ADAPTIVE COMMUNICATIONS AND SIGNAL PROCESSING LABORATORY
CORNELL UNIVERSITY, ITHACA, NY 14853

# On $\mathcal{A}$-distance and Relative $\mathcal{A}$-distance

**Ting He and Lang Tong**

Technical Report No. ACSP-TR-08-04-02

August 2004

## I. Introduction

We give a method to measure the distance between two probability distributions, and based on the distance measure, we bound the probability that the distance between the empirical distribution and the actual distribution exceeds certain level. The direct implication of our result is that for large sample size, one can replace the actual probability with its corresponding empirical probability with arbitrarily small error. The proof of the theorems is based on the Vapnik-Chervonenkis Theory [6] and Anthony & Shawe-Taylor's extension of the Vapnik-Chervonenkis Theory [4].

## II. Distance measure

**$\mathcal{A}$-distance** Fix a measure space and let $\mathcal{A}$ be a collection of measurable sets. Let $P_1$ and $P_2$ be probability distributions over this space. The *$\mathcal{A}$-distance* between $P_1$ and $P_2$ is defined as

$$d_{\mathcal{A}}(P_1, P_2) = \sup_{A \in \mathcal{A}} |P_1(A) - P_2(A)|$$

For finite sample sets $S_1$ and $S_2$, $d_{\mathcal{A}}(S_1, S_2)$ is defined similarly by replacing $P_i(A)$ with $S_i(A) \triangleq |S_i \cap A|/|S_i|$.

The following notion of *relative $\mathcal{A}$-distance* offers a way to take the relative magnitude of a change into account.

**Relative $\mathcal{A}$-distance** Let $P_1, P_2$ be two probability distributions over the same measure space, let $\mathcal{A}$ denote a family of measurable subsets of that space, and $A$ a set in $\mathcal{A}$. The relative $\mathcal{A}$-distance between $P_1$ and $P_2$ is defined as

$$\phi_{\mathcal{A}}(P_1, P_2) \quad = \quad \sup_{A \in \mathcal{A}} \frac{|P_1(A) - P_2(A)|}{\sqrt{\frac{P_1(A)+P_2(A)}{2}}}$$

For empirical distances, simply replace $P_i(A)$ with the empirical measure $S_i(A) = |S_i \cap A|/|S_i|$.

It is easy to see that $\mathcal{A}$-distance is a metric. For the proof of relative $\mathcal{A}$-distance as a metric, see [5].

## III. VC Bounds

The following theorems are derived from [6] and [4] to guarantee the rate that the empirical distance converges to the underlying distance for both distance notions.

For *$\mathcal{A}$-distance*, we have the following theorems:

*Theorem 3.1 (Vapnik-Chervonenkis Inequality):* Let $P$ be a probability distribution over domain $X$ and $S$ be a collection of $n$ i.i.d samples drawn from $P$. Then for a family of subsets of $X$ $\mathcal{A}$ and a constant $\epsilon \in (0,1)$,

$$P^n\{\sup_{A\in\mathcal{A}} |S(A) - P(A)| > \epsilon\} \leq 4\Pi_{\mathcal{A}}(2n)^1 e^{-n\epsilon^2/8}$$

Using Theorem 3.1, it is easy to derive the following corollary.

*Corollary 3.2:* Let $P_1,\ P_2$ be any probability distributions over some domain $X$ and let $\mathcal{A}$ be a family of subsets of $X$ and $\epsilon \in (0,1)$. If $S_1,\ S_2$ are i.i.d $n$ samples drawn by $P_1,\ P_2$ respectively, then,

$$P^{2n}\left[\exists A \in \mathcal{A}\,||P_1(A) - P_2(A)| - |S_1(A) - S_2(A)|| \geq \epsilon\right]$$

$$< 8\Pi_{\mathcal{A}}(2n)e^{-n\epsilon^2/32}$$

Where $P^{2n}$ in the above inequality is the probability over the pairs of samples $(S_1, S_2)$ induced by the sample generating distributions $(P_1, P_2)$.

*Proof:* Simple algebra yields the result.

$$\Pr\{\exists A \in \mathcal{A}, ||P_1(A) - P_2(A)| - |S_1(A) - S_2(A)|| \geq \epsilon\}$$

$$\begin{aligned}
&\leq \Pr\{\sup_{A\in\mathcal{A}} |P_1(A) - P_2(A) - S_1(A) + S_2(A)| \geq \epsilon\} && (1)\\
&\leq \Pr\{\sup_{A\in\mathcal{A}} |P_1(A) - S_1(A)| + |P_2(A) - S_2(A)| \geq \epsilon\} && (2)\\
&\leq \Pr\{\{\sup_{A\in\mathcal{A}} |P_1(A) - S_1(A)| \geq \tfrac{\epsilon}{2}\}\\
&\qquad \cup\{\sup_{A\in\mathcal{A}} |P_2(A) - S_2(A)| \geq \tfrac{\epsilon}{2}\}\} && (3)\\
&\leq \Pr\{\sup_{A\in\mathcal{A}} |P_1(A) - S_1(A)| \geq \tfrac{\epsilon}{2}\}\\
&\qquad + \Pr\{\sup_{A\in\mathcal{A}} |P_2(A) - S_2(A)| \geq \tfrac{\epsilon}{2}\} && (4)\\
&\leq 8\Pi_{\mathcal{A}}(2n)e^{-n\epsilon^2/32} && (5)
\end{aligned}$$

where the last inequality comes from Theorem 3.1. ■

[1] $\Pi_{\mathcal{A}}(2n)$ is the *shatter coefficient* [3]. If $\mathcal{A}$ has a finite VC-dimension $d$, then by Sauer's Lemma, $\Pi_{\mathcal{A}}(2n) < (2n+1)^d$ for all $n$.

We thus have ways to bound the probability that empirical $\mathcal{A}$-distance deviates from true $\mathcal{A}$-distance from both sides.

The theoretical guarantee can be improved by considering *relative $\mathcal{A}$-distance*. We can get results similar to Theorem 3.1 and Corollary 3.2 for the metric $\phi_{\mathcal{A}}(P_1, P_2)$. We start with the following result of Anthony and Shawe-Taylor [4].

*Lemma 3.3:* Let $\mathcal{A}$ be a family of subsets of the domain $X$. $P$ is any probability distribution over $X$. If $S_1$ and $S_2$ are two collections of $n$ samples each, drawn i.i.d. from $P$, then

$$P^{2n}(\phi_{\mathcal{A}}(S_1, S_2) > \epsilon) \leq 2\Pi_{\mathcal{A}}(2n)e^{-n\epsilon^2/4}$$

(where $P^{2n}$ is the probability that $P$ induces over the choice of samples.)

In [4], Anthony and Shawe-Taylor proved that

$$\Pr\{\sup_{A\in\mathcal{A}} \frac{S_1(A) - S_2(A)}{\sqrt{\frac{S_1(A)+S_2(A)}{2}}} > \epsilon\} \leq \Pi_{\mathcal{A}}(2n)e^{-n\epsilon^2/4}$$

By symmetry of $S_1(A)$ and $S_2(A)$, the result in Lemma 3.3 holds.

*Theorem 3.4:* Let $\mathcal{A}$ be a family of subsets of the domain $X$, $P$ be any probability distribution over $X$, and $S$ be a set of $n$ samples, each drawn i.i.d. from $P$. Then

$$P^n(\phi_{\mathcal{A}}(S, P) > \epsilon) \leq 8\Pi_{\mathcal{A}}(2n)e^{-n\epsilon^2/4}$$

(Where $P^n$ is the $n$'th power of $P$ - the probability that $P$ induces over the choice of samples). The proof of this theorem is similar to the proof in [4].

*Proof:* Define

$$Q = \left\{ S \in X^n : \exists A \in \mathcal{A}\ s.t.\ \frac{P(A) - S(A)}{\sqrt{\frac{P(A)+S(A)}{2}}} > \epsilon \right\}$$

$$R = \left\{ SS' \in X^{2n} : \exists A \in \mathcal{A}\ s.t.\ \frac{S(A) - S'(A)}{\sqrt{\frac{S(A)+S'(A)}{2}}} > \epsilon \right\}$$

where $S$, $S'$ are two sets of n samples, drawn i.i.d. from $P$.

Then we claim that $\Pr(Q) \leq 4\Pr(R)$ for $n > \frac{4}{\epsilon^2}$. This is true because of the following. Suppose $S \in Q$, so there is $C \in \mathcal{A}$ s.t. $P(C) - S(C) > \epsilon\sqrt{\frac{P(C)+S(C)}{2}}$. Hence $S(C) < P(C) + \frac{\epsilon^2}{4} - \epsilon\sqrt{\frac{\epsilon^2}{16} + P(C)}$. Noting that $S(C) \geq 0$, some simple calculation shows that $P(C) > \frac{\epsilon^2}{2}$.

If we draw another set of $n$ samples $S'$, each drawn i.i.d. from $P$, and define

$$F = \frac{S'(C) - S(C)}{\sqrt{\frac{S'(C)+S(C)}{2}}}$$

we have $F > \epsilon$ if $S'(C) > P(C)$. This is because the function $f(x,y) = \frac{x-y}{\sqrt{(x+y)/2}}$ is monotone increasing w.r.t. $x$ and monotone decreasing w.r.t. $y$ on $x, y \in (0,1)$(taking derivative easily verifies it). So $\inf F$ is achieved when $S(C) = P(C) + \frac{\epsilon^2}{4} - \epsilon\sqrt{\frac{\epsilon^2}{16} + P(C)}$ and $S'(C) = P(C)$. Plugging in yields the value $\epsilon$, and the strict inequality follows from the strict inequalities about $S(C)$ and $S'(C)$.

The random variable $nS'(C)$ has binomial distribution $\mathcal{B}(n, P(C))$. For $n > \frac{4}{\epsilon^2} > \frac{2}{P(C)}$, $nS'(C) > nP(C)$ with probability$\geq 1/4$( [4]).

Therefore for $n > \frac{4}{\epsilon^2}$, we have $\Pr(Q) \leq 4\Pr(R)$. In [4], it is proved that

$$\Pr(R) \leq \Pi_{\mathcal{A}}(2n)e^{-n\epsilon^2/4}.$$

Thus

$$P^n\left\{\sup_{A\in\mathcal{A}} \frac{P(A) - S(A)}{\sqrt{\frac{P(A)+S(A)}{2}}} > \epsilon\right\} \leq 4\Pi_{\mathcal{A}}(2n)e^{-n\epsilon^2/4}$$

Note that this inequality is trivially satisfied if $n \leq \frac{4}{\epsilon^2}$.

By symmetry we have

$$P^n(\phi_{\mathcal{A}}(S,P) > \epsilon) \leq 8\Pi_{\mathcal{A}}(2n)e^{-n\epsilon^2/4}.$$

■

Similar to Corollary 3.2, we have the following corollary of Theorem 3.4 which bounds the probability that the empirical *relative $\mathcal{A}$-distance* deviates from the true *relative $\mathcal{A}$-distance*.

*Corollary 3.5:* Let $P_1,\ P_2$ be any probability distributions over some domain $X$ and let $\mathcal{A}$ be a family of subsets of $X$ and $\epsilon \in (0,1)$. If $S_1,\ S_2$ are two collections of $n$ samples each, drawn i.i.d. from $P_1,\ P_2$ respectively, then

$$P^{2n}[|\phi_{\mathcal{A}}(P_1,P_2) - \phi_{\mathcal{A}}(S_1,S_2)| > \epsilon]$$

$$\leq 16\Pi_{\mathcal{A}}(2n)e^{-n\epsilon^2/16}$$

Where $P^{2n}$ in the above inequality is the probability over the pairs of samples $(S_1, S_2)$ induced by the sample generating distribution $(P_1, P_2)$.

*Proof:* Because $\phi_{\mathcal{A}}(\cdot,\cdot)$ is a metric on $[0,1]$( [5]), we have

$$\phi_{\mathcal{A}}(P_1,P_2) \leq \phi_{\mathcal{A}}(P_1,S_1) + \phi_{\mathcal{A}}(S_1,S_2) + \phi_{\mathcal{A}}(S_2,P_2)$$

and

$$\phi_{\mathcal{A}}(P_1,P_2) \geq \phi_{\mathcal{A}}(S_1,S_2) - \phi_{\mathcal{A}}(P_1,S_1) - \phi_{\mathcal{A}}(S_2,P_2)$$

Therefore,

$$\Pr\{|\phi_{\mathcal{A}}(P_1,P_2) - \phi_{\mathcal{A}}(S_1,S_2)| > \epsilon\}$$

$$\leq \quad \Pr\{\phi_{\mathcal{A}}(P_1,S_1) + \phi_{\mathcal{A}}(P_2,S_2) > \epsilon\} \tag{6}$$

$$\leq \quad \Pr\{\phi_{\mathcal{A}}(P_1,S_1) > \tfrac{\epsilon}{2}\} + \Pr\{\phi_{\mathcal{A}}(P_2,S_2) > \tfrac{\epsilon}{2}\} \tag{7}$$

$$\leq \quad 16\Pi_{\mathcal{A}}(2n)e^{-n\epsilon^2/16} \tag{8}$$

where the last inequality comes from Theorem 3.4. ■

## REFERENCES


[1] B. Brodsky and B. Darkovsky, *Non-Parametric Methods in Change-Point Problems*, Kluwer Academic, The Netherlands, 1993.

[2] J. Shao, *Mathematical Statistics*, Springer, 1999.

[3] L. Gyorfi, *Principles of Nonparametric Learning*, Springer Wien New York, 2002.

[4] M. Anthony and J. Shawe-Taylor, "A result of Vapnik with applications", in Discrete and Applied Mathematics, vol. 47(2), pp. 207-217, 1993.

[5] S. Ben-David, J. Gehrke and D. Kifer, "Detecting Change in Data Streams", in Proc. 2004 VLDB Conference, (Toronto, Canada), 2004.

[6] V.N. Vapnik and A. Ya. Chervonenkis "On the uniform convergence of relative frequency of events to their probabilities" in Theory of Probability and its Applications, Vol. 16, pp 264-280, 1971.