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

\[
\eta_{KL}(P_X^n, P_{Y|X}^n) = \eta_{KL}(P_X, P_{Y|X})
\]

In particular, if \((X_i, Y_i)\) i.i.d. \(\sim P_{X,Y} \), then for all \(P_U|X^n\) then

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

**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

\[
\begin{array}{c}
U \\
\downarrow \\
X_1 & \rightarrow & Y_1 \\
& \searrow & \\
X_2 & \rightarrow & Y_2
\end{array}
\]

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

\[
I(U; Y_1, Y_2) = I(U; Y_1) + I(U; Y_2|Y_1)
\]

\[
\leq \eta \left[ I(U; X_1) + I(U; X_2|Y_1) \right]
\]

\[
= \eta \left[ I(U; X_1) + I(U; X_2|X_1) + I(U; X_1|Y_1) - I(U; X_1|Y_1, X_2) \right]
\]

\[
\leq \eta \left[ I(U; X_1) + I(U; X_2|Y_1) \right]
\]

\[
= \eta I(U; X_1, X_2)
\]

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 \( Y_1 \).

This tensorisation property can be used for correlation estimation. Suppose Alice have samples \( \{X_i\}_{i \geq 1} \sim B(1/2) \) and Bob have samples \( \{Y_i\}_{i \geq 1} \sim B(1/2) \) 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

\[ X_1 \rightarrow Y_1 \\
\vdots \\
W \rightarrow X_i \rightarrow Y_i \\
\vdots \\
\]

Note that once the messages $W$ are fixed we have a parameter estimation problem $\{Q_\rho, \rho \in [-1,1]\}$ where $Q_\rho$ is a distribution of $(X^\infty, W)$ when $A^\infty, B^\infty$ are $\rho$-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 $X^\infty$ and $W$ independent and thus

\[ D(Q_\rho \| Q_0) = D(P_{X^\infty,W}^\rho \| P_{X^\infty}^\rho) \\
= D(P_{X^\infty,W}^\rho \| P_{X^\infty}^\rho \times P_W^\rho) \\
= I(W; X^\infty) \\
\leq \rho^2 I(W; Y^\infty) \\
\leq \rho^2 B \]

hence $I_F(0) \leq 2 \ln 2 + o(1)$ 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

\[ \eta_{KL}^{(p)}(P_{Y|X}) = \sup_{P_X,P_U|Y} \left\{ \frac{I(U;X)}{I(U;Y)} : X \rightarrow Y \rightarrow U \right\} \\
= \sup_{P_X} \eta_{KL}(P_Y, P_{X|Y}) \]

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

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

\[ X \rightarrow Y \rightarrow U \\
\bar{\epsilon} \rightarrow 0 \rightarrow 0 \rightarrow 0 \\
\epsilon \rightarrow ? \rightarrow 1 \rightarrow 0 \\
\bar{\epsilon} \rightarrow 1 \rightarrow 1 \rightarrow 1 \]

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

2
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

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

This last term tends to 1 when $\varepsilon$ tends to 0 hence

$$
\eta_{KL}^{(p)}(\text{BEC}_\tau) = 1
$$
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 $\eta_{KL}^{(p)}(\text{Unif}, \text{BEC}_\tau) > 1 - \tau$ for $\tau \rightarrow 1$.
Thus, the natural conjecture that for any BMS we should have $\eta_{KL}^{(p)}(\text{Unif}, \text{BMS}) = \eta_{KL}(\text{BMS})$ is incorrect.

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

**Theorem 2.**

$$
\eta_{KL}^{(p)}(\text{BSC}_\delta) = (1 - 2\delta)^2
$$
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 $\eta_{KL}(P_{Y|X}) \leq (1 - 2\delta)^2$

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

$$
\eta_{KL}(P_{Y|X}) \leq H^2(P_{Y|X=0}||P_{Y|X=1}) - \frac{H^4(P_{Y|X=0}||P_{Y|X=1})}{4}
$$
If we let $\phi = \frac{p_{Y|X=0}(y)}{p_{Y|X=1}(y)}$, then we have $p_{Y|X=0} = B \left( \frac{\lambda}{\phi + \lambda} \right)$ and $p_{Y|X=1} = B \left( \frac{1}{1+\phi} \right)$ and a simple check shows that

$$
\max_{\phi} H^2(P_{Y|X=0}||P_{Y|X=1}) - \frac{H^4(P_{Y|X=0}||P_{Y|X=1})}{4} = \frac{H^4(\phi = 1(P_{Y|X=0}||P_{Y|X=1})}{4} - \frac{H^4(\phi = 1(P_{Y|X=0}||P_{Y|X=1})}{4} = (1 - 2\delta)^2
$$
Now observe that $P_{X,Y}$ in Theorem 2 satisfies the property of the lemma with $X$ and $Y$ exchanged, hence $\eta_{KL}(P_Y, P_{X|Y}) \leq (1 - 2\delta)^2$ which implies that $\eta_{KL}^{(p)}(P_{Y|X}) = \sup_{P_X} \eta_{KL}(P_Y, P_{X|Y}) \leq (1 - 2\delta)^2$ with equality if $P_X$ is uniform. \qed

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

$$
I(X;U) \leq \eta_{KL}(P_{Y|X}) \cdot \log |U|
$$
where recall that $\eta_{KL}(P_{Y|X}) = \chi^2(X;Y)$ when $X \sim \text{Bern}(1/2)$. 

3
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 \(X\), \(P_{Y|X,\Delta} = \text{BSC}_\Delta\) and let \(Y = (\bar{Y}, \Delta)\). Then
\[
I(X; U) \leq I(X; U, \Delta) = I(X; U|\Delta = \Delta) = \mathbb{E}_\Delta[I(X; U|\Delta = \Delta)] \\
\leq \mathbb{E}_\Delta[(1 - 2\Delta^2)I(Y; U|\Delta = \Delta)] \\
\leq \mathbb{E}_\Delta[(1 - 2\Delta^2) \log |U|] = \eta_{\text{KL}}(P_{Y|X}) \cdot \log |U|
\]

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 \(\sum_i |U_i| \leq B\) then the diagram is

\[\begin{array}{ccc}
X_1 & \rightarrow & U_1 \\
\theta \left\downarrow \right. & & \left\uparrow \right. \theta \\
\vdots \rightarrow & \vdots \leftarrow & \vdots \\
X_m & \rightarrow & U_m
\end{array}\]

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

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 \(\theta\) 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 \(\theta\) is necessary in order to be able to estimate it with finitely many bits communicated.

- It is easy to establish that \(R^*(m, d, \sigma^2, \infty) = \mathbb{E}\left[\left\|\frac{1}{m} \sum_i Z_i\right\|^2\right] = \frac{d\sigma^2}{m}\) 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 \(U_i = \text{sign}(X_i)\) (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 $m, d, \sigma^2$ if $R^*(m, d, \sigma^2, B) \leq c_1 \frac{\sigma^2 d}{m}$ then $B \geq c_2 dm$.

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

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

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

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 $\varepsilon \leq \varepsilon_0(c)$, and all $P_{U|X}$ we have

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

where $U \rightarrow X \rightarrow Y$.

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 $\varepsilon \leq \varepsilon_0(c)$, and all $P_{U|X^d}$ we have

$$I(U; Y^d) \geq c d \varepsilon^2 \Rightarrow I(U; X^d) \geq c' d.$$

Proof. The proof of the first part is ... TODO: add.

To prove the second part, consider an expansion

$$c \varepsilon^2 d \leq I(U; Y^d) = \sum_{i=1}^{d} I(U; Y_i | Y^{i-1}) .$$

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

$$|\{i : I(U; Y_i | Y^{i-1}) \geq c \varepsilon^2 / 2\}| \geq c d .$$

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

$$I(U; X^d) = \sum_{i} I(U; X_i | X^{i-1}) \geq c' \cdot (cd) .$$

**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 $\leq c_1 \frac{md}{\sigma^2}$, where $c_1 = c_1(c)$ also sufficiently small, can be converted into an estimator of $\theta$ within expected Hamming distance $\leq d/2$. This in turn implies $I(\theta; \hat{\theta}) \geq c_2 d$.

Now notice that $I(\theta; U_i) \leq I(\theta; X_i) = d I(\theta_1 ; X_{i,1})$. Note that the $\theta_1 \mapsto X_{i,1}$ is a BIAWGN channel with $\varepsilon = \frac{\varepsilon}{\sqrt{2}}$. So we have $I(\theta_1; X_{i,1}) \leq \frac{1}{2} \log(1 + (\varepsilon^2 / \sigma^2) \leq \frac{\varepsilon^2}{2 \sigma^2} = \frac{\varepsilon^2}{2m}$. So we have $I(\theta; U_i) \leq \frac{\varepsilon^2 d}{2m}$.

2Note: If we had $\eta_{KL}(\text{BIAWGN}) \leq c_1 \varepsilon^2$, then the statement would follow with $c' = \frac{\varepsilon}{c_1}$. However, we do not yet know what is $\eta_{KL}$ for the BIAWGN.
But the total sum $\sum_i I(\theta; U_i) \geq c_3d$. Therefore, we should have This implies that for some $c_4 > 0$ we must have

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

For each such $i$, we apply the second part of Lemma to get $I(X_i; U_i) > c'_3d$ which implies

$$\sum_i \log |U_i| \geq \sum_i I(X_i; U_i) \geq (c'_3d) \cdot (c_4 m) \approx dm.$$

\[\square\]

References


