h_p, h_q)$, since there is a correspondence between the permutation σ used in Definition 3 and the integer movement in EMD. (The factor of $1/2$ is due to the fact that the variation distance between distributions equals half the L_1-norm between them.)

It therefore remains to prove that $\mathrm{EMD}(h_p, h_q) = \mathrm{IEMD}(h_p, h_q)$; that is, the function m that obtains the minimum of the 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.

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 x, and $d(y)$ is the demand of y.) Note that $\sum_{x \in S(h_p)} s(x) = n = \sum_{y \in S(h_q)} d(y)$. 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 $m(x,y)$. Hence, all intermediate values of m (as well as s and d) are integers. (We note that an equivalent definition of m can be obtained by considering the mapping σ from $[n]$ to $[n]$ that maps the i^{th} smallest p-value to the i^{th} smallest q-value.)[1] By its construction, the function m satisfies the constraints of Definition 2.

To verify that the resulting function m is an optimal setting for EMD, consider any other non-negative function ℓ over $S(h_p) \times S(h_q)$ that satisfies the constraints of Definition 2. Actually, among all such functions ℓ consider only those that agree with m on the longest prefix of pairs (x,y) according to the lexicographical order on pairs, and 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 $(x,y) < (x^*,y^*)$. Furthermore, among all such functions ℓ, select one for which $|\ell(x^*,y^*) - m(x^*,y^*)|$ is minimal. We shall show that $\ell = m$.

Assume towards the contradiction that $\ell \neq m$, and let (x^*,y^*) be as above. We first prove 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 determined; that is, $s(x^*) = h_p(x^*) - \sum_{y<y^*} m(x^*,y)$ and $d(y^*) = h_q(y^*) - \sum_{x<x^*} m(x,y^*)$. Recalling that $m(x^*,y^*) = \min(s(x^*), d(y^*))$, we note that if $\ell(x^*,y^*) > m(x^*,y^*) = s(x^*)$, then $\sum_{y \le y^*} \ell(x^*,y) = \sum_{y<y^*} m(x^*,y) + \ell(x^*,y^*) > h_p(x^*)$, which means that ℓ violates a constraint of Definition 2. A similar contradiction is obtained by assuming that $\ell(x^*,y^*) > m(x^*,y^*) = d(y^*)$, when in this case we get $\sum_{x \le x^*} \ell(x,y^*) > h_q(x^*)$.

Having shown that $\ell(x^*,y^*) < m(x^*,y^*)$. we now derive a function ℓ' that violates the "minimality" of ℓ. Specifically, $\ell(x^*,y^*) < m(x^*,y^*)$ (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

[1] That is, letting π_p and π_q be permutations over $[n]$ such that $p(\pi_p(i)) \le p(\pi_p(i+1))$ and $q(\pi_q(i)) \le q(\pi_q(i+1))$ for every $i \in [n-1]$, define $\sigma(\pi_p(i)) = \pi_q(i)$ for every $i \in [n]$.

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, but $\ell'(x,y) = \ell(x,y) = m(x,y)$ for every $(x,y) < (x^*,y^*)$ and $|\ell'(x^*,y^*) - m(x^*,y^*)| = |\ell(x^*,y^*) - m(x^*,y^*)| - c$, in contradiction to the choice of ℓ, since $c > 0$. ■

3 Comments

As noted in [VV11], there exist distributions p and q for which $\mathrm{VDR}(h_p, h_q) \ll \mathrm{REMD}(h_p^R, h_q^R)$. 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 at least $\frac{n-1}{n} \cdot \log(1/\epsilon)$.

This phenomenon is reflected in the proof of Lemma 5 at the point we used the inequality $1 - (1/z) < \log z$ for $z > 1$. This inequality becomes more crude when z grows.

References

[VV11] 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 *Proceedings of the Fourty-Third Annual ACM Symposium on the Theory of Computing (STOC)*, pages 685–694, 2011. See *ECCC* TR10-180 for the algorithm, and TR10-179 for the lower bound.

[VV15] G. Valiant and P. Valiant. Instance optimal learning. *CoRR*, abs/1504.05321, 2015.