RECENT RESULTS ON THE TOTAL VARIATION DISTANCE

A.M. Zubkov
Steklov Mathematical Institute of RAS
Moscow, RUSSIA
e-mail: zubkov@mi-ras.ru

Abstract

Let $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$ be two sets of independent discrete random variables. Explicit upper and lower bounds for the total variation distance between distributions of these sets are obtained in terms of some functions of distributions of separate components X_k and Y_k, $k = 1, \dots, n$. The cases of identical (inside each set) and arbitrary distributions of random variables are considered. Results may be used to estimate the sample sizes necessary or sufficient for testing two hypotheses with given sum of error probabilities.

Keywords: data science, total variation distance, hypotheses testing

1 Introduction

Let S be a countable set; let $P = \{p_s\}_{s\in S}$ and $Q = \{q_s\}_{s\in S}$ be probability distributions of random variables X and Y with values in S correspondingly. The total variation distance between probability distributions P and Q (or random variables X and Y) is defined by

$$d_{\mathrm{TV}}(P,Q) = d_{\mathrm{TV}}(X,Y) \overset{\mathrm{def}}{=} \sup_{A\subseteq S} |P(A) - Q(A)| = \frac{1}{2}\sum_{s\in S} |p_s - q_s|. \tag{1}$$

The value $1 - d_{\mathrm{TV}}(P,Q)$ is an exact low bound for the sum of error probabilities of two kinds in a problem of testing two simple hypothesis on the observation Z:
$H_0 :$ Z has distribution P,
$H_1 :$ Z has distribution Q.

So, estimates of total variation distance between distributions of sets of independent random variables $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$ considered as a function on n may be used to draw objective conclusions on the sample size necessary or sufficient to distinguish simple hypotheses on such distributions.

Recently the upper and lower estimates of the total variation distance between samples $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$ of independent identically distributed observations were obtained in [1, 2]. Under some conditions these estimates has the order $d_{\mathrm{tv}}(X_1, Y_1)\sqrt{n}$ with coefficients depending on the distributions of X_1 and Y_1. In the general case the upper bound cannot be smaller than $d_{\mathrm{tv}}(X_1, Y_1)\, n$.

Here we state lower and upper estimates of the total variation distance between the distributions of sets $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$ of independent random variables for cases of identically (within each set) or arbitrary distributed random variables. The domains of applicability of our inequalities are wider than that of [1, 2].

2 Identically distributed components

It is easy to see that if P and Q are probability distributions on a countable set S and $C_{P,Q} = \{x \in S \colon P(x) > Q(x)\}$, then $d_{\text{tv}}(P,Q) = \sum_{x \in C_{P,Q}}(P(x) - Q(x))$. Put

$$v = v(P,Q) = \min\{P(C_{P,Q}), Q(C_{P,Q}), P(\Omega \setminus C_{P,Q}), Q(\Omega \setminus C_{P,Q})\},$$

then $0 < v \le \frac{1}{2}(1 - d_{\text{tv}}(P,Q))$, and these estimates are best possible.

Theorem 1. *Let $X_1, \ldots, X_n$ be independent random variables with the same distribution P on the countable set S, $Y_1, \ldots, Y_n$ be independent random variables with the same distribution Q on S and $d_{\text{tv}}(P,Q) = \varepsilon > 0$. Then*

$$d_{\text{tv}}((X_1, \ldots, X_n), (Y_1, \ldots, Y_n)) \ge \frac{e^{-4/nv(1-v)}}{2(1+2/n)} \sqrt{\frac{v}{1-v}} \left(\Phi\left(2\varepsilon\sqrt{n}\right) - \tfrac{1}{2}\right),$$

where $v = v(P,Q) \in [\frac{2}{n}, \frac{n-1}{2n})$ and $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du$ is a standard normal distribution function.

Remark 1. The difference $\Phi\left(2\varepsilon\sqrt{n}\right) - \frac{1}{2}$ does not exceed $\frac{1}{2}$ and is equivalent to $\varepsilon\sqrt{2n/\pi}$ for $\varepsilon\sqrt{n} \to 0$. The estimate of theorem 1 cannot be larger than $\frac{1}{4}$; for $n > 2\varepsilon^{-2} \ln 2$ the known low estimate (see, e. g., [3, 4]])

$$d_{\text{tv}}((X_1, \ldots, X_n), (Y_1, \ldots, Y_n)) \ge 1 - 2e^{-n\varepsilon^2/2},$$

is nontrivial and tends to 1 as $n \to \infty$.

Theorem 2. *Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be independent random variables taking values $1, \ldots, N$:*

$$\mathbf{P}\{X_t = k\} = p_k, \quad \mathbf{P}\{Y_t = k\} = r_k, \quad k \in \{1, \ldots, N\}, \quad t = 1, 2 \ldots, \quad d_{\text{tv}}(X_1, Y_1) = \varepsilon.$$

Then

$$d_{\text{tv}}((X_1, \ldots, X_n), (Y_1, \ldots, Y_n)) \le \varepsilon\sqrt{n} \left(\frac{1}{\sqrt{S_X} + \sqrt{S_X + \varepsilon}} + \frac{1}{\sqrt{S_Y} + \sqrt{S_Y + \varepsilon}} \right),$$

where $S_X = \sum_{k \colon p_k < r_k} p_k$, $S_Y = \sum_{k \colon r_k < p_k} r_k$.

Remark 2. Upper bound may be very large if S_X or S_Y is very small, but in such cases the total variation distance between $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ also may be large. For example, if $N = 3$, $p_1 = r_2 = \varepsilon$, $p_2 = r_1 = 0$, $p_3 = q_3$, then $d_{\text{tv}}(X_1, Y_1) = \varepsilon$, $S_X = S_Y = 0$ and

$$d_{\text{tv}}((X_1, \ldots, X_n), (Y_1, \ldots, Y_n)) \ge \frac{1}{2}(\mathbf{P}\{\exists k \colon X_k = 1\} + \mathbf{P}\{\exists k \colon Y_k = 2\}) \ge n\varepsilon(1-\varepsilon)^{n-1},$$

upper bound of theorem 2 in this case equals $2\sqrt{n\varepsilon}$.

Theorems 1 and 2 were proved in [6]. Another form of theorem 2 may be found in [5].

3 Arbitrary distributions of components

Theorem 3. *Let $X_1, X_2, \dots$ and $Y_1, Y_2, \dots$ be independent random variables taking values* $1, \dots, N$*:*

$$\mathbf{P}\{X_k = j\} = p_j^{(k)}, \quad \mathbf{P}\{Y_k = j\} = r_j^{(k)}, \quad j \in \{1, \dots, N\}, \quad k = 1, 2 \dots, n,$$

and $\rho_k = \rho(X_k, Y_k) = \frac{1}{2}\sum_{j=1}^{N} |p_j^{(k)} - r_j^{(k)}| > 0\,(k = 1, \dots, n)$, $S = \sum_{k=1}^{n} \rho_k$. *Then*

$$d_{\text{tv}}((X_1, \dots, X_n), (Y_1, \dots, Y_n)) \geq \frac{S}{3(2 + S + \sqrt{2n \ln(n/S)})}.$$

If $\rho_1 = \ldots = \rho_n = \rho < 1$, then the estimate takes the form

$$d_{\text{tv}}((X_1, \dots, X_n), (Y_1, \dots, Y_n)) \geq \frac{n\rho}{3\left(2 + n\rho + \sqrt{2n \ln(\frac{1}{\rho})}\right)} = \frac{\rho\sqrt{n}}{3\left(\frac{2}{\sqrt{n}} + \rho\sqrt{n} + \sqrt{2\ln(\frac{1}{\rho})}\right)}.$$

Modifying the proof it is possible to obtain low bound which is arbitrary close to 1 for fixed ρ and sufficiently large n.

Theorem 4. *Let $X_1, X_2, \dots$ and $Y_1, Y_2, \dots$ be independent random variables taking values* $1, \dots, N$*:*

$$\mathbf{P}\{X_t = k\} = p_k^{(t)}, \quad \mathbf{P}\{Y_t = k\} = r_k^{(t)}, \quad k \in \{1, \dots, N\}, \quad t = 1, 2 \dots, n,$$

$\rho_t = \rho(X_t, Y_t) = \frac{1}{2}\sum_{k=1}^{N} |p_k^{(t)} - r_k^{(t)}|$, $t = 1, \dots, n$, *and*

$$\min_{1 \leq t \leq n} \min\{S_X^{(t)}, S_Y^{(t)}\} \geq \delta > 0, \quad S_X^{(t)} = \sum\nolimits_{k:\, p_k^{(t)} < r_k^{(t)}} p_k^{(t)}, \; S_Y^{(t)} = \sum\nolimits_{k:\, r_k^{(t)} < p_k^{(t)}} r_k^{(t)}.$$

Then

$$\rho((X_1, \dots, X_n), (Y_1, \dots, Y_n)) \leq \frac{1}{\sqrt{2\delta}}\sqrt{\sum\nolimits_{t=1}^{n} \rho_t^2}.$$

Condition $\min_{1 \leq t \leq n} \min\{S_X, S_Y\} \geq \delta > 0$ exclude cases mentioned in Remark 2.

References

[1] Reyzin L. (2004). Preprint: A note on the statistical difference of small direct products. *Boston Univ. Computer Science, Techn. Rep. BUCS-TR-2004-032.*

[2] Renner R. (2005). On the variational distance of independently repeated experiments. *arXiv.*

[3] Sahai A., Vadhan S. (1999). Manipulating statistical difference. *Randomization Methods in Algorithm Design. Proc. DIMACS Workshop, December 1997. DIMACS Ser. in Discr. Math. and Theor. Comput. Sci. Vol. 43. Amer. Math. Soc. Providence, R.I.* p. 251–270.

[4] Sahai A., Vadhan S. (2003). A complete problem for statistical zero knowledge. *J. ACM.* Vol. 50, Iss. 2, pp. 196–249.

[5] Zubkov A. M. (2013). New inequalities for the binomial law and for the total variation distance between iid samples. Proc. 10th Int. Conf. “Computer Data Analysis and Modeling”, v. 2. Publ. center of BSU, Minsk, p. 48–50.

[6] Zubkov A. M. (2017). New estimates for the variational distance between two distributions of a sample. *Matematicheskie voprosy kriptografii.* Vol. 9, Iss. 3, pp. 45–60.