The final form of Tao's inequality relating conditional expectation and conditional mutual information

Rudolf Ahlswede*
Department of Mathematics
University of Bielefeld
POB 100131, D-33501 Bielefeld, Germany
Email: ahlswede@math.uni-bielefeld.de

To Jack's memory
crystalline and amorph views
keep always distinct!

Abstract

Recently Terence Tao approached Szemerédi's Regularity Lemma from the perspectives of Probability Theory and of Information Theory instead of Graph Theory and found a stronger variant of this lemma, which involves a new parameter.

To pass from an entropy formulation to an expectation formulation he found the following

Lemma. *Let Y, and X, X' be discrete random variables taking values in $\mathcal{Y}$ and $\mathcal{X}$, respectively, where $\mathcal{Y} \subset [-1,1]$, and with $X' = f(X)$ for a (deterministic) function f.*

Then we have

$$\mathbb{E}\big(|\mathbb{E}(Y|X') - \mathbb{E}(Y|X)|\big) \leq 2I(X \wedge Y|X')^{\frac{1}{2}}.$$

We show that the constant 2 can be improved to $(2\ell n 2)^{\frac{1}{2}}$ and that this is the best possible constant.

A word about notation

We have replaced Tao's X, Y, Y' by Y, X, X', because we are used to have $\mathcal{X}$ as input alphabet and $\mathcal{Y}$ as output alphabet of a channel. We find it also convenient to use P_Z for the distribution of a random variable (RV) Z and an analogous notation for conditional distributions of two RV's.

From Pinsker's inequality to Tao's inequality

Beginning, as Tao did, with the special case $X' = f(X) =$ constant, the inequality can actually readily be derived from Pinsker's inequality [2]. Indeed

$$\begin{aligned}
&\mathbb{E}\big(|\mathbb{E}(Y) - \mathbb{E}(Y|X)|\big) \\
&= \sum_{x\in\mathcal{X}} P_X(x)|\mathbb{E}(Y) - \mathbb{E}(Y|X=x)| \\
&= \sum_{x\in\mathcal{X}} P_X(x)\Big|\sum_{y\in\mathcal{Y}} y\big(P_Y(y) - P_{Y|X}(y|x)\big)\Big| \\
&\leq \sum_{x\in\mathcal{X}} P_X(x)\sum_{y\in\mathcal{Y}} |P_Y(y) - P_{Y|X}(y|x)|, \qquad (1)
\end{aligned}$$

*The author was supported by the 'Finite Structures' Marie Curie Host Fellowship for the Transfer of knowledge project carried out by Alfred Renyi Institute of Mathematics, in the framework of the European Community's Structuring the European Research Area programme.

because $\mathcal{Y} \subset [-1, 1]$.

Since

$$\sum_{y\in\mathcal{Y}} |P_Y(y) - P_{Y|X}(y|x)| = \|P_Y - P_{Y|X=x}\|_1$$
$$\leq (2\ell n 2)^{\frac{1}{2}} D(P_{Y|X=x}\|P_Y)^{\frac{1}{2}} \tag{2}$$

(Pinsker's inequality) and the square-root function is concave, we finally get

$$\begin{aligned}
&\mathbb{E}\big(|\mathbb{E}(Y) - \mathbb{E}(Y|X)|\big) \\
&\leq (2\ell n 2)^{\frac{1}{2}} \left(\sum_{x\in\mathcal{X}} P_X(x) D(P_{Y|X=x}\|P_Y)\right)^{\frac{1}{2}} \\
&= (2\ell n 2)^{\frac{1}{2}} I(X \wedge Y)^{\frac{1}{2}},
\end{aligned} \tag{3}$$

by definition of mutual information.

Now we go to the general case $X' = f(X)$

$$\begin{aligned}
&\mathbb{E}\big(|\mathbb{E}(Y|X') - \mathbb{E}(Y|X)|\big) \\
&= \sum_{x'\in\mathcal{X}} P_{X'}(x') \mathbb{E}\big(|\mathbb{E}(Y|X' = x') \\
&\quad - \mathbb{E}(Y|X; X' = x')|\big) \\
&\leq \sum_{x'\in\mathcal{X}} P_{X'}(x') (2\ell n 2)^{\frac{1}{2}} I(X \wedge Y|X' = x')^{\frac{1}{2}}
\end{aligned}$$

(by (3)) and again by the concavity of the square-root function we get part (a) in the

Lemma (Relation between conditional expectations and conditional mutual information)[1]

Let Y, X, X' be discrete random variables with $X' = f(X)$ and with Y taking values in the interval $[-1, 1]$, then

(a) $\mathbb{E}\big(|\mathbb{E}(Y|X') - \mathbb{E}(Y|X)|\big) \leq (2\ell n 2)^{\frac{1}{2}} I(X \wedge Y|X')^{\frac{1}{2}}$

(b) The constant in the inequality is best.

Proof of (b):

We look at the case $X' =$ constant and pairs of RV's (Y, X) with distributions parametrized by δ $\left(0 < \delta \leq \frac{1}{2}\right)$

$\mathcal{X} = \{x_1, x_2\}$, $\mathcal{Y} = \{y_1, y_2\} = \{+1, -1\}$

$P_X = \left(\frac{1}{2}, \frac{1}{2}\right)$, $P_Y = \left(\frac{1}{2}, \frac{1}{2}\right)$,

$$P_{Y|X}(y_i|x_j) = \begin{cases} 1 - \delta, & \text{if } i = j \\ \delta, & \text{if } i \neq j \end{cases}.$$

Then

$$\mathbb{E}\big(|\mathbb{E}(Y) - \mathbb{E}(Y|X)|\big) = 1 - 2\delta$$
$$= \sum_{x\in\mathcal{X}} P_X(x) \sum_{y\in\mathcal{Y}} |P_Y(y) - P_{Y|X}(y|x)|$$

[1] Actually the inequality more generally holds if the condition $X' = f(X)$ is weakened to Y, X, X' satisfy a Markov relation, but that is not needed by Tao

and there is equality in (1).

Moreover,

$$\sum_{y\in\mathcal{Y}} |P_Y(y) - P_{Y|X}(y|x_1)|$$
$$= \sum_{y\in\mathcal{Y}} |P_Y(y) - P_{Y|X}(y|x_2)|$$
$$= \|P_Y - P_{Y|X=x_1}\|_1 = \|P_Y - P_{Y|X=x_2}\|_1.$$

Also

$$D(P_{Y|X=x_1}\|P_Y) = D(P_{Y|X=x_2}\|P_Y) = I(X \wedge Y).$$

It suffices to show that

$$\sup_{0<\delta<\frac{1}{2}} (1-2\delta) D\left((1-\delta,\delta)\| \left(\frac{1}{2},\frac{1}{2}\right)\right)^{-\frac{1}{2}} = (2\ell n 2)^{\frac{1}{2}}. \tag{4}$$

Since $D\left((1-\delta,\delta)\| \left(\frac{1}{2},\frac{1}{2}\right)\right) = (1-\delta)\log 2(1-\delta) + \delta \log 2\delta = 1 + (1-\delta)\log(1-\delta) + \delta\log\delta$, we already know that for $c = \frac{1}{2\ell n 2}$

$$g(\delta) \triangleq 1 + (1-\delta)\log(1-\delta) + \delta\log\delta - c(1-2\delta)^2 \geq 0 \tag{5}$$

and that $g\left(\frac{1}{2}\right) = 0$.

It suffices now to show that the biggest value for c such that (5) holds for all $\delta < \frac{1}{2}$ is $\frac{1}{2\ell n 2}$.

Since g is differentiable and $g\left(\frac{1}{2}\right) = 0$ it suffices to show that

$$\frac{dg(\delta)}{d\delta} = \frac{1}{\ell n 2} \ell n \left(\frac{\delta}{1-\delta}\right) + 4c(1-2\delta) > 0$$

for all δ in a neighbourhood of $\frac{1}{2}$, if $c = \frac{1+\varepsilon}{2\ell n 2}$, $\varepsilon > 0$.

For this sufficient is, using $t - 1 - \frac{(t-1)^2}{2}$ as lower bound of $\ell n t$, that

$$\frac{\delta}{1-\delta} - 1 - \frac{\left(\frac{\delta}{1-\delta} - 1\right)^2}{2} + 2(1+\varepsilon)(1-2\delta) > 0$$

for all δ in a neighbourhood of $\frac{1}{2}$.

Equivalently,

$$\frac{2\delta-1}{1-\delta} - \frac{1}{2}\left(\frac{2\delta-1}{1-\delta}\right)^2 + 2(1+\varepsilon)(1-2\delta) > 0$$

or

$$-\frac{1}{1-\delta} + \frac{1}{2}\frac{2\delta-1}{(1-\delta)^2} + 2(1+\varepsilon) > 0$$

or

$$\frac{-2+2\delta+2\delta-1}{2(1-\delta)^2} + 2(1+\varepsilon) = \frac{4\delta-3}{2(1-\delta)^2} + 2(1+\varepsilon) > 0$$

or

$$2(1+\varepsilon) > \frac{3-4\delta}{2(1-\delta)^2},$$

which is the case for $\delta = \frac{1}{2}$ and a neighbourhood of smaller δ's.

References

[1] T. Tao, Szemeredi's regularity lemma revisited, preprint 2005.

[2] M.S. Pinsker, Information and Information Stability of Random Variables and Processes (in Russian), Vol. 7 of the series Problemy Peredaci Informacii, AN SSSR, Moscow, 1960; English translation: Holden-Day, San Francisco, 1964.