Scribe: Pierre Quinton

1 Strong Data Processing Inequality and Distributed Estimation

1.1 More on Strong Data Processing Inequalities

Proposition 1 (Tensorisation). *For a given number n, two measures P_X and $P_{Y|X}$, the following tensorisation holds*

$$\eta_{\mathrm{KL}}(P_X^{\otimes n}, P_{Y|X}^{\otimes n}) = \eta_{\mathrm{KL}}(P_X, P_{Y|X})$$

In particular, if $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P_{X,Y}$, then $\forall P_{U|X^n}$ then

$$I(U; Y^n) \leq \eta_{\mathrm{KL}}(P_X, P_{Y|X}) \times I(U; X^n)$$

Proof. Without loss of generality (by induction) it is sufficient to prove the proposition for $n = 2$. It is always useful to keep in mind the following diagram

$$\begin{tikzcd}
 & X_1 \arrow[r] & Y_1 \\
U \arrow[ur] \arrow[dr] & & \\
 & X_2 \arrow[r] & Y_2
\end{tikzcd}$$

Let $\eta = \eta_{\mathrm{KL}}(P_X, P_{Y|X})$

$$\begin{aligned}
I(U; Y_1, Y_2) &= I(U; Y_1) + I(U; Y_2|Y_1) \\
&\leq \eta \left[I(U; X_1) + I(U; X_2|Y_1) \right] && (1.1) \\
&= \eta \left[I(U; X_1) + I(U; X_2|X_1) + I(U; X_1|Y_1) - I(U; X_1|Y_1, X_2) \right] && (1.2) \\
&\leq \eta \left[I(U; X_1) + I(U; X_2|Y_1) \right] && (1.3) \\
&= \eta I(U; X_1, X_2)
\end{aligned}$$

Where 1.1 is due to the fact that conditioned on Y_1, $U - X_2 - Y_2$ is still a Markov chain, 1.2 is because $U - X_1 - Y_1$ is a Markov chain and 1.3 follows from the fact that $X_2 - U - X_1$ is a Markov chain even when condition Y_1. $\square$

This tensorisation property can be used for correlation estimation. Suppose Alice have samples $\{X_i\}_{i\geq 1} \overset{\text{i.i.d.}}{\sim} B(1/2)$ and Bob have samples $\{Y_i\}_{i\geq 1} \overset{\text{i.i.d.}}{\sim} B(1/2)$ such that the (X_i, Y_i) are i.i.d. with $\mathbb{E}[X_i Y_i] = \rho \in [-1, 1]$. The goal is for Bob to send W to Alice with $H(W) = B$ bits and for Alice to estimate $\hat{\rho} = \hat{\rho}(X^\infty, W)$ with objective

$$R^*(B) = \inf_{W, \hat{\rho}} \sup_{\rho} \mathbb{E}[(\rho - \hat{\rho})^2]$$

Notice that if Bob sends $W = (Y_1, \ldots, Y_B)$ then the optimal estimator is $\hat{\rho}(X^\infty, W) = \frac{1}{n} \sum_{i=1}^{B} X_i Y_i$ which has error $\frac{1}{B}$, hence $R^*(B) \leq \frac{1}{B}$.

Theorem 1. *The optimal rate when $B \to \infty$ is given by*

$$R^*(X^\infty, W) = \frac{1 + o(1)}{2 \ln 2} \cdot \frac{1}{B}$$

Proof. Fix $P_{W|Y^\infty}$, we get the following decomposition

$$\begin{tikzcd}
 & X_1 \arrow[r] & Y_1 \\
 & \vdots & \vdots \\
W \arrow[uur] \arrow[r] & X_i \arrow[r] & Y_i \\
 & \vdots & \vdots
\end{tikzcd}$$

Note that once the messages W are fixed we have a parameter estimation problem $\{Q_\rho, \rho \in [-1,1]\}$ where Q_ρ is a distribution of (X^∞, W) when A^∞, B^∞ are ρ-correlated. Since we minimize MMSE, we know from the Bayesian Cramer-Rao lower bound (van Trees inequality)[1] that $R^*(B) \geq \frac{1+o(1)}{\min_\rho I_F(\rho)} \geq \frac{1+o(1)}{I_F(0)}$ where $I_F(\rho)$ is the Fisher Information of the family $\{Q_\rho\}$.

Recall, that we also know from the local approximation that

$$D(Q_\rho \| Q_0) = \frac{\rho^2}{2\ln(2)} I_F(0) + o(\rho^2)$$

Furthermore, notice that under $\rho = 0$ we have X^∞ and W independent and thus

$$\begin{aligned}
D(Q_\rho \| Q_0) &= D(P^\rho_{X^\infty,W} \| P^0_{X^\infty,W}) \\
&= D(P^\rho_{X^\infty,W} \| P^\rho_{X^\infty} \times P^\rho_W) \\
&= I(W; X^\infty) \\
&\leq \rho^2 I(W; Y^\infty) \\
&\leq \rho^2 B
\end{aligned}$$

hence $I_F(0) \leq 2\ln 2B + o(1)$ which in turns implies the theorem. For full details, the upper bound and the extension to interactive communication between Alice and Bob see [Had+19]. $\square$

1.2 Post Strong Data Processing Inequality (Post-SDPI)

Definition 1. Given a conditional measure $P_{Y|X}$, define

$$\begin{aligned}
\eta^{(p)}_{\mathrm{KL}}(P_{Y|X}) &= \sup_{P_X, P_{U|Y}} \left\{ \frac{I(U;X)}{I(U;Y)} : X \to Y \to U \right\} \\
&= \sup_{P_X} \eta_{\mathrm{KL}}(P_Y, P_{X|Y})
\end{aligned}$$

where $P_Y(\cdot) = P_X \times P_{Y|X}(\mathcal{X}, \cdot)$.

It is easy to see that by the data processing inequality, $\eta^{(p)}_{\mathrm{KL}}(P_{Y|X}) \leq 1$. This bound can be achieved with equality in some non trivial cases, in example let $P_Y|X = \mathrm{BEC}_\tau$ and $X \to Y \to U$ be given by

$$\begin{tikzcd}
 & X & Y & U \\
\bar{\varepsilon} & 0 \arrow[r, "\bar{\tau}"] \arrow[dr, "\tau"] & 0 \arrow[r] & 0 \\
 & & ? \arrow[ur] & \\
\varepsilon & 1 \arrow[r, "\bar{\tau}"] \arrow[ur, "\tau"'] & 1 \arrow[r] & 1
\end{tikzcd}$$

[1] This requires some technical justification about smoothness of fisher information $I_F(\rho)$.

Then we can compute $I(Y;U) = H(U) = h(\varepsilon\bar{\tau})$ and $I(X;U) = H(U) - H(U|X) = h(\varepsilon\bar{\tau}) - \varepsilon h(\tau)$ hence

$$\begin{aligned}\eta_{\mathrm{KL}}^{(p)}(P_{Y|X}) &\geq \frac{I(X;U)}{I(Y;U)} \\ &= 1 - h(\tau)\frac{\varepsilon}{h(\varepsilon\bar{\tau})}\end{aligned}$$

This last term tends to 1 when ε tends to 0 hence

$$\eta_{\mathrm{KL}}^{(p)}(\mathrm{BEC}_\tau) = 1$$

even though Y is not a one to one function of X.

The second bad news is that by taking $\varepsilon = \frac{1}{2}$, we have that $\eta_{\mathrm{KL}}^{(p)}(\mathrm{Unif}, \mathrm{BEC}_\tau) > 1 - \tau$ for $\tau \to 1$. Thus, the natural conjecture that for any BMS we should have $\eta_{\mathrm{KL}}^{(p)}(\mathrm{Unif}, \mathrm{BMS}) = \eta_{\mathrm{KL}}(\mathrm{BMS})$ is *incorrect*.

Nevertheless, the post-SDPI constant is often non-trivial, most importantly for the BSC:

Theorem 2.

$$\eta_{\mathrm{KL}}^{(p)}(\mathrm{BSC}_\delta) = (1-2\delta)^2$$

to prove the theorem, the following lemma is of help.

Lemma 1. *If for any X and Y in $\{0,1\}$ we have*

$$p_{X,Y}(x,y) = f(x)\left(\frac{\delta}{1-\delta}\right)^{1(x\neq y)} g(Y)$$

for some functions f and g, then $\eta_{\mathrm{KL}}(P_{Y|X}) \leq (1-2\delta)^2$

Proof. It is known that for binary input chanels $P_{Y|X}$ [PW17].

$$\eta_{\mathrm{KL}}(P_{Y|X}) \leq H^2(P_{Y|X=0}\|P_{Y|X=1}) - \frac{H^4(P_{Y|X=0}\|P_{Y|X=1})}{4}$$

If we let $\phi = \frac{g(0)}{g(1)}$, then we have $p_{Y|X=0} = B\left(\frac{\lambda}{\phi+\lambda}\right)$ and $p_{Y|X=1} = B\left(\frac{1}{1+\phi\lambda}\right)$ and a simple check shows that

$$\begin{aligned}\max_\phi H^2(P_{Y|X=0}\|P_{Y|X=1}) - \frac{H^4(P_{Y|X=0}\|P_{Y|X=1})}{4} &\overset{\phi=1}{=} H^2_{\phi=1}(P_{Y|X=0}\|P_{Y|X=1}) - \frac{H^4_{\phi=1}(P_{Y|X=0}\|P_{Y|X=1})}{4} \\ &= (1-2\delta)^2\end{aligned}$$

Now observe that $P_{X,Y}$ in Theorem 2 satisfies the property of the lemma with X and Y exchanged, hence $\eta_{\mathrm{KL}}(P_Y, P_{X|Y}) \leq (1-2\delta)^2$ which implies that $\eta_{\mathrm{KL}}^{(p)}(P_{Y|X}) = \sup_{P_X} \eta_{\mathrm{KL}}(P_Y, P_{X|Y}) \leq (1-2\delta)^2$ with equality if P_X is uniform. $\square$

Theorem 3. *Let $P_{Y|X} = \mathrm{BMS}$, then for any $X \to Y \to U$*

$$I(X;U) \leq \eta_{\mathrm{KL}}(P_{Y|X}) \cdot \log|\mathcal{U}|$$

where recall that $\eta_{\mathrm{KL}}(P_{Y|X}) = I_{\chi^2}(X;Y)$ when $X \sim \mathrm{Bern}(1/2)$.

Proof. Every BMS channel is a mixture of BSCs, it can be represented as follow, let $(0,1) \ni \Delta \sim P_\Delta$ independently of X, $P_{\tilde{Y}|X,\Delta} = \mathrm{BSC}_\Delta$ and let $Y = (\tilde{Y}, \Delta)$. Then

$$\begin{aligned} I(X;U) &\le I(X;U,\Delta) \\ &= I(X;U|\Delta) \\ &= \mathbb{E}_\Delta[I(X;U|\Delta=\Delta)] \\ &\le \mathbb{E}_\Delta[(1-2\Delta^2) I(Y;U|\Delta=\Delta)] \\ &\le \mathbb{E}_\Delta[(1-2\Delta^2)\log|\mathcal{U}|] \\ &= \eta_{\mathrm{KL}}(P_{Y|X}) \cdot \log|\mathcal{U}| \end{aligned}$$

□

1.3 Distributed Mean Estimation

We want to estimate $\theta \in [-1,1]^d$ and we have m machines observing $X_i = \theta + \sigma Z_i$ where $Z_i \sim \mathcal{N}(0, I_d)$ independently. They can send a total of B bits to a remote estimator. The goal of the estimator is to minimize $\sup_\theta \mathbb{E}[\|\theta - \hat{\theta}\|^2]$ over $\hat{\theta}$. If we denote by $U_i \in \mathcal{U}_i$ the messages then $\sum_i |\mathcal{U}_i| \le B$ then the diagram is

$$\begin{tikzcd} & X_1 \arrow[r] & U_1 \arrow[dr] & \\ \theta \arrow[ur] \arrow[dr] & \vdots & \vdots & \hat{\theta} \\ & X_m \arrow[r] & U_m \arrow[ur] & \end{tikzcd}$$

Finally, let

$$R^*(m,d,\sigma^2,B) = \inf_{U_1,\ldots,U_m,\hat{\theta}} \sup_\theta \mathbb{E}[\|\theta - \hat{\theta}\|^2]$$

Observations:

- Without constraint on the magnitude of $\theta \in [-1,1]^d$, we could give $\theta \sim \mathcal{N}(0, bI_d)$ and from rate-distortion quickly conclude that estimating θ within risk R requires communicating at least $\frac{d}{2}\log\frac{bd}{R}$ bits, which diverges as $b \to \infty$. Thus, restricting the magnitude of θ is necessary in order to be able to estimate it with finitely many bits communicated.
- It is easy to esstablish that $R^*(m,d,\sigma^2,\infty) = \mathbb{E}\left[\left\|\frac{\sigma}{m}\sum_i Z_i\right\|^2\right] = \frac{d\sigma^2}{m}$ by taking $U_i = X_i$ and $\hat{\theta} = \frac{1}{m}\sum_i U_i$.
- In order to approach the risk of order $\frac{d\sigma^2}{m}$ we could do the following. Let $U_i = \mathrm{sign}(X_i)$ (coordinate-wise sign). This yields $B = md$ and it is easy to show that the achievable risk is $O(\frac{d\sigma^2}{m})$.
- Our main result is that this is optimal. This simplifies the proofs (in the non-interactive case) of [Duc+14]; [Bra+16].
- We want to point out, however, that all of these results (again in the non-interactive case, but with essentially sharp constants) are contained in the long line of work in the information theoretic literature on the so-called *Gaussian CEO problem.* We recommend consulting [EG19]. In particular, Theorem 3 there implies the $B \gtrsim dm$ lower bound. The Gaussian CEO work uses a lot more sophisticated machinery (the entropy power inequality and related results). The advantage of our SDPI proof is simplicity.

Theorem 4. *There exists a $c_1, c_2 > 0$ such that for all m, d, σ^2 if $R^*(m, d, \sigma^2, B) \le c_1 \frac{\sigma^2 d}{m}$ then $B \ge c_2 dm$.*

Proof. for $d = 1$, if we have $\hat{\theta}$ with risk $\mathbb{E}[(\theta - \hat{\theta})^2] \le c \cdot \frac{\sigma^2}{m}$ for all $\theta \in [-1, 1]$ then picking $\theta \sim \mathcal{U}(\{-\varepsilon, \varepsilon\})$ we get that if $\varepsilon \gtrsim \sqrt{\frac{\sigma^2}{m}}$ then $I(\theta; \hat{\theta}) \gtrsim 1$, now

$$I(\theta; \hat{\theta}) \le I(\theta; U^m) \le \sum_{i=1}^{m} I(\theta; U_i) \le \sum_{i=1}^{m} \frac{\varepsilon^2}{\sigma^2} \log |\mathcal{U}_i|$$

since $\eta_{\mathrm{KL}}(P_{X_i|\theta}) = \frac{\varepsilon^2}{\sigma^2}$. Hence $I(\theta; \hat{\theta}) \le \frac{\varepsilon^2}{\sigma^2} \cdot B$. Hence if $\varepsilon \lesssim \sqrt{c \cdot \frac{\sigma^2}{m}}$ then $B \gtrsim m$ $\square$

This proof does not extend to the d-dimensional case because the variant of the post-SDPI in Theorem 3 does not tensorize. So we need a more refined version.

Lemma 2 (restricted post-SDPI for the BIAWGN channel)**.** *if $X = \pm 1$ uniformly and $Y = \varepsilon X + Z$ with $Z \sim \mathcal{N}(0, 1)$. Then for all $c > 0$, there exist $c' > 0$ such that for all $\varepsilon \le \varepsilon_0(c)$, and all $P_{U|Y}$ we have*[2]

$$I(U; X) \ge c \cdot \varepsilon^2 \Rightarrow I(U; Y) \ge c'$$

where $X \to Y \to U$.

Furthermore, we have tensorization: Let $X^d = (\pm 1)^d$ uniformly and $Y^d = \varepsilon X^d + Z^d$ with $Z^d \sim \mathcal{N}(0, I_d)$. Then for all $c > 0$, there exist $c' > 0$ such that for all $\varepsilon \le \varepsilon_0(c)$, and all $P_{U|Y^d}$ we have

$$I(U; Y^d) \ge cd\varepsilon^2 \implies I(U; X^d) \ge c' d \,.$$

where $X^d \to Y^d \to U$.

Proof. The proof of the first part is as follows. Let us represent the channel as a mixture of BSC with the output $(\tilde{Y}, \Delta)$, $\tilde{Y} = \mathrm{BSC}_\Delta(X)$. The relation between Y and Δ is $\Delta(Y) = \frac{1}{1+e^{2|Y|\epsilon}}$. Thus, we have

$$(1 - 2\Delta)^2 = f^2(\epsilon Y), f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \tanh(x) \,.$$

Note that $|f(x)| \le x$. Fix $\eta_1 = c_1 \epsilon^2$ with $c_1 > 1$ to be specified. We have

$$\begin{aligned}
\mathbb{E}[(1-2\Delta)^2 1\{(1-2\Delta)^2 > \eta_1\}] &= \mathbb{E}[f^2(\epsilon Y) 1\{f^2(\epsilon Y) > \eta_1\}] \\
&\le \epsilon^2 \mathbb{E}[Y^2 1\{f^2(\epsilon Y) > \eta_1\} \\
&\le \epsilon^2 \mathbb{E}[Y^2 1\{Y^2 > c_1\}] \\
&\le \epsilon^2 \mathbb{E}[2(\epsilon^2 + Z^2) 1\{\epsilon^2 + Z^2 > c_1/2\} \\
&\le 2\epsilon^2 (\epsilon^2 + \mathbb{E}[Z^2 1\{Z^2 > c_1/4\}] \quad \text{assuming } \epsilon^2 < c_1/4 \\
&\le \frac{c\epsilon^2}{4 \log 2} \,, \qquad (1.4)
\end{aligned}$$

where in the last step we selected c_1 so large and ϵ_0 so small that $\epsilon^2 + \mathbb{E}[Z^2 1\{Z^2 > c_1/4\}] < c_1/(8 \log 2)$.

Now let $F = 1\{(1-2\Delta)^2 > \eta_1\}$. We have

$$\begin{aligned}
I(X; U) &\le I(X; U, F) = I(X; U|F) \\
&= \mathbb{P}[F = 0] I(X; U|F = 0) + \mathbb{P}[F = 1] I(X; U|F = 1) \\
&\le \eta_1 \mathbb{P}[F = 0] I(Y; U|F = 0) + \mathbb{P}[F = 1] I(X; Y|F = 1) \qquad (1.5) \\
&\le \eta_1 I(Y; U|F) + \log 2 \mathbb{E}[(1-2\Delta)^2 1\{F = 1\}] \qquad (1.6) \\
&\le \eta_1 I(Y; U) + \frac{c\epsilon^2}{4} \qquad (1.7)
\end{aligned}$$

[2]Note: If we had $\eta_{KL}^{(p)}(\mathrm{BIAWGN}_\epsilon) \le c_1 \epsilon^2$, then the statement would follow with $c' = \frac{c}{c_1}$. However, we do not yet know what is $\eta_{KL}^{(p)}$ for the BIAWGN.

where in (1.5) we applied BSC Post-SDPI conditioned on $\Delta = \delta$ and noted that under $F = 0$ the $\eta_{KL}^{(post)} \le \eta_1$, in (1.6) we applied Theorem 3, and in (1.7) we noted that $I(Y, F; U) = I(Y; U)$ and invoked our estimate (1.4). In all we see from (1.7) that if $I(X; U) > c\epsilon^2$ then $I(U; Y) \ge \frac{3}{4}\frac{c\epsilon^2}{\eta_1} \triangleq c'$, as required.

To prove the second part, consider an expansion

$$c\varepsilon^2 d \le I(U; X^d) = \sum_{i=1}^{d} I(U; X_i | X^{i-1}) \,.$$

Note that $I(U; X_i|X^{i-1}) \le I(X_i; Y_i) \le \frac{1}{2}\log(1+\varepsilon^2) \le \frac{\varepsilon^2}{2}$. Hence, we must have

$$|\{i : I(U; X_i|X^{i-1}) \ge c\varepsilon^2/2\}| \ge cd \,.$$

Now for every such i we can apply the first part of the lemma which guarantees then $I(U; Y_i|X^{i-1}) \ge c'$. Thus, also $I(U; Y_i|Y^{i-1}) \ge c'$. And hence, we should have

$$I(U; Y^d) = \sum_i I(U; Y_i|Y^{i-1}) \ge c' \cdot (cd) \,.$$

□

General d proof. To see how Lemma implies the result, let again $\theta \sim \mathcal{U}(\{-\varepsilon, \varepsilon\}^d)$ with $\varepsilon = c\sqrt{\frac{\sigma^2}{m}}$ for some fixed (sufficiently small) $c > 0$. Then the estimator $\hat\theta$ with risk $\le c_1 \frac{md}{\sigma^2}$, where $c_1 = c_1(c)$ also sufficiently small, can be converted into an estimator of θ within expected Hamming distance $\le d/2$. This in turn implies $I(\theta; \hat\theta) \ge c_3 d$.

Now notice that $I(\theta; U_i) \le I(\theta; X_i) = dI(\theta_1; X_{i,1})$. Note that the $\theta_1 \mapsto X_{i,1}$ is a BIAWGN_ϵ channel with $\epsilon = \frac{\varepsilon}{\sigma}$. So we have $I(\theta_1; X_{i,1}) \le \frac{1}{2}\log(1+\varepsilon^2/\sigma^2) \le \frac{\varepsilon^2}{2\sigma^2} = \frac{c^2}{2m}$. So we have $I(\theta; U_i) \le \frac{c^2 d}{2m}$. But the total sum $\sum_i I(\theta; U_i) \ge c_3 d$. Therefore, we should have This implies that for some $c_4 > 0$ we must have

$$|\{i \in [m] : I(\theta; U_i) > c_3 \frac{d}{2m}\}| \ge c_4 m$$

For each such i, we apply the second part of Lemma to get $I(X_i; U_i) > c_3' d$ which implies

$$\sum_i \log |\mathcal{U}_i| \ge \sum_i I(X_i; U_i) \ge (c_3' d) \cdot (c_4 m) \asymp dm \,.$$

□

References

[Bra+16] Mark Braverman, Ankit Garg, Tengyu Ma, Huy L Nguyen, and David P Woodruff. "Communication lower bounds for statistical estimation problems via a distributed data processing inequality". In: *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. ACM. 2016, pp. 1011–1020.

[Duc+14] John C Duchi, Michael I Jordan, Martin J Wainwright, and Yuchen Zhang. "Optimality guarantees for distributed statistical estimation". In: *arXiv preprint arXiv:1405.0782* (2014).

[EG19] Krishnan Eswaran and Michael Gastpar. "Remote source coding under Gaussian noise: Dueling roles of power and entropy power". In: *IEEE Transactions on Information Theory* (2019).

[Had+19] Uri Hadar, Jingbo Liu, Yury Polyanskiy, and Ofer Shayevitz. "Communication complexity of estimating correlations". In: *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. ACM. 2019, pp. 792–803.

[PW17] Yury Polyanskiy and Yihong Wu. "Strong Data-Processing Inequalities for Channels and Bayesian Networks". In: *Convexity and Concentration*. Ed. by Eric Carlen, Mokshay Madiman, and Elisabeth M. Werner. New York, NY: Springer New York, 2017, pp. 211–249. ISBN: 978-1-4939-7005-6.