Strong Data Processing Constant is Achieved by Binary Inputs

Or Ordentlich and Yury Polyanskiy

***Abstract*—For any channel $P_{Y|X}$ the strong data processing constant is defined as the smallest number $\eta_{KL} \in [0,1]$ such that $I(U;Y) \le \eta_{KL} I(U;X)$ holds for any Markov chain $U - X - Y$. It is shown that the value of η_{KL} is given by that of the best binary-input subchannel of $P_{Y|X}$. The same result holds for any f-divergence, verifying a conjecture of Cohen, Kemperman and Zbaganu (1998).**

Consider an arbitrary channel $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ with countable $\mathcal{X}$. We define the strong data processing inequality (SDPI) constant [1]

$$\eta_{\mathrm{KL}} = \sup \frac{D(P_{Y|X} \circ P \| P_{Y|X} \circ Q)}{D(P\|Q)}\,, \tag{1}$$

where optimization is over all pairs of distributions on $\mathcal{X}$, denoted $P, Q \in \mathcal{P}(\mathcal{X})$, such that $0 < D(P\|Q) < \infty$, and $P_{Y|X} \circ P$ is the distribution of the output Y when the input X is distributed according to $P \in \mathcal{P}(\mathcal{X})$. We refer to [2] for a survey of the properties and importance of the SDPI, in particular for showing equivalence to the definition in the abstract, and advertise [3] as a recent application in statistics.

When the input alphabet $\mathcal{X}$ is binary, the value of η_{KL} is relatively easy to compute, cf. [2, Appendix B]. Here we prove that for general $\mathcal{X}$ determination of η_{KL} can be reduced to the binary case.

Theorem 1: Optimization in (1) can be restricted to pairs P, Q supported on two points in $\mathcal{X}$ (same for both).

Proof. For two distributions P and Q on $\mathcal{X}$ and $\lambda \in (0,1)$ define

$$L_\lambda(P,Q) \triangleq D(P_{Y|X} \circ P \| P_{Y|X} \circ Q) - \lambda D(P\|Q).$$

We assume that $0 < D(P\|Q) < \infty$ as required by the definition of η_{KL}. We will show that we can find two distributions $\hat{P}$ and $\hat{Q}$ where $\hat{Q}$ is supported on two letters in $\mathrm{supp}(Q) \triangleq \{x \in \mathcal{X} : Q(x) > 0\}$, and $L_\lambda(\hat{P},\hat{Q}) \ge L_\lambda(P,Q)$. This implies the statement, since $\eta_{\mathrm{KL}} = \sup\left\{\lambda : \sup_{P,Q} L_\lambda(P,Q) \ge 0\right\}$.

O. Ordentlich is with the Hebrew University of Jerusalem, Israel (`or.ordentlich@mail.huji.ac.il`). Y. Polyanskiy is with the MIT, USA (`yp@mit.edu`). This work was supported, in part, by ISF under Grant 1791/17, by the MIT-IBM Watson AI Lab and by the National Science Foundation under the grants agreement ECCS-1808692 and CCF-2131115.

To that end define the convex set of distributions

$$\mathcal{S} \triangleq \left\{\hat{Q} : \mathrm{supp}(\hat{Q}) \subseteq \mathrm{supp}(Q), \sum_{x \in \mathrm{supp}(Q)} \frac{P(x)}{Q(x)} \cdot \hat{Q}(x) = 1\right\}.$$

Consider the function $g : \mathcal{S} \to \mathbb{R}$ defined as $g(\hat{Q}) = L_\lambda\left(\frac{P}{Q}\hat{Q}, \hat{Q}\right)$. Note that $Q \in \mathcal{S}$ and $g(Q) = L_\lambda(P,Q)$. Consequently, $\max_{\hat{Q} \in \mathcal{S}} g(\hat{Q}) \ge L_\lambda(P,Q)$. Note that

$$\hat{Q} \mapsto D\left(P_{Y|X} \circ \frac{P}{Q}\hat{Q} \middle\| P_{Y|X} \circ \hat{Q}\right)$$

is convex by convexity of $(P,Q) \mapsto D(P\|Q)$, and that

$$\hat{Q} \mapsto D\left(\frac{P}{Q}\hat{Q} \middle\| \hat{Q}\right) = \sum_x \hat{Q}(x) \frac{P(x)}{Q(x)} \log \frac{P(x)}{Q(x)}$$

is linear. Thus, $\hat{Q} \mapsto g(\hat{Q})$ is convex on $\mathcal{S}$. It therefore follows that $\max_{\hat{Q} \in \mathcal{S}} g(\hat{Q})$ is obtained at an extreme point of $\mathcal{S}$. Since $\mathcal{S}$ is the intersection of the simplex with a hyperplane, its extreme points are supported on at most two atoms. ■

Paired with [2, Appendix B] we get a corollary bounding η_{KL} in terms of the Hellinger-diameter of the channel:

$$\frac{1}{2}\mathrm{diam}_{\mathrm{H}^2}(P_{Y|X}) \le \eta_{\mathrm{KL}} \le g\left(\frac{1}{2}\mathrm{diam}_{\mathrm{H}^2}(P_{Y|X})\right) \le \mathrm{diam}_{\mathrm{H}^2}(P_{Y|X}) \tag{2}$$

where $g(t) \triangleq 2t\left(1 - \frac{t}{2}\right)$, $\mathrm{diam}_{\mathrm{H}^2}(P_{Y|X}) = \sup_{x,x'} H^2(P_{Y|X=x}, P_{Y|X=x'})$ and $H^2(P,Q) = 2 - 2\int\sqrt{dPdQ}$.

Note that the only property of divergence that we have used in the proof of Theorem 1 is convexity of $(P,Q) \mapsto D(P,Q)$. This property is shared by all f-divergences, cf. [4]. In other words we proved:

Theorem 2: Let $\eta_f = \sup \frac{D_f(P_{Y|X} \circ P \| P_{Y|X} \circ Q)}{D_f(P\|Q)}$ optimized over all $P, Q \in \mathcal{P}(\mathcal{X})$ with $0 < D_f(P,Q) < \infty$. Then the optimization can be restricted to pairs P, Q supported on two common points in $\mathcal{X}$.

This fact was conjectured in [5, Open Problem 7.4].

There are two other noteworthy results that our technique entails. First, a moment of reflection confirms that we, in fact, have shown that the upper concave envelope of the set $\cup_{P_X,Q_X}\{(D_f(P_X\|Q_X), D_f(P_Y\|Q_Y))\}$ is unchanged if we restrict the union to pairs P_X, Q_X supported on two points.

Second, a similar argument holds for the post-SDPI coefficient of a channel [6], defined as

$$\eta_{KL}^{(p)}(P_{Y|X}) = \inf\{\eta : I(U;X) \le \eta I(U;Y), X{-}Y{-}U\}$$

Namely, we have that $\eta_{KL}^{(p)}$ can be computed by restricting X to take two values. Indeed, fix an arbitrary $P_{X,Y,U}$ s.t. $X - Y - U$. As shown in [2, Theorem 4] one can safely assume U to be binary. Now, consider a set $\mathcal{S}$ of all $\hat{P}_X$ such that the joint distribution $\hat{P}_{X,Y,U} = \hat{P}_X P_{Y|X} P_{U|Y}$ satisfies $\hat{P}_U = P_U$. Since U is binary, $\mathcal{S}$ is an intersection of a hyperplane with a simplex. Now, the function $\hat{P}_X \mapsto \hat{I}(U;X) - \lambda \hat{I}(U;Y)$ is linear in $\hat{P}_X$ over $\mathcal{S}$. Consequently, the maximum (and the minimum) of this function is attained at a binary $\hat{P}_X$.

References

[1] R. Ahlswede and P. Gács, "Spreading of sets in product spaces and hypercontraction of the markov operator," *The annals of probability*, pp. 925–939, 1976.

[2] Y. Polyanskiy and Y. Wu, "Strong data-processing inequalities for channels and Bayesian networks," in *Convexity and Concentration*. Springer, 2017, pp. 211–249.

[3] ——, "Application of the information-percolation method to reconstruction problems on graphs," *Mathematical Statistics and Learning*, vol. 2, no. 1, pp. 1–24, 2020.

[4] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observation," *studia scientiarum Mathematicarum Hungarica*, vol. 2, pp. 229–318, 1967.

[5] J. Cohen, J. H. Kempermann, and G. Zbaganu, *Comparisons of stochastic matrices with applications in information theory, statistics, economics and population*. Springer Science & Business Media, 1998.

[6] Y. Polyanskiy, "Post-SDPI and distributed estimation", *Lecture 5, Information-Theoretic Methods in Statistics and Computer Science*, EPFL, 2019. http://people.lids.mit.edu/yp/homepage/data/LN_sdpi3.pdf

Or Ordentlich (Member, IEEE) is an associate professor in the School of Computer Science and Engineering at the Hebrew University of Jerusalem. He received the B.Sc. (cum laude), M.Sc. (summa cum laude), and Ph.D. degrees from Tel Aviv University, Israel, in 2010, in 2011, and 2016, respectively, all in electrical engineering. During the years 2015-2017 he was a postdoctoral fellow in the Laboratory for Information and Decision Systems at the Massachusetts Institute of Technology (MIT), and in the Department of Electrical and Computer Engineering at Boston University.

Yury Polyanskiy (S'08-M'10-SM'14) is an Associate Professor of Electrical Engineering and Computer Science and a member of IDSS and LIDS at MIT. Yury received M.S. degree in applied mathematics and physics from the Moscow Institute of Physics and Technology, Moscow, Russia in 2005 and Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ in 2010. His research interests span information theory, statistical learning, error-correcting codes, wireless communication and fault tolerance. Dr. Polyanskiy won the 2020 IEEE Information Theory Society James Massey Award, 2013 NSF CAREER award and 2011 IEEE Information Theory Society Paper Award.