1 Information Theory

1.1 Entropy

Definition 1.1 (Entropy)**.** Let X be a discrete random variable taking values in a finite set $\mathcal{X}$ with probability mass function $p(x) = P(X = x)$. The *entropy* of X, denoted $H(X)$, is defined as:

$$H(X) \coloneqq -\sum_{x \in \mathcal{X}} p(x) \log p(x),$$

where the logarithm is typically taken base 2 (bits) or base e (nats).

Remark 1.1. If $p(x) = 0$, we set $p(x) \log p(x) \coloneqq 0$. This ensures that $p(x) \log p(x)$ is continuous on $[0, 1]$.

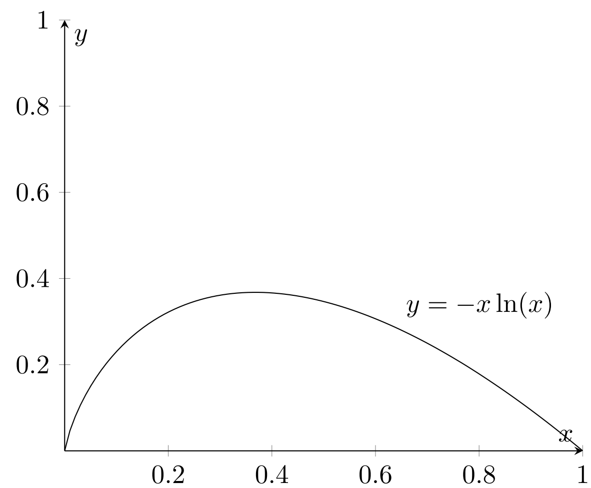

Figure 1: Plot of the function $y = -x \ln(x)$.

Remark 1.2. Entropy measures the uncertainty or information content of a random variable. Higher entropy indicates more unpredictability.

Proposition 1.1 (Non-Negativity of Entropy)**.** *For any discrete random variable X, we have $H(X) \geq 0$.*

Proof. Since $0 \leq p(x) \leq 1$ and $-\log p(x) \geq 0$, each term in the sum is non-negative, so their total sum is non-negative. $\square$

Lemma 1.1 (Jensen's Inequality)**.** *Let $X \in \mathcal{X}$ be a random variable over a finite set $\mathcal{X}$, and let ϕ be a convex function defined for all X. Then:*

$$\phi(E[X]) \leq E[\phi(X)] \quad .$$

Proof. We use induction over $n = |\mathcal{X}|$. The base case $n = 1$ is trivial. Hence, assume that the claim holds for some n. We now prove the claim for $n + 1$. Clearly, for $n > 1$, we must have $P(X = x_k) < 1$ for some $x_k \in \mathcal{X}$. Without loss of generality, we assume $k = n + 1$. Hence:

$$\begin{aligned}
\phi(E[X]) &= \phi\left(\sum_{i=1}^{n+1} p(x_i)x_i\right) \\
&= \phi\left(\left[(1 - p(x_{n+1}))\sum_{i=1}^{n} \frac{p(x_i)}{1 - p(x_{n+1})}x_i\right] + p(x_{n+1})x_{n+1}\right) \\
&\underset{\text{convexity}}{\leq} (1 - p(x_{n+1}))\phi\left(\sum_{i=1}^{n} \frac{p(x_i)}{1 - p(x_{n+1})}x_i\right) + p(x_{n+1})\phi(x_{n+1}) \\
&\underset{\text{inductive hypothesis}}{\leq} (1 - p(x_{n+1}))\sum_{i=1}^{n} \frac{p(x_i)}{1 - p(x_{n+1})}\phi(x_i) + p(x_{n+1})\phi(x_{n+1}) \\
&= \sum_{i=1}^{n+1} p(x_i)\phi(x_i) = E[\phi(X)] \quad .
\end{aligned}$$

□

Remark 1.3. For strictly convex ϕ, it can be shown that

$\phi(E[X]) = E[\phi(X)]$ is maximized $\iff$ X is sampled from a uniform distribution .

Proposition 1.2 (Maximum Entropy)**.** *For a discrete random variable X over n outcomes, entropy is maximized when X is uniform:*

$$H(X) \leq \log n \quad .$$

Proof. We have:

$$\begin{aligned}
-H(X) &= -E[-\log(p(X))] \\
&= E\left[-\log\left(\frac{1}{p(X)}\right)\right] \\
&\underset{\text{Jensen's Inequality}}{\geq} -\log\left(E\left[\frac{1}{p(X)}\right]\right) \\
&= -\log n \quad ,
\end{aligned}$$

where we assumed $p(X) > 0$. Of course, the cases where $p(X) = 0$ follow directly, since $p(X) \log p(X) = 0$.

$H(X) \leq \log n$ follows directly. Note that we have equality iff X has uniform distribution (since $-\log(x)$ is strictly convex). $\square$

1.1.1 Joint, Conditional, and Cross Entropy

Definition 1.2 (Joint Entropy)**.** For a pair of discrete random variables X and Y, the joint entropy is:

$$H(X, Y) \coloneqq -\sum_{x,y} p(x, y) \log p(x, y) \quad .$$

Definition 1.3 (Conditional Entropy)**.** The conditional entropy of Y given X is defined as:

$$H(Y \mid X) \coloneqq \sum_{x} p(x) H(Y \mid X = x) = -\sum_{x,y} p(x, y) \log p(y \mid x).$$

Corollary 1.1. *We immediately see from the first equation that $H(Y \mid X) \geq 0$.*

Theorem 1.1 (Chain Rule for Entropy)**.**

$$H(X,Y) = H(X) + H(Y \mid X) \quad .$$

Proof. We have:

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} p(x,y) \log p(x,y) \\ &= -\sum_{x,y} p(x,y) \log \left(p(x)p(y \mid x)\right) \\ &= -\sum_{x,y} p(x,y) \log p(x) - \sum_{x,y} p(x,y) \log p(y \mid x) \\ &= H(X) + H(Y \mid X) \quad . \end{aligned}$$

□

Corollary 1.2. $H(X,Y) \geq 0$ *follows directly.*

Definition 1.4 (Cross-Entropy)**.** Let p and q be two probability distributions over a finite set $\mathcal{X}$, with $p(x) > 0 \Rightarrow q(x) > 0$. The *cross-entropy* of p relative to q is defined as:

$$H_q(p) := -\sum_{x \in \mathcal{X}} p(x) \log q(x) \quad .$$

Remark 1.4. Cross-entropy measures the expected number of bits required to encode samples from p using a code optimized for the distribution q.

Remark 1.5. Cross-entropy is non-negative (see section 1.2).

1.1.2 Properties of Entropy

Proposition 1.3. *Conditional entropy satisfies:*

$$H(Y \mid X) \leq H(Y) \quad ,$$

with equality if and only if X and Y are independent.

Proof. From the chain rule:

$$H(X,Y) = H(Y) + H(X \mid Y) = H(X) + H(Y \mid X) \quad ,$$

which implies:

$$H(Y \mid X) = H(Y) + H(X \mid Y) - H(X) = H(Y) - I(X;Y) \quad ,$$

with mutual information $I(X;Y) \geq 0$ (see section 1.3). Equality holds if and only if $I(X;Y) = 0$, i.e., X and Y are independent. □

Corollary 1.3 (Subadditivity of Entropy)**.** *For any two random variables X and Y,*

$$H(X,Y) \leq H(X) + H(Y) \quad ,$$

with equality if and only if X and Y are independent.

Proof. From the chain rule:

$$H(X,Y) = H(X) + H(Y \mid X) \leq H(X) + H(Y) \quad ,$$

since $H(Y \mid X) \leq H(Y)$ based on proposition 1.3. Equality holds if and only if $H(Y \mid X) = H(Y)$, i.e., X and Y are independent. □

Theorem 1.2 (Concavity of Entropy)**.** *The entropy function $H(p)$, where $p \in \Delta$ is a probability vector, is concave on the probability simplex Δ.*

Proof. This follows from the fact that $f(x) = -x \log x$ is concave for $x \in [0,1]$, and entropy is the sum of such terms. Therefore, for every convex combination $p = \lambda p_1 + (1-\lambda)p_2$:

$$H(p) \geq \lambda H(p_1) + (1-\lambda)H(p_2) \quad .$$

□

Summary of Key Properties

- Non-negativity: $H(X) \geq 0$
- Maximum entropy: $H(X) \leq \log |\mathcal{X}|$
- Chain rule: $H(X,Y) = H(X) + H(Y \mid X)$
- Subadditivity: $H(X,Y) \leq H(X) + H(Y)$
- Conditioning reduces entropy: $H(Y \mid X) \leq H(Y)$
- Concavity: $H(p)$ is concave in the distribution p

1.2 Kullback-Leibler Divergence

Definition 1.5 (KL Divergence)**.** Let P and Q be two discrete probability distributions over the same finite set $\mathcal{X}$, with $P(x) > 0 \Rightarrow Q(x) > 0$. The Kullback-Leibler divergence (or relative entropy) from P to Q is defined as:

$$\begin{aligned} D_{\mathrm{KL}}(P\|Q) &:= \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)} \\ &= -\sum_{x} P(x)\log Q(x) + \sum_{x} P(x)\log P(x) \\ &= H_Q(P) - H(P) \quad . \end{aligned}$$

Remark 1.6. If $P(x) = Q(x) = 0$, we set $P(x)\log\dfrac{P(x)}{Q(x)} := 0$.

Remark 1.7. KL divergence measures the inefficiency of assuming that the distribution is Q when the true distribution is P. It is not a metric: it is not symmetric and does not satisfy the triangle inequality.

Lemma 1.2 (Gibb's Inequality)**.** *Suppose that $P = \{p_1, \ldots, p_n\}$ and $Q = \{q_1, \ldots, q_n\}$ are discrete probability distributions. Then:*

$$-\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} p_i \log q_i \quad .$$

Proof. The claim is equivalent to $\sum_{i=1}^{n} p_i \log p_i - \sum_{i=1}^{n} p_i \log q_i \geq 0$. We have:

$$\begin{aligned} \sum_{i=1}^{n} p_i \log p_i - \sum_{i=1}^{n} p_i \log q_i &= \sum_{i=1}^{n} p_i \log\frac{p_i}{q_i} \\ &= \sum_{i=1}^{n} p_i \left(-\log\frac{q_i}{p_i}\right) \\ &\underset{\text{Jensen's Inequality}}{\geq} -\log\left(\sum_{i=1}^{n} p_i \frac{q_i}{p_i}\right) \\ &= -\log(1) = 0 \quad . \end{aligned}$$

$\square$

Corollary 1.4. *It directly follows from the proof that $D_{\mathrm{KL}}(P\|Q) \geq 0$ and $0 \leq H(P) \leq H_Q(P)$.*

Proposition 1.4 (Additivity). *Let $P = P_1 \times P_2$, $Q = Q_1 \times Q_2$. Then:*

$$D_{\mathrm{KL}}(P\|Q) = D_{\mathrm{KL}}(P_1\|Q_1) + D_{\mathrm{KL}}(P_2\|Q_2) \quad .$$

Proof.

$$\begin{aligned}
D_{\mathrm{KL}}(P_1 \times P_2\|Q_1 \times Q_2) &= \sum_{x,y} P_1(x)P_2(y) \log \frac{P_1(x)P_2(y)}{Q_1(x)Q_2(y)} \\
&= \sum_{x,y} P_1(x)P_2(y) \left(\log \frac{P_1(x)}{Q_1(x)} + \log \frac{P_2(y)}{Q_2(y)} \right) \\
&= \sum_{x} P_1(x) \log \frac{P_1(x)}{Q_1(x)} + \sum_{y} P_2(y) \log \frac{P_2(y)}{Q_2(y)} \\
&= D_{\mathrm{KL}}(P_1\|Q_1) + D_{\mathrm{KL}}(P_2\|Q_2) \quad .
\end{aligned}$$

□

Proposition 1.5 (Entropy Representation via KL Divergence). *Let U be the uniform distribution over $\mathcal{X}$, where $|\mathcal{X}| = n$. Then for any distribution P,*

$$H(P) = \log n - D_{\mathrm{KL}}(P\|U) \quad .$$

Proof.

$$\begin{aligned}
D_{\mathrm{KL}}(P\|U) &= \sum_{x} P(x) \log \frac{P(x)}{1/n} = \sum_{x} P(x) \log P(x) + \sum_{x} P(x) \log n \\
&= -H(P) + \log n \quad .
\end{aligned}$$

□

Summary of Key Properties

- $D_{\mathrm{KL}}(P\|Q) \geq 0$
- $D_{\mathrm{KL}}(P\|Q) = 0 \iff P = Q$
- Asymmetric: $D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$
- Additive over independent distributions
- Connection to entropy: $H(P) = \log n - D_{\mathrm{KL}}(P\|U)$

1.3 Mutual Information

Definition 1.6 (Mutual Information)**.** Let X and Y be discrete random variables with joint distribution $p(x, y)$ and marginals $p(x)$, $p(y)$. The *mutual information* between X and Y is defined as:

$$I(X;Y) \coloneqq \sum_{x,y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \quad .$$

Remark 1.8. Mutual information quantifies how much knowing X reduces uncertainty about Y, and vice versa. Per definition, it is symmetric: $I(X;Y) = I(Y;X)$.

Proposition 1.6 (Equivalent Expressions)**.** *Mutual information can also be expressed as:*

$$\begin{aligned}
I(X;Y) &= D_{\mathrm{KL}}(p(x,y) \,\|\, p(x)p(y)) \\
&= H_{p(x)p(y)}(p(x,y)) - H(X,Y) \\
&= \left[-\sum_{x,y} p(x,y) \log(p(x)p(y)) \right] - H(X,Y) \\
&= H(X) + H(Y) - H(X,Y) \\
&= H(X) - H(X \mid Y) \\
&= H(Y) - H(Y \mid X)
\end{aligned}$$

Proof. Each follows from basic entropy identities and the definition of KL divergence. □

Corollary 1.5. $I(X;Y) \geq 0$*, since* $I(X;Y) = D_{\mathrm{KL}}(p(x,y) \| p(x)p(y))$ *and KL divergence is always non-negative.*

Definition 1.7 (Conditional Mutual Information)**.** Let X, Y, Z be discrete random variables. The *conditional mutual information* of X and Y given Z is defined as:

$$I(X;Y \mid Z) \coloneqq \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y \mid z)}{p(x \mid z)p(y \mid z)} \quad .$$

Equivalently, in terms of entropy:

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid (Y,Z)) \quad .$$

Proof.

$$\begin{aligned}
&H(X \mid Z) - H(X \mid (Y,Z)) \\
&= \sum_z p(z) H(X \mid Z = z) - \sum_{y,z} p(y,z) H(X \mid Y = y, Z = z) \\
&= -\sum_z p(z) \sum_x p(x \mid z) \log p(x \mid z) \;+\; \sum_{y,z} p(y,z) \sum_x p(x \mid y,z) \log p(x \mid y,z) \\
&= \sum_{x,y,z} p(x,y,z) \log \frac{p(x \mid y,z)}{p(x \mid z)} \\
&= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y \mid z)}{p(x \mid z) p(y \mid z)} \\
&= I(X;Y \mid Z) \quad .
\end{aligned}$$

□

Remark 1.9. Conditional mutual information measures how much knowing Y reduces the uncertainty of X, *given* that we already know Z.

Proposition 1.7 (Chain Rule for Mutual Information)**.** *Let X, Y, and Z be random variables. Then:*

$$I(X;Y,Z) = I(X;Z) + I(X;Y \mid Z) \quad .$$

Proof. We use entropy-based expressions for mutual information:

$$\begin{aligned}
I(X;Y,Z) &= H(X) - H(X \mid (Y,Z)) \\
&= I(X;Z) + H(X \mid Z) - H(X \mid (Y,Z)) \\
&= I(X;Z) + H(X \mid Z) - (H(X \mid Z) - I(X;Y \mid Z)) \\
&= I(X;Z) + I(X;Y \mid Z) \quad .
\end{aligned}$$

□

Proposition 1.8 (Non-Negativity of Conditional Mutual Information)**.** *It holds true that*

$$I(X;Y \mid Z) \geq 0 \quad .$$

Proof. We have:

$$\begin{aligned}
I(X;Y \mid Z) &= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y \mid z)}{p(x \mid z) p(y \mid z)} \\
&= \sum_z p(z) \sum_{x,y} p(x,y \mid z) \log \frac{p(x,y \mid z)}{p(x \mid z) p(y \mid z)} \\
&= \sum_z p(z) D_{\mathrm{KL}} \left(p(x,y \mid z) \,\|\, p(x \mid z) p(y \mid z) \right) \geq 0 \quad .
\end{aligned}$$

□

Corollary 1.6. *As a direct consequence, we have*

$$I(X;Z) \leq I(X;Y,Z) \quad .$$

Definition 1.8 (Conditional Independence)**.** Let X, Y, Z be discrete random variables. We say that X is *conditionally independent* of Z given Y, and write:

$$X \perp Z \mid Y$$

if and only if

$$p(z \mid x, y) = p(z \mid y) \quad \text{for all } x, y, z \quad .$$

Equivalently:

$$p(x, z \mid y) = p(x \mid y)p(z \mid y) \quad .$$

Proposition 1.9. *If $X \perp Z \mid Y$, then the conditional mutual information between X and Z given Y is zero:*

$$I(X;Z \mid Y) = 0 \quad .$$

Proof. By definition of conditional mutual information:

$$I(X;Z \mid Y) = \sum_{x,z,y} p(x,z,y) \log \frac{p(x,z \mid y)}{p(x \mid y)p(z \mid y)} \quad .$$

If $X \perp Z \mid Y$, then:

$$p(x, z \mid y) = p(x \mid y)p(z \mid y) \quad ,$$

so the logarithm becomes:

$$\log \frac{p(x \mid y)p(z \mid y)}{p(x \mid y)p(z \mid y)} = \log 1 = 0 \quad .$$

Hence, each term in the sum is zero, and:

$$I(X;Z \mid Y) = 0 \quad .$$

□

1.3.1 Data Processing Inequality

Lemma 1.3 (Markov Chain). *Let X, Y, Z be random discrete random variables forming the Markov chain $X \to Y \to Z$. Then:*

$$X \perp Z \mid Y \quad .$$

Proof. Per definition from Markov chains, we have:

$$p(z \mid x, y) = p(z \mid y) \quad ,$$

and hence $X \perp Z \mid Y$. □

Theorem 1.3 (Data Processing Inequality). *If $X \to Y \to Z$ is a Markov chain, then:*

$$I(X; Z) \leq I(X; Y) \quad .$$

Proof. We use the chain rule and conditional independence:

$$\begin{aligned} I(X; Z) &= I(X; Z, Y) - I(X; Y \mid Z) \\ &= I(X; Y) + I(X; Z \mid Y) - I(X; Y \mid Z) \quad . \end{aligned}$$

Since $X \to Y \to Z$, we have $I(X; Z \mid Y) = 0$. Thus:

$$I(X; Z) = I(X; Y) - I(X; Y \mid Z) \leq I(X; Y) \quad ,$$

because $I(X; Y \mid Z) \geq 0$. □

Corollary 1.7 (No Gain in Processing). *Any function $f(Y)$ of Y cannot increase information about X:*

$$I(X; f(Y)) \leq I(X; Y) \quad .$$

Proof. This follows by applying the DPI to the chain $X \to Y \to f(Y)$. □

Summary of Key Properties

- $I(X; Y) \geq 0$
- $I(X; Y) = 0$ if and only if $X \perp Y$
- $I(X; Y) = D_{\mathrm{KL}}(p(x, y) \| p(x)p(y))$
- Chain rule: $I(X; Y, Z) = I(X; Z) + I(X; Y \mid Z)$
- Data Processing Inequality: $X \to Y \to Z \Rightarrow I(X; Z) \leq I(X; Y)$

1.4 Bounding Mutual Information via Matrix Rank of the Joint Distribution

Theorem 1.4. *Let X, Y be random variables from finite sets $\mathcal{X}, \mathcal{Y}$, and let matrix $\boldsymbol{P}$ denote their joint probability distribution, i.e. $\boldsymbol{P}_{ij} = p(x_i, y_j)$. Let $r \coloneqq \operatorname{rank} \boldsymbol{P}$ denote the rank of matrix $\boldsymbol{P}$. Then we have*

$$I(X;Y) \leq \log r \quad .$$

Proof. Let $n \coloneqq |\mathcal{X}|$ and $m \coloneqq |\mathcal{Y}|$. If $\boldsymbol{P}$ has rank r, then so must the transition matrix $\boldsymbol{P}_{Y|X} \in \mathbb{R}^{m \times n}$ defined as $(\boldsymbol{P}_{Y|X})_{ij} \coloneqq p(y_i \mid x_j) = \frac{p(x_j, y_i)}{\sum_k p(x_k, y_i)}$, since $\boldsymbol{P}_{Y|X}$ is created from $\boldsymbol{P}$ by transposing and column scaling. If one column consisted of only zeros, i.e. $\sum_k p(x_k, y_i) = 0$, we may just copy a different scaled column vector to this column.

Now, let's analyze matrix $\boldsymbol{P}_{Y|X}$. First, note that it is a Markov chain transition matrix, and hence all its columns lie in the m-dimensional unit simplex. Consider the convex hull of the column vectors, it is a r-dimensional convex polytope in the m-dimensional unit simplex. Thus, we can find a r-dimensional simplex with corners collected by matrix $\boldsymbol{U}$ s.t. it is a superset of this polytope and still a subset of the (potentially) higher dimensional unit simplex.

Thus, every column vector in $\boldsymbol{P}_{Y|X}$ can be written as a convex combination of the column vectors in $\boldsymbol{U}$. It follows that $\boldsymbol{P}_{Y|X}$ can be decomposed as

$$\boldsymbol{P}_{Y|X} = \boldsymbol{U}\boldsymbol{V} \ , \quad \boldsymbol{U} \in \mathbb{R}^{m \times r}, \boldsymbol{V} \in \mathbb{R}^{r \times n} \quad ,$$

where both $\boldsymbol{U}$ and $\boldsymbol{V}$ are Markov chain transition matrices as well.

Hence, we can introduce a latent variable $Z \in \{1, \ldots, r\}$, which forms the Markov chain

$$X \underset{\boldsymbol{V}}{\to} Z \underset{\boldsymbol{U}}{\to} Y \quad .$$

Finally, based on theorem 1.3 it follows that

$$I(X;Y) \leq I(X;Z) = H(Z) - H(Z \mid X) \leq H(Z) \leq \log r \quad .$$

$\square$

1.5 Convergence of Mutual Information

Theorem 1.5 (Element-Wise Exponential Convergence Implies Exponential Convergence)**.** *Let $f : \mathcal{D} \to \mathbb{R}^m$ be a function defined on a convex domain $\mathcal{D} \subseteq \mathbb{R}^n$ that is a Cartesian product of real intervals, i.e., $\mathcal{D} = \mathcal{D}_1 \times \mathcal{D}_2 \times \cdots \times \mathcal{D}_n$ where each $\mathcal{D}_i \subseteq \mathbb{R}$ is an interval. Let $\{\boldsymbol{x}_k\}_{k=1}^{\infty} \subset \mathcal{D}$ be a sequence converging exponentially fast to $\boldsymbol{x}_0 \in \mathcal{D}$.*

Let $\mathbf{e}_j$ denote the j-th standard basis vector in $\mathbb{R}^n$. Suppose that for each input coordinate $j \in \{1, 2, \ldots, n\}$ there exists functions $K^j(C_j, \rho_j)$, $C^j(C_j, \rho_j)$, $P^j(C_j, \rho_j)$ s.t. for every sequence $\{\boldsymbol{u}_k\}_{k=1}^{\infty} \subset \mathcal{D}$ converging to $\boldsymbol{u}_0$ where the difference $\boldsymbol{u}_\ell - \boldsymbol{u}_{\ell'}$ is parallel to $\boldsymbol{e}_j$ (i.e., they only differ in the j-th coordinate) that satisfies $|\boldsymbol{u}_0 - \boldsymbol{u}_k| \leq C_j \rho_j^k$ for all k and some $\rho_j \in [0, 1)$, $C_j > 0$, we have for all $k \geq K^j(C_j, \rho_j)$:

$$\|f(\boldsymbol{u}_0) - f(\boldsymbol{u}_k)\| \leq C^j(C_j, \rho_j) \rho_j^k k^{P^j(C_j, \rho_j)} \quad .$$

Then, there exist constants $C > 0$ and $\rho \in [0, 1)$ such that for all sufficiently large k:

$$\|f(\boldsymbol{x}_0) - f(\boldsymbol{x}_k)\| \leq C \rho^k \quad .$$

Proof. Let the sequence $\{\boldsymbol{x}_k\}_{k=1}^{\infty} \subset \mathcal{D}$ converge exponentially to $\boldsymbol{x}_0 \in \mathcal{D}$. By definition, there exist constants $C_x > 0$ and $\rho \in [0, 1)$ such that for all k,

$$\|\boldsymbol{x}_k - \boldsymbol{x}_0\| \leq C_x \rho^k \quad .$$

Let $\boldsymbol{x}_k = (x_{k,1}, \ldots, x_{k,n})^T$ and $\boldsymbol{x}_0 = (x_{0,1}, \ldots, x_{0,n})^T$. An immediate consequence is that each coordinate also converges exponentially, i.e., for each $j \in \{1, \ldots, n\}$:

$$|x_{k,j} - x_{0,j}| \leq \|\boldsymbol{x}_k - \boldsymbol{x}_0\|_\infty \leq \|\boldsymbol{x}_k - \boldsymbol{x}_0\| \leq C_x \rho^k \quad ,$$

where we use the equivalence of norms in $\mathbb{R}^n$.

To bound $\|f(\boldsymbol{x}_0) - f(\boldsymbol{x}_k)\|$, we define a sequence of $n+1$ intermediate points that form a path from $\boldsymbol{x}_k$ to $\boldsymbol{x}_0$ by changing one coordinate at a time. For each k, let:

$$\begin{aligned}
\boldsymbol{z}_{k,0} &:= \boldsymbol{x}_k = (x_{k,1}, x_{k,2}, \ldots, x_{k,n}) \\
\boldsymbol{z}_{k,1} &:= (x_{0,1}, x_{k,2}, \ldots, x_{k,n}) \\
&\vdots \\
\boldsymbol{z}_{k,j} &:= (x_{0,1}, \ldots, x_{0,j}, x_{k,j+1}, \ldots, x_{k,n}) \\
&\vdots \\
\boldsymbol{z}_{k,n} &:= (x_{0,1}, \ldots, x_{0,n}) = \boldsymbol{x}_0 \quad .
\end{aligned}$$

Since $\mathcal{D}$ is a cartesian product intervals and both $\boldsymbol{x}_k$ and $\boldsymbol{x}_0$ are in $\mathcal{D}$, all intermediate points $\boldsymbol{z}_{k,j}$ are also contained in $\mathcal{D}$. We can express the total difference $f(\boldsymbol{x}_0) - f(\boldsymbol{x}_k)$ as a telescoping sum:

$$f(\boldsymbol{x}_0) - f(\boldsymbol{x}_k) = f(\boldsymbol{z}_{k,n}) - f(\boldsymbol{z}_{k,0}) = \sum_{j=1}^{n} \left(f(\boldsymbol{z}_{k,j}) - f(\boldsymbol{z}_{k,j-1})\right) \quad .$$

By the triangle inequality, we have:

$$\|f(\boldsymbol{x}_0) - f(\boldsymbol{x}_k)\| \leq \sum_{j=1}^{n} \|f(\boldsymbol{z}_{k,j}) - f(\boldsymbol{z}_{k,j-1})\| \quad .$$

Now, we analyze each term $\|f(\boldsymbol{z}_{k,j}) - f(\boldsymbol{z}_{k,j-1})\|$ for a fixed $j \in \{1, \dots, n\}$. The points $\boldsymbol{z}_{k,j}$ and $\boldsymbol{z}_{k,j-1}$ differ only in their j-th coordinate.

Let us define a sequence $\{\boldsymbol{u}_m\}_{m=1}^{\infty}$ and a limit point $\boldsymbol{u}_0$ that fit the condition in the theorem's hypothesis. For the given j and k, let

$$\begin{aligned} \boldsymbol{u}_m &:= (x_{0,1}, \dots, x_{0,j-1}, x_{m,j}, x_{k,j+1}, \dots, x_{k,n}) \\ \boldsymbol{u}_0 &:= (x_{0,1}, \dots, x_{0,j-1}, x_{0,j}, x_{k,j+1}, \dots, x_{k,n}) \quad . \end{aligned}$$

Note that $\boldsymbol{u}_0 = \boldsymbol{z}_{k,j}$ and by setting $m = k$, we get $\boldsymbol{u}_k = \boldsymbol{z}_{k,j-1}$. The sequence $\{\boldsymbol{u}_m\}$ lies on a line parallel to the j-th coordinate axis. As $m \to \infty$, $\boldsymbol{u}_m \to \boldsymbol{u}_0$ because $x_{m,j} \to x_{0,j}$. The convergence is exponential:

$$\|\boldsymbol{u}_m - \boldsymbol{u}_0\| = |x_{m,j} - x_{0,j}| \leq C_x \rho^m \quad .$$

The hypothesis states that for any such sequence, there exist constants $K^j(C_x, \rho)$, $C^j(C_x, \rho)$, $P^j(C_x, \rho)$ which are independent of the specific line, such that for all $k \geq K^j(C_x, \rho)$ we have $\|f(\boldsymbol{u}_0) - f(\boldsymbol{u}_m)\| \leq C^j(C_x, \rho)\rho^m m^{P^j(C_x,\rho)}$. Applying this for $m = k$:

$$\|f(\boldsymbol{z}_{k,j}) - f(\boldsymbol{z}_{k,j-1})\| = \|f(\boldsymbol{u}_0) - f(\boldsymbol{u}_k)\| \leq C^j(C_x, \rho)\rho^k k^{P^j(C_x,\rho)} \quad .$$

This inequality holds for each $j = 1, \dots, n$. Substituting these bounds back into the sum:

$$\|f(\boldsymbol{x}_0) - f(\boldsymbol{x}_k)\| \leq \sum_{j=1}^{n} C^j(C_x, \rho)\rho^k k^{P^j(C_x,\rho)} \quad .$$

Let $K := \max_{j\in\{1,\dots,n\}}\{K^j(C_x, \rho)\}$, $C := \sum_{j=1}^{n} C^j(C_x, \rho)$ and
$P := \max_{j\in\{1,\dots,n\}}\{P^j(C_x, \rho)\}$. Hence, for all $k \geq K$:

$$\|f(\boldsymbol{x}_0) - f(\boldsymbol{x}_k)\| \leq \sum_{j=1}^{n} C^j(C_x, \rho)\rho^k k^P = \left(\sum_{j=1}^{n} C^j(C_x, \rho)\right)\rho^k k^P = C\rho^k k^P \quad .$$

This shows that $\{f(\boldsymbol{x}_k)\}$ converges exponentially to $f(\boldsymbol{x}_0)$, which completes the proof. $\square$

Lemma 1.4. *Let the function $f : [0,1] \to \mathbb{R}$ be defined as $f(x) = x \log x$, with the convention $f(0) = 0$. If a sequence $\{x_k\}_{k=1}^{\infty} \subset [0,1]$ converging to a limit $x_\infty \in [0,1]$ satisfies $|x_k - x_\infty| \le C\rho^k$ for some $C \in \mathbb{R}_{>0}$, $\rho \in [0,1)$, then the sequence $\{f(x_k)\}$ converges to $f(x_\infty)$ with $|f(x_k) - f(x_\infty)| \le C\rho^k \ |\log C + k \log \rho|$ for $k \ge \log_\rho \left(\frac{e^{-1}}{C}\right) =: K$.*

Proof. We consider two cases for the limit x_∞:

Case 1: $x_\infty = 0$
In this case, $|x_k - 0| = x_k \le C\rho^k$. We want to bound the difference $|f(x_k) - f(0)| = |x_k \log x_k|$. Note that $|x \log x|$ is monotonically increasing on $[0, e^{-1}]$. For $k \ge K$ we have $x_k \le e^{-1}$. Thus, for all $k \ge K$:

$$|x_k \log x_k| \le |C\rho^k \log(C\rho^k)| = C\rho^k \ |\log C + k \log \rho| \quad .$$

Case 2: $x_\infty > 0$
Similarly, for $k \ge K$ we have $|x_k - x_\infty| \le C\rho^k \le e^{-1}$. For $k \ge K$ it follows that:

$$|f(x_k) - f(x_\infty)| \le |f(C\rho^k) - f(0)| = |f(C\rho^k)| = C\rho^k \ |\log C + k \log \rho| \quad .$$

$\square$

Theorem 1.6 (Element-Wise Exponential Convergence Property of Mutual Information)**.** *Let the function $f : [0,1] \to \mathbb{R}$ be defined as $f(x) = x \log x$, with the convention $f(0) = 0$. Define the function $I : [0,1]^{mn} \mapsto \mathbb{R}$ for a matrix $\boldsymbol{M}$ as:*

$$I(\boldsymbol{M}) = \sum_{i=1}^{m} \sum_{j=1}^{n} f(M_{ij}) - \sum_{i=1}^{m} f\left(\sum_{j=1}^{n} M_{ij}\right) - \sum_{j=1}^{n} f\left(\sum_{i=1}^{m} M_{ij}\right) \quad .$$

This function exhibits element-wise exponential convergence. That is, for any single component (i_0, j_0), if a sequence of matrices $\{\boldsymbol{U}_k\}_{k=1}^{\infty} \subset [0,1]^{mn}$ converges to a limit $\boldsymbol{U}_\infty$, varies only in the (i_0, j_0)-th component and satisfies $\|\boldsymbol{U}_k - \boldsymbol{U}_\infty\| \le C\rho^k$ for some $C > 0$, $\rho \in [0,1)$ and all k, then the sequence of values $\{I(\boldsymbol{U}_k)\}$ converges to $I(\boldsymbol{U}_\infty)$ with $|I(\boldsymbol{U}_k) - I(\boldsymbol{U}_\infty)| \le C'(C,\rho)\rho^k n^{P(C,\rho)}$ for all $k \ge K(C,\rho)$.

Proof. The function $I(\boldsymbol{M})$ is a sum of terms involving f applied to the matrix entries and their row and column sums. Let $m_i(\boldsymbol{M}) = \sum_j M_{ij}$ and $m'_j(\boldsymbol{M}) = \sum_i M_{ij}$. We have:

$$I(\boldsymbol{M}) = \sum_{i,j} f(M_{ij}) - \sum_i f(m_i(\boldsymbol{M})) - \sum_j f(m'_j(\boldsymbol{M})) \quad .$$

We are given a sequence $\{\boldsymbol{U}_k\}$ that varies only in the (i_0, j_0)-th component, $u_k = U_k(i_0, j_0)$. All other components are constant. The exponential convergence of $\{\boldsymbol{U}_k\}$ means $|u_k - u_\infty| \leq C\rho^k$.

The difference $I(\boldsymbol{U}_k) - I(\boldsymbol{U}_\infty)$ consists only of terms whose arguments change with k. These are:

1. The entry term: $f(u_k)$.
2. The row-sum term: $f(m_{i_0}(\boldsymbol{U}_k))$, where $m_{i_0}(\boldsymbol{U}_k) = u_k + \text{const}$.
3. The column-sum term: $f(m'_{j_0}(\boldsymbol{U}_k))$, where $m'_{j_0}(\boldsymbol{U}_k) = u_k + \text{const}$.

By the triangle inequality, the total error is bounded by the sum of the absolute errors of these three terms:

$$\begin{aligned} |I(\boldsymbol{U}_k) - I(\boldsymbol{U}_\infty)| \leq\ & |f(u_k) - f(u_\infty)| \\ & + |f(m_{i_0}(\boldsymbol{U}_k)) - f(m_{i_0}(\boldsymbol{U}_\infty))| \\ & + |f(m'_{j_0}(\boldsymbol{U}_k)) - f(m'_{j_0}(\boldsymbol{U}_\infty))| \quad . \end{aligned}$$

The arguments to the function f in each of these three terms converge exponentially to their limits with rate ρ and constant C, since $|m_{i_0}(\boldsymbol{U}_k) - m_{i_0}(\boldsymbol{U}_\infty)| = |u_k - u_\infty|$ and $|m'_{j_0}(\boldsymbol{U}_k) - m'_{j_0}(\boldsymbol{U}_\infty)| = |u_k - u_\infty|$.

Hence, by lemma 1.4, we have:

$$\begin{aligned} |I(\boldsymbol{U}_k) - I(\boldsymbol{U}_\infty)| &\leq 3C\rho^k \ |\log C + k \log \rho| \\ &\leq 3C\rho^k k \ (|\log C| + |\log \rho|) \\ &\leq C'\rho^k k \quad , \end{aligned}$$

with $C' := 3C(|\log C| + |\log \rho|)$ and for all $k \geq K(C, \rho)$. Note that K, C' and P only depend on C and ρ. ☐

Corollary 1.8. *Using theorem 1.5, we see that if a sequence $\{\boldsymbol{P}_k\}$ of joint probability distributions converges exponentially fast, then $\{I(\boldsymbol{P}_k)\}$ converges exponentially fast as well.*

Corollary 1.9. *A joint probability matrix $\boldsymbol{P}_{X,Y}$ can be calculated from the conditional probability matrix $\boldsymbol{P}_{Y|X}$ and the diagonal matrix $\boldsymbol{P}_X$ with the probabilities for X on its diagonal using $\boldsymbol{P}_{X,Y} = \boldsymbol{P}_{Y|X}\boldsymbol{P}_X$. Hence, if $\boldsymbol{P}_{Y|X}$ converges exponentially fast while $\boldsymbol{P}_X$ stays constant, the mutual information $I(X;Y)$ will converge exponentially fast as well.*