Communication Complexity Notes

Amit Chakrabarti

February 10, 2019

1 Recap of Information Theory Basics

For the purposes of these notes, we shall only consider probability distributions on finite sets. Thus, a distribution on a set of size n may be thought of as an n-dimensional vector with nonnegative entries and ℓ_1-norm 1.

1.1 Setup and Basic Definitions

Let $\mathbf{p}$ and $\mathbf{q}$ be distributions on finite sets $\mathcal{X}$ and $\mathcal{Y}$ respectively. Let $X \sim \mathbf{p}$ and $Y \sim \mathbf{q}$ be random variables such that $XY \sim \mathbf{r}$, where $\mathbf{r}$ is some joint distribution on $\mathcal{X} \times \mathcal{Y}$. Note that we're writing XY instead of (X, Y)—this is a useful shorthand.

The distributions $\mathbf{p}$ and $\mathbf{q}$ are *marginals* of $\mathbf{r}$ and we have

$$p_x = \sum_{y \in \mathcal{Y}} r_{xy}\,, \quad q_y = \sum_{x \in \mathcal{X}} r_{xy}\,. \tag{1}$$

Further, for the conditional distributions of X given a particular realization $Y = y$ (or vice versa), we have

$$\Pr[X = x \mid Y = y] = \frac{r_{xy}}{q_y}\,, \quad \Pr[Y = y \mid X = x] = \frac{r_{xy}}{p_x}\,. \tag{2}$$

We define

$$\text{Entropy:} \quad \mathrm{H}(X) = \mathrm{H}(\mathbf{p}) = \sum_{x \in \mathcal{X}} p_x \log \frac{1}{p_x}\,; \tag{3}$$

$$\text{Conditional entropy:} \quad \mathrm{H}(X \mid Y) = \mathbb{E}_y\, \mathrm{H}(X \mid Y = y) = \sum_{y \in \mathcal{Y}} q_y\, \mathrm{H}(X \mid Y = y)\,; \tag{4}$$

$$\text{Mutual information:} \ \mathrm{I}(X : Y) = \mathrm{H}(X) - \mathrm{H}(X \mid Y)\,. \tag{5}$$

If Z is another random variable, with XYZ having some joint distribution, we define the conditional mutual information, $\mathrm{I}(X : Y \mid Z) = \mathbb{E}_z I(X : Y \mid Z = z) = \mathrm{H}(X \mid Z) - \mathrm{H}(X \mid YZ)$.

In the definitions above, we must interpret $0 \log 0 = 0$. Clearly, $\mathrm{H}(X) \geq 0$ with equality iff X is a constant; likewise, $\mathrm{H}(X \mid Y) \geq 0$ with equality iff Y fully determines X.

1.2 Closeness of Distributions

For our work, we shall need several notions of closeness or "distance" between probability distributions. There are several meaningful notions, some of which are *not* metrics.

Suppose that $\mathbf{p}$ and $\mathbf{q}$ are distributions on the *same* set $\mathcal{X}$. We make the following definitions.

$$\text{Total variation distance:} \qquad \mathrm{D_{TV}}(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_1 = \frac{1}{2}\sum_{x \in \mathcal{X}} |p_x - q_x| \tag{6}$$

$$\text{Kullback-Leibler divergence:} \qquad \mathrm{D_{KL}}(\mathbf{p} \,\|\, \mathbf{q}) = \sum_{x \in \mathcal{X}} p_x \log \frac{p_x}{q_x} \tag{7}$$

$$\text{Jensen-Shannon divergence:} \qquad \mathrm{D_{JS}}(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\left(\mathrm{D_{KL}}\left(\mathbf{p} \,\Big\|\, \frac{\mathbf{p} + \mathbf{q}}{2}\right) + \mathrm{D_{KL}}\left(\mathbf{q} \,\Big\|\, \frac{\mathbf{p} + \mathbf{q}}{2}\right)\right) \tag{8}$$

$$\begin{aligned} \text{Squared Hellinger distance:} \qquad \mathrm{h}^2(\mathbf{p}, \mathbf{q}) &= \frac{1}{2}\|\sqrt{\mathbf{p}} - \sqrt{\mathbf{q}}\|_2^2 = \frac{1}{2}\sum_{x \in \mathcal{X}} \left(\sqrt{p_x} - \sqrt{q_x}\right)^2 && (9) \\ &= 1 - \sum_{x \in \mathcal{X}} \sqrt{p_x q_x}\,. && (10) \end{aligned}$$

Total variation distance is also called *statistical distance*; KL divergence is also called *informational divergence* or *relative entropy*. In eq. (7), we interpret $0 \log \frac{0}{0} = 0$. If there exists $x \in \mathcal{X}$ such that $p_x \neq 0 = q_x$, then $\mathrm{D_{KL}}(\mathbf{p} \,\|\, \mathbf{q}) = \infty$. With the exception of KL divergence, all of the above "distances" are symmetric in $\mathbf{p}$ and $\mathbf{q}$. Obviously, total variation distance and Hellinger distance are both metrics (they arise from norms). It is a somewhat deep fact that the square root of JS divergence is also a metric; we won't need this.

Theorem 1.1. $\mathrm{D_{KL}}(\mathbf{p} \,\|\, \mathbf{q}) \geq 0$, *with equality iff* $\mathbf{p} = \mathbf{q}$.

Proof. The function $t \mapsto -\log t$ is strictly convex on $(0, \infty)$. By Jensen's inequality,

$$\mathrm{D_{KL}}(\mathbf{p} \,\|\, \mathbf{q}) = \sum_{x \in \mathcal{X}} p_x \left(-\log \frac{q_x}{p_x} \right) \geq -\log \left(\sum_{x \in \mathcal{X}} p_x \cdot \frac{q_x}{p_x} \right) = -\log 1 = 0\,,$$

with equality iff q_x / p_x is the same for each x, which requires $\mathbf{p} = \mathbf{q}$, since $\|\mathbf{p}\|_1 = \|\mathbf{q}\|_1 = 1$. □

1.3 Basic Properties of Entropy and Mutual Information

The nonnegativity of KL divergence has many important consequences.

Theorem 1.2. $\mathrm{H}(X) \leq \log |\mathcal{X}|$, *with equality iff* $X \sim \mathbf{u}$, *the uniform distribution on* $\mathcal{X}$.

Proof. Use Theorem 1.1 and the calculation

$$\mathrm{D_{KL}}(\mathbf{p} \,\|\, \mathbf{u}) = \sum_{x \in \mathcal{X}} p_x \log \frac{p_x}{1/|\mathcal{X}|} = \log |\mathcal{X}| - \mathrm{H}(X)\,.$$ □

We now return to the setup of Section 1.1, i.e., X and Y may be distributed on distinct sets, $X \sim \mathbf{p}$, $Y \sim \mathbf{q}$, $XY \sim \mathbf{r}$.

Theorem 1.3. $\mathrm{I}(X : Y) = \mathrm{D_{KL}}(\mathbf{r} \,\|\, \mathbf{p} \otimes \mathbf{q})$.

Proof. Combining eqs. (1) to (5),

$$\begin{aligned} \mathrm{I}(X : Y) = \mathrm{H}(X) - \mathrm{H}(X \mid Y) &= \sum_{x \in \mathcal{X}} \Big(\sum_{y \in \mathcal{Y}} r_{xy} \Big) \log \frac{1}{p_x} - \sum_{y \in \mathcal{Y}} q_y \sum_{x \in \mathcal{X}} \frac{r_{xy}}{q_y} \log \frac{q_y}{r_{xy}} \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} r_{xy} \log \frac{r_{xy}}{p_x q_y} = \mathrm{D_{KL}}(\mathbf{r} \,\|\, \mathbf{p} \otimes \mathbf{q})\,. \end{aligned}$$ □

Corollary 1.4. $\mathrm{I}(X : Y) = \mathbb{E}_y \mathrm{D_{KL}}(\mathbf{p}^{(y)} \,\|\, \mathbf{p})$, *where* $\mathbf{p}^{(y)}$ *is the conditional distribution of* X *given* $Y = y$.

Proof. By eqs. (1) and (2),

$$\mathbb{E}_y \mathrm{D_{KL}}(\mathbf{p}^{(y)} \,\|\, \mathbf{p}) = \sum_{y \in \mathcal{Y}} q_y \sum_{x \in \mathcal{X}} p_x^{(y)} \log \frac{p_x^{(y)}}{p_x} = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} r_{xy} \log \frac{r_{xy}}{q_y p_x} = \mathrm{I}(X : Y)\,.$$ □

Corollary 1.5. $\mathrm{I}(X : Y) = \mathrm{I}(Y : X) \geq 0$, *with equality iff* $X \perp Y$.

Proof. Use Theorem 1.1 and the observation that $\mathbf{r} = \mathbf{p} \otimes \mathbf{q}$ iff $X \perp Y$. □

Corollary 1.6. $\mathrm{H}(X \mid Y) \leq \mathrm{H}(X)$, *with equality iff* $X \perp Y$. □

The next theorem does not itself depend on nonnegativity of KL divergence but, together with the above, leads to further important results.

Theorem 1.7 (Chain rule for entropy)**.** $\mathrm{H}(XY) = \mathrm{H}(X) + \mathrm{H}(Y \mid X)$.

Proof. Direct calculation, similar to that in Theorem 1.3. □

Corollary 1.8 (Subadditivity of entropy)**.** $\mathrm{H}(XY) \le \mathrm{H}(X) + \mathrm{H}(Y)$. □

Theorem 1.7 and its corollary clearly extend to a joint distribution of more than two random variables: if $X_1, \ldots, X_t$ are random variables, then

$$\mathrm{H}(X_1 \ldots X_t) = \sum_{j=1}^{t} \mathrm{H}(X_j \mid X_1 \ldots X_{j-1}) \le \sum_{j=1}^{t} \mathrm{H}(X_j)\,. \tag{11}$$

Using eq. (5), we have some further corollaries.

Corollary 1.9 (Chain rule for mutual info)**.** $\mathrm{I}(X_1 \ldots X_t : Y) = \sum_{j=1}^{t} \mathrm{I}(X_j : Y \mid X_1 \ldots X_{j-1})$. □

Corollary 1.10. *If $X_1, \ldots, X_t$ are mutually independent, then* $\mathrm{I}(X_1 \ldots X_t : Y) \ge \sum_{j=1}^{t} \mathrm{I}(X_j : Y)$. □

We record an important special case of mutual information, where one of the variables involved is a uniformly random bit.

Lemma 1.11. *Let A and X be random variables such that $A \in_R \{0,1\}$. For each $i \in \{0,1\}$, let $\mathbf{p}^i$ be the distribution of X conditioned on $A = i$. Then*

$$\mathrm{I}(A : X) = \mathrm{D}_{\mathrm{JS}}(\mathbf{p}^0, \mathbf{p}^1)\,.$$

Proof. Direct consequence of Corollary 1.4 and the definition of JS divergence. □

1.4 Relations Between Distance Measures

We return to the setting of distributions $\mathbf{p}$ and $\mathbf{q}$ on the same set $\mathcal{X}$. Define the one-variable binary entropy function $\mathrm{H}_2 \colon [0,1] \to [0,1]$ by

$$\mathrm{H}_2(t) = t \log \frac{1}{t} + (1-t) \log \frac{1}{1-t}\,,$$

where, as usual, we take $0 \log 0 = 0$. Thus, the function vanishes at 0 and 1 and is symmetric around $\frac{1}{2}$: $\mathrm{H}_2(t) = \mathrm{H}_2(1-t)$. Some straightforward calculus shows that

- H_2 is concave on $(0,1)$;
- $\mathrm{H}_2(t)$ is maximized at $t = \frac{1}{2}$, where $\mathrm{H}_2(\frac{1}{2}) = 1$;
- $\mathrm{H}_2(t)$ has infinite derivative at $t = 0$;
- $\mathrm{H}_2(\frac{1}{2} - \delta) = 1 - \Theta(\delta^2)$, for small positive δ.

The following is a less straightforward property of H_2.

Lemma 1.12 (Lin's inequality [Lin91])**.** $\mathrm{H}_2(t) \leq 2\sqrt{t(1-t)}$. □

Theorem 1.13 (Pinsker's inequality)**.** $\mathrm{D}_{\mathrm{KL}}(\mathbf{p} \,\|\, \mathbf{q}) \geq \frac{2}{\ln 2}\mathrm{D}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})$.

Proof. This is worked out as a homework problem. □

Theorem 1.14 (Vajda's inequality)**.** $\mathrm{D}_{\mathrm{JS}}(\mathbf{p}, \mathbf{q}) \geq 1 - \mathrm{H}_2\left(\frac{1-\mathrm{D}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})}{2}\right)$.

Proof. We start with two useful formulas.

$$\begin{aligned}
\mathrm{D}_{\mathrm{JS}}(\mathbf{p}, \mathbf{q}) &= \frac{1}{2}\sum_{x\in\mathcal{X}}\left(p_x \log\frac{2p_x}{p_x+q_x} + q_x \log\frac{2q_x}{p_x+q_x}\right) && (12)\\
&= \frac{1}{2}\sum_{x\in\mathcal{X}}(p_x+q_x)\left(\log 2 + \frac{p_x}{p_x+q_x}\log\frac{p_x}{p_x+q_x} + \frac{q_x}{p_x+q_x}\log\frac{q_x}{p_x+q_x}\right)\\
&= \frac{1}{2}\sum_{x\in\mathcal{X}}(p_x+q_x)\left(1 - \mathrm{H}_2\left(\frac{p_x}{p_x+q_x}\right)\right)\\
&= 1 - \sum_{x\in\mathcal{X}}\frac{p_x+q_x}{2}\mathrm{H}_2\left(\frac{p_x}{p_x+q_x}\right), && (13)
\end{aligned}$$

where (12) follows directly from eqs. (7) and (8); and

$$\mathrm{D}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) = \sum_{x\in\mathcal{X}}\frac{|p_x-q_x|}{2} = \sum_{x\in\mathcal{X}}\left(\frac{p_x+q_x}{2} - \min\{p_x, q_x\}\right) = 1 - \sum_{x\in\mathcal{X}}\min\{p_x, q_x\}. \qquad (14)$$

Since H_2 is symmetric around $\frac{1}{2}$ and concave on $(0, 1)$, Jensen's inequality gives

$$\begin{aligned}
\sum_{x\in\mathcal{X}}\frac{p_x+q_x}{2}\mathrm{H}_2\left(\frac{p_x}{p_x+q_x}\right) &= \sum_{x\in\mathcal{X}}\frac{p_x+q_x}{2}\mathrm{H}_2\left(\frac{\min\{p_x, q_x\}}{p_x+q_x}\right)\\
&\leq \mathrm{H}_2\left(\sum_{x\in\mathcal{X}}\frac{p_x+q_x}{2}\frac{\min\{p_x, q_x\}}{p_x+q_x}\right) = \mathrm{H}_2\left(\frac{1-\mathrm{D}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})}{2}\right),
\end{aligned}$$

where the final step uses (14). Combining this with eq. (13) completes the proof. □

Theorem 1.15. $\mathrm{D}_{\mathrm{JS}}(\mathbf{p}, \mathbf{q}) \geq \mathrm{h}^2(\mathbf{p}, \mathbf{q})$.

Proof. By eq. (13) and Lin's inequality (Lemma 1.12),

$$\mathrm{D}_{\mathrm{JS}}(\mathbf{p}, \mathbf{q}) \geq 1 - \sum_{x\in\mathcal{X}}\frac{p_x+q_x}{2}\cdot 2\sqrt{\frac{p_x}{p_x+q_x}\frac{q_x}{p_x+q_x}} = 1 - \sum_{x\in\mathcal{X}}\sqrt{p_x q_x} = \mathrm{h}^2(\mathbf{p}, \mathbf{q}),$$

where the final step uses eq. (10). □

2 A Unified Treatment of Communication Complexity Measures

Consider a general two-player communication protocol Π on input space $\mathcal{X} \times \mathcal{Y}$, where the players may use both public and private randomness. Given an input $(x, y) \in \mathcal{X} \times \mathcal{Y}$, a coin string $r_A \in \{0,1\}^{\ell_A}$ private to Alice, a coin string $r_B \in \{0,1\}^{\ell_B}$ private to Bob, and a public coin string $r \in \{0,1\}^{\ell}$, the protocol Π will generate a certain transcript; let $\Pi(x, y, r_A, r_B, r)$ denote this transcript. By definition, this transcript ends with the output of the protocol explicitly given; let $\mathrm{out}(\tau)$ denote the output corresponding to a transcript τ. Let (R_A, R_B, R) denote a random setting of the coin strings; w.l.o.g., we may assume that $(R_A, R_B, R) \in_R \{0,1\}^{\ell_A+\ell_B+\ell}$. A public-coin protocol is one with $\ell_A = \ell_B = 0$ and a private-coin protocol is one with $\ell = 0$. A deterministic protocol is one with $\ell_A = \ell_B = \ell = 0$.

Consider a function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ and a distribution μ on $\mathcal{X} \times \mathcal{Y}$. Let $(X, Y) \sim \mu$. The worst-case and μ-distributional errors of Π in computing f are defined as follows, respectively.

$$\mathrm{err}(\Pi, f) := \max\{\Pr[\mathrm{out}(\Pi(x, y, R_A, R_B, R)) \neq f(x, y)] : (x, y) \in \mathcal{X} \times \mathcal{Y}\}\,, \tag{15}$$

$$\mathrm{err}^{\mu}(\Pi, f) := \Pr[\mathrm{out}(\Pi(X, Y, R_A, R_B, R)) \neq f(X, Y)]\,. \tag{16}$$

Observe that if Π is deterministic then $\mathrm{err}(\Pi, f) \in \{0,1\}$, and moreover $\mathrm{err}(\Pi, f) = 0$ iff Π computes f correctly on all inputs. When f is clear from context, we sometimes drop f from these notations and simply write $\mathrm{err}(\Pi)$ and $\mathrm{err}^{\mu}(\Pi)$.

Let $M = \Pi(X, Y, R_A, R_B, R)$; recall that $(X, Y) \sim \mu$. Let $\|\tau\|$ denote the length of the transcript τ. We define the cost, the μ-distributional cost, the μ-information cost, and the μ-external information cost of Π as follows, respectively.

$$\mathrm{cost}(\Pi) := \max\left\{\|\Pi(x, y, r_A, r_B, r)\| : (x, y) \in \mathcal{X} \times \mathcal{Y}, (r_A, r_B, r) \in \{0,1\}^{\ell_A+\ell_B+\ell}\right\}, \tag{17}$$

$$\mathrm{cost}^{\mu}(\Pi) := \mathbb{E}\|M\|\,, \tag{18}$$

$$\mathrm{icost}^{\mu}(\Pi) := \mathrm{I}(M : X \mid YR) + \mathrm{I}(M : Y \mid XR)\,, \tag{19}$$

$$\mathrm{eicost}^{\mu}(\Pi) := \mathrm{I}(M : XY \mid R)\,. \tag{20}$$

The following lemma relates these cost measures.

Lemma 2.1. *For all Π and μ as above:* $\mathrm{icost}^{\mu}(\Pi) \leq \mathrm{eicost}^{\mu}(\Pi) \leq \mathrm{cost}^{\mu}(\Pi) \leq \mathrm{cost}(\Pi)$.

Proof (Sketch). The final inequality is obvious.

We prove the first inequality. Let $M_1, M_2, \ldots$ denote the successive bits of M. Let $\mathcal{T}_A = \{t :$ Alice sends the bit $M_t\}$ and $\mathcal{T}_B = \{t :$ Bob sends the bit $M_t\}$. Notice that when $t \in \mathcal{T}_B$, we have $\mathrm{I}(X : M_t \mid M_1 \ldots M_{t-1} YR) = 0$. Therefore, using the chain rule for mutual information,

$$\begin{aligned}\mathrm{icost}^{\mu}(\Pi) &= \sum_{t \geq 1} \mathrm{I}(X : M_t \mid M_1 \ldots M_{t-1} YR) + \sum_{t \geq 1} \mathrm{I}(Y : M_t \mid M_1 \ldots M_{t-1} XR) \\ &= \sum_{t \in \mathcal{T}_A} \mathrm{I}(X : M_t \mid M_1 \ldots M_{t-1} YR) + \sum_{t \in \mathcal{T}_B} \mathrm{I}(Y : M_t \mid M_1 \ldots M_{t-1} XR)\,. \end{aligned} \tag{21}$$

At this point, we could prove the weaker statement $\mathrm{icost}^{\mu}(\Pi) \leq \mathrm{cost}(\Pi)$ by noting that each term in (21) is at most $\mathrm{H}(M_t)$, for some t, and since M_t is a single bit, $\mathrm{H}(M_t) \leq 1$. Instead, to prove the first inequality of the lemma, we use the chain rule and split the resulting sum into two parts:

$$\begin{aligned}\mathrm{eicost}^{\mu}(\Pi) &= \sum_{t \in \mathcal{T}_A} \mathrm{I}(XY : M_t \mid M_1 \ldots M_{t-1} R) + \sum_{t \in \mathcal{T}_B} \mathrm{I}(XY : M_t \mid M_1 \ldots M_{t-1} R) \\ &\geq \sum_{t \in \mathcal{T}_A} \mathrm{I}(X : M_t \mid M_1 \ldots M_{t-1} YR) + \sum_{t \in \mathcal{T}_B} \mathrm{I}(Y : M_t \mid M_1 \ldots M_{t-1} XR)\,,\end{aligned}$$

where we have used $\mathrm{I}(AB : C) \geq \mathrm{I}(A : C \mid B)$. Comparing the above with (21) completes the proof.

We prove the second inequality. Since, for each $r \in \{0,1\}^\ell$, the set of transcripts of Π conditioned on $R = r$ must be a prefix code,

$$\mathrm{eicost}^\mu(\Pi) = \mathrm{I}(M : XY \mid R) \leq \mathrm{H}(M \mid R) \leq \mathbb{E}\|M\|\,,$$

where the final inequality uses the (nontrivial) theorem that the expected length of a prefix code is at least the entropy of the source. □

Based on our various cost measures we define communication complexity measures for functions that define two-party communication problems.

$$\begin{aligned}
\mathrm{D}(f) &:= \min\{\mathrm{cost}(\Pi) :\ \Pi \text{ is deterministic}, \mathrm{err}(\Pi, f) = 0\} && (22)\\
\mathrm{R}_\varepsilon(f) &:= \min\{\mathrm{cost}(\Pi) :\ \mathrm{err}(\Pi, f) \leq \varepsilon\}\,, && (23)\\
\mathrm{D}^\mu_\varepsilon(f) &:= \min\{\mathrm{cost}(\Pi) :\ \mathrm{err}^\mu(\Pi, f) \leq \varepsilon\}\,, && (24)\\
\mathrm{IC}^\mu_\varepsilon(f) &:= \inf\{\mathrm{icost}^\mu(\Pi) :\ \mathrm{err}(\Pi, f) \leq \varepsilon\}\,, && (25)\\
\mathrm{EIC}^\mu_\varepsilon(f) &:= \inf\{\mathrm{eicost}^\mu(\Pi) :\ \mathrm{err}(\Pi, f) \leq \varepsilon\}\,. && (26)
\end{aligned}$$

The minimum in (23) is achieved by a public-coin protocol because privateness of randomness plays no role in the definition. In contrast, eqs. (25) and (26) require fully general randomized protocols because the information cost measures treat public and private randomness differently. By an averaging argument, the minimum in (24) is achieved by a *deterministic* protocol, justifying the "D" in the notation.

The following corollary of Lemma 2.1 is immediate.

Corollary 2.2. *For all f and μ as above and all $\varepsilon \geq 0$: $\mathrm{IC}^\mu_\varepsilon(f) \leq \mathrm{EIC}^\mu_\varepsilon(f) \leq \mathrm{R}_\varepsilon(f) \leq \mathrm{D}(f)$.* □

Yao's minimax lemma, applied to these notions of cost and error, gives us another relation.

Lemma 2.3. *For all f as above and all $\varepsilon \geq 0$: $\mathrm{R}_\varepsilon(f) = \max\{\mathrm{D}^\mu_\varepsilon(f) :\ \mu \text{ is a distribution on } \mathcal{X} \times \mathcal{Y}\}$.* □

3 Lower Bound for Disjointness via Information Complexity

We shall lower bound $\mathrm{R}(\textsc{disj}_n)$ by thinking of $\textsc{disj}_n$ as a direct-summed version of $\textsc{and}_1$, where $\textsc{and}_1(x, y) := x \wedge y$ defines the communication problem of computing the logical AND of two bits: x, held by Alice, and y, held by Bob. Note that, for $x = x_1 \ldots x_n$ and $y = y_1 \ldots y_n$, we have

$$\textsc{disj}_n(x, y) = \neg \bigvee_{j=1}^{n} x_j \wedge y_j = \neg \bigvee_{j=1}^{n} \textsc{and}_1(x_j, y_j)\,.$$

Let Π be a protocol for $\textsc{disj}_n$ with $\mathrm{err}(\Pi) \leq \varepsilon$. Let λ denote the following distribution on the input space for $\textsc{and}_1$:

$$\lambda(0,0) = \lambda(0,1) = \lambda(1,0) = \frac{1}{3}; \quad \lambda(1,1) = 0\,. \tag{27}$$

Let λ^n denote the distribution obtained by taking n i.i.d. copies of λ (i.e., $\lambda \otimes \cdots \otimes \lambda$); it is a distribution on the input space for $\textsc{disj}_n$. Our proof strategy will be to construct certain protocols $\Pi_1, \ldots, \Pi_n$ for $\textsc{and}_1$, each with error at most ε, and establish that

$$\mathrm{icost}^{\lambda^n}(\Pi) \overset{①}{\geq} \sum_{j=1}^{n} \mathrm{icost}^{\lambda}(\Pi_j) \geq n \cdot \mathrm{IC}_{\varepsilon}^{\lambda}(\textsc{and}_1) \overset{②}{\geq} n \cdot \phi(\varepsilon) = \Omega(n)\,, \tag{28}$$

for an appropriate function ϕ. Taking the infimum over all ε-error $\textsc{disj}_n$ protocols then establishes that $\mathrm{IC}_{\varepsilon}^{\lambda^n}(\textsc{disj}_n) = \Omega(n)$ and, by Corollary 2.2, that $\mathrm{R}_{\varepsilon}(\textsc{disj}_n) = \Omega(n)$.

The middle inequality in (28) is trivial. The inequality marked ①, establishing a direct sum property, is proved using a simulation argument. The inequality marked ② is proved using an analytic argument involving distances between probability distributions.

3.1 Direct Sum via Simulation

This part has very little to do with the $\textsc{disj}_n$ function, so we shall generalize the setup considerably. Let $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be a function and let μ be a probability distribution on $\mathcal{X} \times \mathcal{Y}$. Let $g : \mathcal{X}^n \times \mathcal{Y}^n \to \mathcal{Z}'$ be another function. Our generalization will relate certain information complexities of f and g under the following circumstance: suppose there exists a "recovery" function $h : \mathcal{Z}' \to \mathcal{Z}$ such that

$$\forall\, j \in [n] : \quad \Pr[h(g(X_{1:j-1} x X_{j+1:n}, Y_{1:j-1} y Y_{j+1:n})) \neq f(x, y)] \leq \delta\,, \tag{29}$$

where the pairs $\{(X_i, Y_i)\}_{i \in [n] \setminus \{j\}}$ are i.i.d. and each $(X_i, Y_i) \sim \mu$. The notation $X_{a:b}$ is shorthand for the sequence $X_a X_{a+1} \cdots X_b$, and the juxtaposition of the X_is denotes concatenation (rather than multiplication). This models our particular situation with $\textsc{disj}_n$ using the following instantiation:

$$\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \mathcal{Z}' = \{0,1\}, \quad f = \textsc{and}_1, \quad g = \textsc{disj}_n, \quad h = \textsc{not}, \quad \mu = \lambda, \quad \delta = 0\,.$$

Let $\widehat{\Pi}$ be an ε-error protocol for g. We shall construct protocols $\widehat{\Pi}_1, \ldots, \widehat{\Pi}_n$ for f, each with error at most $\varepsilon + \delta$, such that

$$\mathrm{icost}^{\mu^n}(\widehat{\Pi}) \geq \sum_{j=1}^{n} \mathrm{icost}^{\mu}(\widehat{\Pi}_j)\,, \tag{30}$$

implying

$$\mathrm{IC}_{\varepsilon}^{\mu^n}(g) \geq n \cdot \mathrm{IC}_{\varepsilon+\delta}^{\mu}(f)\,.$$

We now specify the protocols $\widehat{\Pi}_j$.

Protocol $\widehat{\Pi}_j(x, y)$

- The players publicly sample bits $X_{1:j-1}$ and $Y_{j+1:n}$, all mutually independent, so that each X_i and each Y_i are distributed according to the first and second marginals of μ, respectively.
- Alice privately samples $X_{j+1:n}$ from an appropriately conditioned distribution so that $((X_{j+1}, Y_{j+1}), \ldots, (X_n, Y_n)) \sim \mu^{n-j}$.
- Bob privately samples $Y_{1:j-1}$ from an appropriately conditioned distribution so that $((X_1, Y_1), \ldots, (X_{j-1}, Y_{j-1})) \sim \mu^{j-1}$.
- The players run $\widehat{\Pi}(X_{1:j-1}xX_{j+1:n}, Y_{1:j-1}yY_{j+1:n})$, modifying the output from z (say) to $h(z)$.

The construction of $\widehat{\Pi}_j$, together with eq. (29), ensures that $\mathrm{err}(\widehat{\Pi}_j) \le \mathrm{err}(\widehat{\Pi}) + \delta \le \varepsilon + \delta$.

We turn to proving eq. (30). Let R and $R^{(j)}$ denote the public random strings of $\widehat{\Pi}$ and $\widehat{\Pi}_j$, respectively. Let $(X_{1:n}, Y_{1:n}) \sim \mu^n$ be a random input to $\widehat{\Pi}$, leading to a transcript M. Let $M^{(j)}$ denote the transcript when (X_j, Y_j) is fed to $\widehat{\Pi}_j$. Then

$$\begin{aligned}
\mathrm{I}(M^{(j)} : X_j \mid Y_j R^{(j)}) &= \mathrm{I}(M : X_j \mid Y_j X_{1:j-1} Y_{j+1:n} R) \\
&\le \mathrm{I}(M Y_{1:j-1} : X_j \mid Y_j X_{1:j-1} Y_{j+1:n} R) \\
&= \mathrm{I}(Y_{1:j-1} : X_j \mid Y_j X_{1:j-1} Y_{j+1:n} R) + \mathrm{I}(M : X_j \mid Y_{1:j-1} Y_j X_{1:j-1} Y_{j+1:n} R) \\
&= 0 + \mathrm{I}(M : X_j \mid X_{1:j-1} Y_{1:n} R)\,,
\end{aligned}$$

where the vanishing of the first term is because the (X_i, Y_i) pairs are mutually independent. Using the chain rule for mutual information,

$$\sum_{j=1}^{n} \mathrm{I}(M^{(j)} : X_j \mid Y_j R^{(j)}) \le \sum_{j=1}^{n} \mathrm{I}(M : X_j \mid X_{1:j-1} Y_{1:n} R) = \mathrm{I}(M : X_{1:n} \mid Y_{1:n} R)\,. \tag{31}$$

An analogous argument that flips the roles of Xs and Ys and uses the chain rule "from the other end" gives us

$$\sum_{j=1}^{n} \mathrm{I}(M^{(j)} : Y_j \mid X_j R^{(j)}) \le \sum_{j=1}^{n} \mathrm{I}(M : Y_j \mid X_{1:n} Y_{j+1:n} R) = \mathrm{I}(M : Y_{1:n} \mid X_{1:n} R)\,. \tag{32}$$

Using eqs. (31) and (32), we get

$$\begin{aligned}
\mathrm{icost}^{\mu^n}(\widehat{\Pi}) &= \mathrm{I}(M : X_{1:n} \mid Y_{1:n} R) + \mathrm{I}(M : Y_{1:n} \mid X_{1:n} R) \\
&\ge \sum_{j=1}^{n} \Big(\mathrm{I}(M^{(j)} : X_j \mid Y_j R^{(j)}) + \mathrm{I}(M^{(j)} : Y_j \mid X_j R^{(j)}) \Big) = \sum_{j=1}^{n} \mathrm{icost}^{\mu}(\widehat{\Pi}_j)\,.
\end{aligned}$$

which proves the desired generalization of step ① in eq. (28).

3.2 Relating Information Complexity to Distances Between Distributions

We now justify step ② in eq. (28). We need to show that $\mathrm{IC}^{\lambda}_{\varepsilon}(\mathrm{AND}_1) = \Omega(1)$, for small enough ε. To this end, let Ξ be an ε-error protocol for AND_1 that uses only private coins. We shall prove that $\mathrm{icost}^{\lambda}(\Xi) \geq \phi(\varepsilon)$ for some definite function ϕ. This inequality will then extend to general protocols by averaging over all settings of its public random string.

Let $(X, Y) \sim \lambda$ be a random input to Ξ and let M be the resulting random transcript. For $i, j \in \{0, 1\}$, let $\mathbf{p}^{ij}$ denote the distribution of M conditioned on $(X, Y) = (i, j)$. Recalling the definition of λ in eq. (27),

$$\mathrm{I}(M : X \mid Y) = \frac{2}{3}\,\mathrm{I}(M : X \mid Y = 0) + \frac{1}{3}\,\mathrm{I}(M : X \mid Y = 1) = \frac{2}{3}\,\mathrm{I}(M : X \mid Y = 0)\,.$$

Using this and an analogous calculation for $\mathrm{I}(M : Y \mid X)$,

$$\begin{aligned}
\mathrm{icost}^{\lambda}(\Xi) &= \mathrm{I}(M : X \mid Y) + \mathrm{I}(M : Y \mid X) \\
&= \frac{2}{3}\left(\mathrm{I}(M : X \mid Y = 0) + \mathrm{I}(M : Y \mid X = 0)\right) \\
&= \frac{2}{3}\left(\mathrm{D}_{\mathrm{JS}}(\mathbf{p}^{00}, \mathbf{p}^{10}) + \mathrm{D}_{\mathrm{JS}}(\mathbf{p}^{00}, \mathbf{p}^{01})\right) \\
&\geq \frac{2}{3}\left(\mathrm{h}^2(\mathbf{p}^{00}, \mathbf{p}^{10}) + \mathrm{h}^2(\mathbf{p}^{00}, \mathbf{p}^{01})\right) \\
&\geq \frac{1}{3}\left(\mathrm{h}(\mathbf{p}^{00}, \mathbf{p}^{10}) + \mathrm{h}(\mathbf{p}^{00}, \mathbf{p}^{01})\right)^2 \\
&\geq \frac{1}{3}\mathrm{h}^2(\mathbf{p}^{01}, \mathbf{p}^{10}) \\
&= \frac{1}{3}\mathrm{h}^2(\mathbf{p}^{00}, \mathbf{p}^{11})\,,
\end{aligned}$$

where the last step uses the Cut-and-Paste Lemma.

Since Ξ solves AND_1 with error at most ε, it follows that $\mathrm{D}_{\mathrm{TV}}(\mathbf{p}^{00}, \mathbf{p}^{11}) \geq 1 - 2\varepsilon$. By the result of a homework exercise, we conclude that $\mathrm{h}^2(\mathbf{p}^{00}, \mathbf{p}^{11}) \geq 1 - 2\sqrt{\varepsilon}$. Thus, we have

$$\mathrm{icost}^{\lambda}(\Xi) \geq 1 - 2\sqrt{\varepsilon}\,.$$

3.3 Final Result and Discussion

Returning to the initial outline in eq. (28), we have shown the following

Theorem 3.1 (Disjointness lower bound). *$\mathrm{R}_{\varepsilon}(\mathrm{DISJ}_n) \geq \frac{1}{3}(1 - 2\sqrt{\varepsilon})n$.* □

This is a rather elegant bound! However, there is no reason to believe that it is tight. In fact, we know that is *not* tight at either of the two extremes: when $\varepsilon \approx 0$ or when $\varepsilon \approx \frac{1}{2}$.

In general, let $C_{\varepsilon}^{\mathrm{DISJ}}$ denote the "correct" constant in the asymptotic bound on $\mathrm{R}_{\varepsilon}(\mathrm{DISJ}_n)$, i.e.,

$$C_{\varepsilon}^{\mathrm{DISJ}} = \limsup_{n \to \infty} \frac{\mathrm{R}_{\varepsilon}(\mathrm{DISJ}_n)}{n}\,.$$

The above proof shows that $C_{\varepsilon}^{\mathrm{DISJ}} \geq \frac{1}{3}(1 - 2\sqrt{\varepsilon})$. As $\varepsilon \to 0$, this bound approaches $\frac{1}{3}$. Remarkably, Braverman et al. [BGPW13] showed that

$$\lim_{\varepsilon \to 0} C_{\varepsilon}^{\mathrm{DISJ}} = \max\left\{\frac{x}{\ln 2} + \frac{x^2}{1 - 2x}\log\frac{x}{1 - x} + (1 - 2x)\log\frac{1 - x}{1 - 2x} \,:\, x \in [0, 1]\right\} \approx 0.482702\,.$$

Let us denote the above constant by C^{DISJ}. Dagan et al. [DFHL18] extended this result to all small enough ε by showing that, as $\varepsilon \to 0$, one has $C_{\varepsilon}^{\mathrm{DISJ}} = C^{\mathrm{DISJ}} - \Theta(\mathrm{H}_2(\varepsilon))$.

At the other end, when $\varepsilon \geq \frac{1}{4}$, Theorem 3.1 becomes trivial. To obtain a meaningful lower bound on $\mathrm{R}_\varepsilon(\mathrm{DISJ}_n)$, with $\varepsilon \approx \frac{1}{2}$, we can turn to the standard technique of error reduction by repetition. Based on standard Chernoff bounds, this gives $\mathrm{R}_{\frac{1}{2}-\delta}(\mathrm{DISJ}_n) = \Omega(\delta^2 n)$. Remarkably, this is not tight. Braverman and Moitra [BM13] used further information theoretic techniques to show that the correct asymptotics are $\mathrm{R}_{\frac{1}{2}-\delta}(\mathrm{DISJ}_n) = \Theta(\delta n)$.

References

[BGPW13] Mark Braverman, Ankit Garg, Denis Pankratov, and Omri Weinstein. From information to exact communication. In *Proc. 45th Annual ACM Symposium on the Theory of Computing*, pages 151–160, 2013.

[BM13] Mark Braverman and Ankur Moitra. An information complexity approach to extended formulations. In *Proc. 45th Annual ACM Symposium on the Theory of Computing*, pages 161–170, 2013.

[DFHL18] Yuval Dagan, Yuval Filmus, Hamed Hatami, and Yaqiao Li. Trading information complexity for error. *Theor. Comput.*, 14(1):1–73, 2018. Preliminary version in *Proc. 32nd Annual IEEE Conference on Computational Complexity*, pages 16:1–16:59, 2017.

[Lin91] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theory*, 37(1):145–151, 1991.