The Perturbed Variation

Maayan Harel
Department of Electrical Engineering
Technion, Haifa, Israel
maayanga@tx.technion.ac.il

Shie Mannor
Department of Electrical Engineering
Technion, Haifa, Israel
shie@ee.technion.ac.il

Abstract

We introduce a new discrepancy score between two distributions that gives an indication on their *similarity*. The new score gives an intuitive interpretation of similarity by applying perturbations; namely, two distributions are similar if they may be matched by perturbations. The rate of perturbation is inversely proportional to the rate of similarity to be tested. The score is defined between distributions, and can be efficiently estimated from samples. We provide convergence bounds of the estimated score, and develop hypothesis testing procedures that test if two data sets come from similar distributions. The statistical power of these procedures was tested in simulations. We also compared the score's capacity to detect similarity with that of other known measures on real data [1].

1 Introduction

The question of similarity between two sets of examples is common to many fields, including statistics, data mining, machine learning and computer vision. For example, in machine learning, a standard assumption is that the training and test data are generated from the same distribution. However, in some scenarios, such as Domain Adaptation (DA), this is not the case and the distributions are only assumed similar. It is quite intuitive to denote when two inputs are similar in nature, yet the following question remains open: given two sets of examples, how do we test whether or not they were generated by similar distributions? The main focus of this work is providing a similarity score and a corresponding statistical procedure that gives one possible answer to this question.

Discrepancy between distributions has been studied for decades, and a wide variety of distance scores have been proposed. However, not all proposed scores can be used for testing similarity. The main difficulty is that most scores have not been designed for statistical testing of similarity but equality, known as the Two-Sample Problem (TSP). Formally, let P and Q be the generating distributions of the data; the TSP tests the null hypothesis $H_0 : P = Q$ against the general alternative $H_1 : P \neq Q$. This is one of the classical problems in statistics. Sometimes, however, the interesting question is with regards to similarity rather than equality. By design, most equality tests may not be transformed to test similarity (see Section 3 for a review of representative works). In this work, we construct two types of complementary procedures for hypothesis testing of similarity and dissimilarity that relax the equality condition of the TSP.

We suggest to quantify similarity using a new score, the Perturbed Variation (PV). We propose that similarity is related to some predefined value of permitted variations. Consider the gait of two male subjects as an example. If their physical characteristics are similar, we expect their walk to be similar, and thus assume the examples representing the two are from similar distributions. This intuition applies when the distribution of our measurements only endures small changes for people with similar characteristics. Put more generally, similarity depends on what "small changes" are in a given application, and implies that similarity is domain specific. The PV, as hinted by its name,

[1] The presented work is a summarized version of a paper accepted to NIPS 2012 proceedings. Further details including the experiments are presented in the full version [1].

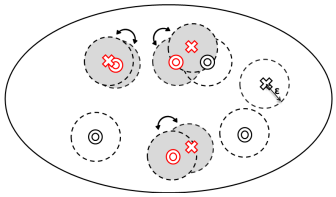

Figure 1: X and O identify samples from two distributions, doted circles denote allowed perturbations. Samples marked in red are matched with neighbors, while the unmatched samples indicate the PV discrepancy.

measures the discrepancy between two distributions while allowing for some perturbation of each distribution; that is, it allows small differences between the distributions. What accounts for small differences is a parameter of the PV, and may be defined by the user with regard to a specific domain. Figure 1 illustrates the PV. Note that, like perceptual similarity, the PV turns a blind eye to variations of some rate.

2 The Perturbed Variation

The PV on continuous distributions is defined as follows:

Definition 1. *Let P and Q be two distributions on a Banach space $\mathcal{X}$, and let $M(P,Q)$ be the set of all joint distributions on $\mathcal{X} \times \mathcal{X}$ with marginals P and Q. The PV, with respect to a distance function $\mathbf{d} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and ϵ, is defined by*

$$PV(P, Q, \epsilon, \mathbf{d}) \doteq \inf_{\mu \in M(P,Q)} \mathbb{P}_\mu[d(X, Y) > \epsilon], \tag{1}$$

over all pairs $(X, Y) \sim \mu$, such that the marginal of X is P and the marginal of Y is Q.

Put into words, Equation (1) defines the joint distribution μ that couples the two distributions such that the probability of the event of a pair $(X, Y) \sim \mu$ being within a distance grater than ϵ is minimized.

The solution to (1) is a special case of the classical mass transport problem of Monge [2] and its version by Kantorovich: $\inf_{\mu \in M(P,Q)} \int_{\mathcal{X} \times \mathcal{X}} c(x, y) d\mu(x, y)$, where $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a measurable cost function. When c is a metric, the problem describes the 1^{st} Wasserstein metric. Problem (1) may be rephrased as the optimal mass transport problem with the cost function $c(x, y) = 1_{[d(x,y)>\epsilon]}$, and may be rewritten as $\inf_\mu \iint 1_{[d(x,y)>\epsilon]} \mu(y|x) dy \, P(x) dx$. The probability $\mu(y|x)$ defines the transportation plan of x to y. The PV optimal transportation plan is obtained by perturbing the mass of each point x in its ϵ neighborhood so that it redistributes to the distribution of Q. These small perturbations do not add any cost, while transportation of mass to further areas is equally costly. Note that when $P = Q$ the PV is zero as the optimal plan is simply the identity mapping. Due to its cost function, the PV it is not a metric, as it is symmetric but does not comply with the triangle inequality and may be zero for distributions $P \neq Q$. Despite this limitation, this cost function fully quantifies the intuition that small variations should not be penalized when similarity is considered. In this sense, similarity is not unique by definition, as more than one distribution can be similar to a reference distribution.

The PV is also closely related to the Total Variation distance (TV) that may be written, using a coupling characterization, as $TV(P, Q) = \inf_{\mu \in M(P,Q)} \mathbb{P}_\mu [X \neq Y]$ [3]. This formulation argues that any transportation plan, even to a close neighbor, is costly. Due to this property, the TV is known to be an overly sensitive measure that overestimates the distance between distributions. For example, consider two distributions defined by the dirac delta functions $\delta(a)$ and $\delta(a + \epsilon)$. For any ϵ, the TV between the two distributions is 1, while they are intuitively similar. The PV resolves this problem by adding perturbations, and therefore is a natural extension of the TV. Notice, however, that the ϵ used to compute the PV need not be infinitesimal, and is defined by the user.

The PV can be seen as a conciliatory between the Wasserstein distance and the TV. As explained, it relaxes the sensitivity of the TV; however, it does not "over optimize" the transportation plan. Specif-

$PV(\mu_1, \mu_2, \epsilon) = \frac{1}{2}$:

$a_1 = 0, a_2 = 1, a_3 = 2, a_4 = 2.1$
$w_1 = w_2 = \frac{1}{4}, w_3 = w_4 = 0$
$v_4 = \frac{1}{2}, v_1 = v_2 = v_3 = 0$

$$Z = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

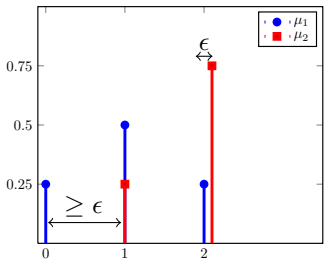

Figure 2.1: Illustration of the PV score between discrete distributions.

ically, distances larger than the allowed perturbation are discarded. This aspect also contributes to the efficiency of estimation of the PV from samples; see Section 2.2.

2.1 The Perturbed Variation on Discrete Distributions

It can be shown that for two discrete distributions Problem (1) is equivalent to the following problem.

Definition 2. *Let μ_1 and μ_2 be two discrete distributions on the unified support $\{a_1, ..., a_N\}$. Define the neighborhood of a_i as $ng(a_i, \epsilon) = \{z\,; d(z, a_i) \leq \epsilon\}$. The PV$(\mu_1, \mu_2, \epsilon, \mathbf{d})$ between the two distributions is:*

$$\min_{w_i \geq 0, v_i \geq 0, Z_{ij} \geq 0} \frac{1}{2}\sum_{i=1}^{N} w_i + \frac{1}{2}\sum_{j=1}^{N} v_j \tag{2}$$

$$\text{s.t.} \sum_{a_j \in ng(a_i,\epsilon)} Z_{ij} + w_i = \mu_1(a_i),\ \forall i$$

$$\sum_{a_i \in ng(a_j,\epsilon)} Z_{ij} + v_j = \mu_2(a_j),\ \forall j$$

$$Z_{ij} = 0\,, \quad \forall (i,j) \notin ng(a_i, \epsilon).$$

Each row in the matrix $Z \in \mathbb{R}^{N \times N}$ corresponds to a point mass in μ_1, and each column to a point mass in μ_2. For each i, $Z(i,:)$ is zero in columns corresponding to non neighboring elements, and non-zero only for columns j for which transportation between $\mu_2(a_j) \to \mu_1(a_i)$ is performed. The discrepancies between the distributions are depicted by the scalars w_i and v_i that count the "leftover" mass in $\mu_1(a_i)$ and $\mu_2(a_j)$. The objective is to minimize these discrepancies, therefore matrix Z describes the optimal transportation plan constrained to ϵ-perturbations. An example of an optimal plan is presented in Figure 2.1.

2.2 Estimation of the Perturbed Variation

Typically, we are given samples from which we would like to estimate the PV. Given two samples $S_1 = \{x_1, ..., x_n\}$ and $S_2 = \{y_1, ..., y_m\}$, generated by distributions P and Q respectively, $\widehat{\text{PV}}(S_1, S_2, \epsilon, d)$ is:

$$\min_{w_i \geq 0, v_i \geq 0, Z_{ij} \geq 0} \frac{1}{2n}\sum_{i=1}^{n} w_i + \frac{1}{2m}\sum_{j=1}^{m} v_j \tag{3}$$

$$\text{s.t.} \sum_{y_j \in \text{ng}(x_i,\epsilon)} Z_{ij} + w_i = 1, \qquad \sum_{x_i \in \text{ng}(y_j,\epsilon)} Z_{ij} + v_j = 1, \quad \forall i,j$$

$$Z_{ij} = 0\,, \qquad \forall (i,j) \notin \text{ng}(x_i, \epsilon),$$

where $Z \in \mathbb{R}^{n \times m}$. When $n = m$, the optimization in (3) is identical to (2), as in this case the samples define a discrete distribution. However, when $n \neq m$ Problem (3) also accounts for the difference in the size of the two samples.

Problem (3) is a linear program with constraints that may be written as a totally unimodular matrix. It follows that one of the optimal solutions of (3) is integral [4]; that is, the mass of each sample

Algorithm 1 *Compute* $\widehat{PV}(S_1, S_2, \epsilon, \mathbf{d})$

Input: $S_1 = \{x_1, ..., x_n\}$ and $S_2 = \{y_1, ..., y_m\}$, ϵ rate, and distance measure $\mathbf{d}$.
1. Define $\hat{G} = (\hat{V} = (\hat{A}, \hat{B}), \hat{E})$: $\hat{A} = \{x_i \in S_1\}$, $\hat{B} = \{y_j \in S_2\}$,
   Connect an edge $e_{ij} \in \hat{E}$ if $d(x_i, y_j) \leq \epsilon$.
2. Compute the maximum matching on $\hat{G}$.
3. Define S_w and S_v as number of unmatched edges in sets S_1 and S_2 respectively.

Output: $\widehat{PV}(S_1, S_2, \epsilon, d) = \frac{1}{2}(\frac{S_w}{n} + \frac{S_v}{m})$.

is transferred as a whole. This solution may be found by solving the optimal assignment on an appropriate bipartite graph [4]. Let $G = (V = (A, B), E)$ define this graph, with $A = \{x_i, w_i\,; i = 1, ..., n\}$ and $B = \{y_j, v_j\,; j = 1, ..., m\}$ as its bipartite partition. The vertices $x_i \in A$ are linked with edge weight zero to $y_j \in \text{ng}(x_i)$ and with weight ∞ to $y_j \notin \text{ng}(x_i)$. In addition, every vertex x_i (y_j) is linked with weight 1 to w_i (v_j). To make the graph complete, assign zero cost edges between all vertices x_i and w_k for $k \neq i$ (and vertices y_j and v_k for $k \neq j$).

We note that the Earth Mover Distance (EMD) [5], a sampled version of the transportation problem, is also formulated by a linear program that may be solved by optimal assignment. For the EMD and other typical assignment problems, the computational complexity is more demanding, for example using the Hungarian algorithm it has an $O(N^3)$ complexity, where $N = n+m$ is the number of vertices [6]. Contrarily, graph G, which describes $\widehat{\text{PV}}$, is a simple bipartite graph for which maximum cardinality matching, a much simpler problem, can be applied to find the optimal assignment. To find the optimal assignment, first solve the maximum matching on the partial graph between vertices x_i, y_j that have zero weight edges (corresponding to neighboring vertices). Then, assign vertices x_i and y_j for whom a match was not found with w_i and v_j respectively; see Algorithm 1 and Figure 1 for an illustration of a matching. It is easy to see that the solution obtained solves the assignment problem associated with $\widehat{\text{PV}}$.

The complexity of Algorithm 1 amounts to the complexity of the maximal matching step and of setting up the graph, i.e., additional $O(nm)$ complexity of computing distances between all points. Let k be the average number of neighbors of a sample, then the average number of edges in the bipartite graph $\hat{G}$ is $|\hat{E}| = n \times k$. The maximal cardinality matching of this graph is obtained in $O(kn\sqrt{(n+m)})$ steps, in the worst case [6].

3 Related Work

Many scores have been defined for testing discrepancy between distributions. We focus on representative works for nonparametric tests that are most related to our work. First, we consider statistics for the Two Sample Problem (TSP), i.e., equality testing, that are based on the asymptotic distribution of the statistic conditioned on the equality. Among these tests is the well known Kolmogorov-Smirnov test (for one dimensional distributions), and its generalization to higher dimensions by minimal spanning trees [7]. A different statistic is defined by the portion of k-nearest neighbors of each sample that belongs to different distributions; larger portions mean the distributions are closer [8]. These scores are well known in the statistical literature but cannot be easily changed to test similarity, as their analysis relies on testing equality.

As discussed above, the 1^{st} Wasserstein metric and the TV metric have some relation to the PV. The EMD and histogram based L_1 distance are the sample based estimates of these metrics respectively. In both cases, the distance is not estimated directly on the samples, but on a higher level partition of the space: histogram bins or signatures (cluster centers). As a result, these estimators have inaccuracies. Contrarily, the PV is estimated directly on the samples and converges to its value between the underlying continuous distributions. We note that after a good choice of signatures, the EMD captures perceptual similarity, similar to that of the PV. It is possible to consider the PV as a refinement of the EMD notion of similarity; instead of clustering the data to signatures and moving the signatures, it perturbs each sample. In this manner it captures a finer notion of the perceptual similarity.

The last group of statistics are scores established in machine learning: the d_A distance presented by Kifer et al. that is based on the maximum discrepancy on a chosen subset of the support [9], and Maximum Mean Discrepancy (MMD) by Gretton et al., which define discrepancy after embeddings the distributions to a Reproducing Kernel Hilbert Space (RKHS)[10]. These scores have corresponding statistical tests for the TSP; however, since their analysis is based on finite convergence bounds, in principle they may be modified to test similarity. The d_A captures some intuitive notion of similarity, however, to our knowledge, it is not known how to compute it for a general subset class [2]. The MMD captures the distance between the samples in some RKHS. The MMD may be used to define a similarity test, yet this would require defining two parameters, σ and the similarity rate, whose dependency is not intuitive. Namely, for any similarity rate the result of the test is highly dependent on the choice of σ, but it is not clear how it should be made. Contrarily, the PV's parameter ϵ is related to the data's input domain and may be chosen accordingly.

4 Analysis

We present sample rate convergence analysis of the PV. Hypothesis testing for similarity may be preformed by applying the presented bounds. The proofs and the testing procedures are presented in [1]. When no clarity is lost, we omit $\mathbf{d}$ from the notation. Our main theorem is stated as follows:

Theorem 3. *Suppose we are given two i.i.d. samples $S_1 = \{x_1, ..., x_n\} \in \mathbb{R}^d$ and $S_2 = \{y_1, ..., y_m\} \in \mathbb{R}^d$ generated by distributions P and Q, respectively. Let the ground distance be $\mathbf{d} = \|\cdot\|_\infty$ and let $\mathcal{N}(\epsilon)$ be the cardinality of a disjoint cover of the distributions' support. Then, for any $\delta \in (0,1)$, $N = \min(n,m)$, and $\eta = \sqrt{\frac{2(\log(2(2^{\mathcal{N}(\epsilon)}-2))+\log(1/\delta))}{N}}$ we have that*

$$\mathbb{P}\left(\left|\widehat{PV}(S_1, S_2, \epsilon) - PV(P, Q, \epsilon)\right| \leq \eta\right) \geq 1 - \delta.$$

The theorem is defined using $\|\cdot\|_\infty$, but can be rewritten for other metrics (with a slight change of constants). The proof of the theorem exploits the form of the optimization Problem 3. We use the bound of Theorem 3 construct hypothesis tests. A weakness of this bound is its strong dependency on the dimension. Specifically, it is dependent on $\mathcal{N}(\epsilon)$, which for $\|\cdot\|_\infty$ is $O((1/\epsilon)^d)$: the number of disjoint boxes of volume ϵ^d that cover the support. Unfortunately, this convergence rate is inherent; namely, without making any further assumptions on the distribution, this rate is unavoidable and is an instance of the "curse of dimensionality". In the following theorem, we present a lower bound on the convergence rate.

Theorem 4. *Let $P = Q$ be the uniform distribution on $\mathbb{S}^{d-1}$, a unit $(d-1)$–dimensional hypersphere. Let $S_1 = \{x_1, ..., x_N\} \sim P$ and $S_2 = \{y_1, ..., y_N\} \sim Q$ be two i.i.d. samples. For any $\epsilon, \epsilon', \delta \in (0,1)$, $0 \leq \eta < 2/3$ and sample size $\frac{\log(1/\delta)}{2(1-3\eta/2)^2} \leq N \leq \frac{\eta}{2} e^{d(1-\frac{\epsilon^2}{2})/2}$, we have $PV(P, Q, \epsilon') = 0$ and*

$$\mathbb{P}(\widehat{PV}(S_1, S_2, \epsilon) > \eta) \geq 1 - \delta. \tag{4}$$

To remedy this dependency we propose a bootstrapping bias correcting technique, and a scheme to project the data to one dimension; see [1] for details.

5 Discussion

We present a new score that measures the similarity between two multivariate distributions, and assigns to it a value in the range [0,1]. The score is defined by the optimal permutation between the distributions, and its sensitivity is reflected by the parameter ϵ that allows for flexibility. The PV is efficiently estimated from samples. Its low computational complexity relies on its simple binary classification of points as neighbors or non-neighbor points, such that optimization of distances of faraway points is not needed. In this manner, the PV captures only the essential information to describe similarity. An added value of the PV is that its computation also gives insight to the areas of discrepancy; namely, the areas of the unmatched samples. In future work we plan to further explore this information, which may be valuable on its own merits.

[2] Most work with the d_A has been with the subset of characteristic functions, and approximated by the error of a classifier.

References

[1] Maayan Harel and Shie Mannor. The perturbed variation. *arXiv:1210.4006*, 2012.

[2] G. Monge. Mémoire sur la théorie des déblais et de remblais. Histoire de l'Academie Royale des Sciences de Paris, avec les Memoires de Mathematique et de Physique pour la meme annee, 1781.

[3] L. Rüschendorf. Monge–kantorovich transportation problem and optimal couplings. *Jahresbericht der DMV*, 3:113–137, 2007.

[4] A. Schrijver. *Theory of linear and integer programming*. John Wiley & Sons Inc, 1998.

[5] Y. Rubner, C. Tomasi, and L.J. Guibas. A metric for distributions with applications to image databases. In *Computer Vision, 1998. Sixth International Conference on*, pages 59–66. IEEE, 1998.

[6] R.K. Ahuja, L. Magnanti, and J.B. Orlin. *Network Flows: Theory, Algorithms, and Applications*, chapter 12, pages 469–473. Prentice Hall, 1993.

[7] J.H. Friedman and L.C. Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Annals of Statistics*, 7:697–717, 1979.

[8] M.F. Schilling. Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, pages 799–806, 1986.

[9] D. Kifer, S. Ben-David, and J. Gehrke. Detecting change in data streams. In *Proceedings of the Thirtieth international conference on Very large data bases*, pages 180–191. VLDB Endowment, 2004.

[10] A. Gretton, K. Borgwardt, B. Schölkopf, M. Rasch, and E. Smola. A kernel method for the two sample problem. In *Advances in Neural Information Processing Systems 19*, 2007.