Previously...

- Joint typical sequences
- Covering and Packing Lemmas
- Channel Coding Theorem
- Capacity of Gaussian channel
- Capacity of additive white Gaussian channel
- Forward proof of Channel Coding Theorem

This time

- Converse Proof of Channel Coding Theorem
- Non-white Gaussian Channel
- Rate-distortion problems
- Rate-distortion Theorem

Converse proof

We want to say that whenever the code rate is larger than the capacity, the probability of error will be non-zero

Converse proof

We want to say that whenever the code rate is larger than the capacity, the probability of error will be non-zero

Equivalently...

As long as the probability of error is 0, the rate of the code R has to be larger than the capacity

Converse proof

We want to say that whenever the code rate is larger than the capacity, the probability of error will be non-zero

Equivalently...

As long as the probability of error is 0, the rate of the code R has to be larger than the capacity

To continue the converse proof, we will need to introduce a simple result from Fano

Fano's inequality

Fano's inequality

Denote $Pr(error) = P_e = Pr(M \neq \hat{M})$, then $H(M|Y^N) \leq 1 + P_e H(M)$
Intuitively, if $P_e \to 0$, on average we will know M for certain given y and thus $\frac{1}{N} H(M|Y^N) \to 0$

Fano's inequality

Fano's inequality

Denote $Pr(error) = P_e = Pr(M \neq \hat{M})$, then $H(M|Y^N) \leq 1 + P_e H(M)$
Intuitively, if $P_e \to 0$, on average we will know M for certain given y and thus $\frac{1}{N} H(M|Y^N) \to 0$

Proof: Let $E = I(M \neq \hat{M})$, then

Fano's inequality

Fano's inequality

Denote $Pr(error) = P_e = Pr(M \neq \hat{M})$, then $H(M|Y^N) \leq 1 + P_e H(M)$
Intuitively, if $P_e \to 0$, on average we will know M for certain given y and thus $\frac{1}{N} H(M|Y^N) \to 0$

Proof: Let $E = I(M \neq \hat{M})$, then

$$H(M|Y^N) = H(M, E|Y^N) - H(E|Y^N, M)$$

Fano's inequality

Fano's inequality

Denote $Pr(error) = P_e = Pr(M \neq \hat{M})$, then $H(M|Y^N) \leq 1 + P_e H(M)$
Intuitively, if $P_e \to 0$, on average we will know M for certain given y and thus $\frac{1}{N}H(M|Y^N) \to 0$

Proof: Let $E = I(M \neq \hat{M})$, then

$$H(M|Y^N) = H(M,E|Y^N) - H(E|Y^N,M)$$
$$= H(M,E|Y^N) = H(E|Y^N) + H(M|Y^N,E)$$

Fano's inequality

Fano's inequality

Denote $Pr(error) = P_e = Pr(M \neq \hat{M})$, then $H(M|Y^N) \leq 1 + P_e H(M)$
Intuitively, if $P_e \to 0$, on average we will know M for certain given y and thus $\frac{1}{N} H(M|Y^N) \to 0$

Proof: Let $E = I(M \neq \hat{M})$, then

$$
\begin{aligned}
H(M|Y^N) &= H(M,E|Y^N) - H(E|Y^N,M) \\
&= H(M,E|Y^N) = H(E|Y^N) + H(M|Y^N,E) \\
&\leq H(E) + H(M|Y^N,E)
\end{aligned}
$$

Fano's inequality

Fano's inequality

Denote $Pr(error) = P_e = Pr(M \neq \hat{M})$, then $H(M|Y^N) \leq 1 + P_e H(M)$
Intuitively, if $P_e \to 0$, on average we will know M for certain given y and thus $\frac{1}{N}H(M|Y^N) \to 0$

Proof: Let $E = I(M \neq \hat{M})$, then

$$
\begin{aligned}
H(M|Y^N) &= H(M,E|Y^N) - H(E|Y^N,M) \\
&= H(M,E|Y^N) = H(E|Y^N) + H(M|Y^N,E) \\
&\leq H(E) + H(M|Y^N,E) \\
&\leq 1 + P(E=0)H(M|Y^N,E=0) + P(E=1)H(M|Y^N,E=1)
\end{aligned}
$$

Fano's inequality

Fano's inequality

Denote $Pr(error) = P_e = Pr(M \neq \hat{M})$, then $H(M|Y^N) \leq 1 + P_e H(M)$
Intuitively, if $P_e \to 0$, on average we will know M for certain given y and thus $\frac{1}{N} H(M|Y^N) \to 0$

Proof: Let $E = I(M \neq \hat{M})$, then

$$
\begin{aligned}
H(M|Y^N) &= H(M, E|Y^N) - H(E|Y^N, M) \\
&= H(M, E|Y^N) = H(E|Y^N) + H(M|Y^N, E) \\
&\leq H(E) + H(M|Y^N, E) \\
&\leq 1 + P(E=0)H(M|Y^N, E=0) + P(E=1)H(M|Y^N, E=1) \\
&\leq 1 + 0 + P_e H(M|Y^N, E=1) \overset{(d)}{\leq} 1 + P_e H(M)
\end{aligned}
$$

Converse proof

$$R = \frac{H(M)}{N} = \frac{1}{N}\left[I(M;Y^N) + H(M|Y^N)\right]$$

Converse proof

$$R = \frac{H(M)}{N} = \frac{1}{N}\left[I(M;Y^N) + H(M|Y^N)\right] \le \frac{1}{N}\left[I(X^N;Y^N) + H(M|Y^N)\right]$$

Converse proof

$$R = \frac{H(M)}{N} = \frac{1}{N}\left[I(M;Y^N) + H(M|Y^N)\right] \leq \frac{1}{N}\left[I(X^N;Y^N) + H(M|Y^N)\right]$$
$$= \frac{1}{N}\left[H(Y^N) - H(Y^N|X^N) + H(M|Y^N)\right]$$

Converse proof

$$R = \frac{H(M)}{N} = \frac{1}{N}\left[I(M;Y^N) + H(M|Y^N)\right] \leq \frac{1}{N}\left[I(X^N;Y^N) + H(M|Y^N)\right]$$
$$= \frac{1}{N}\left[H(Y^N) - H(Y^N|X^N) + H(M|Y^N)\right]$$
$$= \frac{1}{N}\left[H(Y^N) - \sum_i H(Y_i|X^N, Y^{i-1}) + H(M|Y^N)\right]$$

Converse proof

$$R = \frac{H(M)}{N} = \frac{1}{N}\left[I(M;Y^N) + H(M|Y^N)\right] \le \frac{1}{N}\left[I(X^N;Y^N) + H(M|Y^N)\right]$$

$$= \frac{1}{N}\left[H(Y^N) - H(Y^N|X^N) + H(M|Y^N)\right]$$

$$= \frac{1}{N}\left[H(Y^N) - \sum_i H(Y_i|X^N, Y^{i-1}) + H(M|Y^N)\right]$$

$$= \frac{1}{N}\left[H(Y^N) - \sum_i H(Y_i|X_i) + H(M|Y^N)\right]$$

Converse proof

$$R = \frac{H(M)}{N} = \frac{1}{N}\left[I(M;Y^N) + H(M|Y^N)\right] \le \frac{1}{N}\left[I(X^N;Y^N) + H(M|Y^N)\right]$$

$$= \frac{1}{N}\left[H(Y^N) - H(Y^N|X^N) + H(M|Y^N)\right]$$

$$= \frac{1}{N}\left[H(Y^N) - \sum_i H(Y_i|X^N, Y^{i-1}) + H(M|Y^N)\right]$$

$$= \frac{1}{N}\left[H(Y^N) - \sum_i H(Y_i|X_i) + H(M|Y^N)\right]$$

$$\le \frac{1}{N}\left[\sum_i H(Y_i) - \sum_i H(Y_i|X_i) + H(M|Y^N)\right]$$

Converse proof

$$R = \frac{H(M)}{N} = \frac{1}{N}\left[I(M;Y^N) + H(M|Y^N)\right] \leq \frac{1}{N}\left[I(X^N;Y^N) + H(M|Y^N)\right]$$

$$= \frac{1}{N}\left[H(Y^N) - H(Y^N|X^N) + H(M|Y^N)\right]$$

$$= \frac{1}{N}\left[H(Y^N) - \sum_i H(Y_i|X^N, Y^{i-1}) + H(M|Y^N)\right]$$

$$= \frac{1}{N}\left[H(Y^N) - \sum_i H(Y_i|X_i) + H(M|Y^N)\right]$$

$$\leq \frac{1}{N}\left[\sum_i H(Y_i) - \sum_i H(Y_i|X_i) + H(M|Y^N)\right]$$

$$= \frac{1}{N}\left[\sum_i I(X_i;Y_i) + H(M|Y^N)\right] = I(X;Y) + \frac{H(M|Y^N)}{N}$$

Converse proof

$$R = \frac{H(M)}{N} = \frac{1}{N}\left[I(M;Y^N) + H(M|Y^N)\right] \leq \frac{1}{N}\left[I(X^N;Y^N) + H(M|Y^N)\right]$$

$$= \frac{1}{N}\left[H(Y^N) - H(Y^N|X^N) + H(M|Y^N)\right]$$

$$= \frac{1}{N}\left[H(Y^N) - \sum_i H(Y_i|X^N, Y^{i-1}) + H(M|Y^N)\right]$$

$$= \frac{1}{N}\left[H(Y^N) - \sum_i H(Y_i|X_i) + H(M|Y^N)\right]$$

$$\leq \frac{1}{N}\left[\sum_i H(Y_i) - \sum_i H(Y_i|X_i) + H(M|Y^N)\right]$$

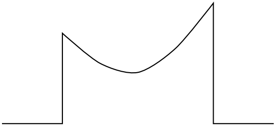
$$= \frac{1}{N}\left[\sum_i I(X_i;Y_i) + H(M|Y^N)\right] = I(X;Y) + \frac{H(M|Y^N)}{N} \to I(X;Y)$$

as $N \to \infty$ by Fano's inequality

Color channels

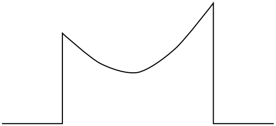
- We look into capacity of white Gaussian channel last time

Color channels



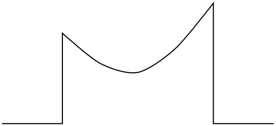
- We look into capacity of white Gaussian channel last time
- But sometimes noise power can be different for different band, consequently, "color" channels

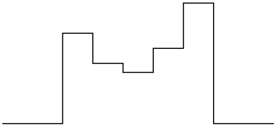
Color channels



- We look into capacity of white Gaussian channel last time
- But sometimes noise power can be different for different band, consequently, "color" channels
- Intuitively, we should assign different amount of power to different band. Hence, we have an allocation problem

Color channels





- We look into capacity of white Gaussian channel last time
- But sometimes noise power can be different for different band, consequently, "color" channels
- Intuitively, we should assign different amount of power to different band. Hence, we have an allocation problem
- Without loss of generality, let's consider the discrete approximation, parallel Gaussian channel

Parallel Gaussian channels

- Consider that we have K parallel channels (K bands) and the corresponding noise powers are $\sigma_1^2, \sigma_2^2, \cdots, \sigma_K^2$

Parallel Gaussian channels

- Consider that we have K parallel channels (K bands) and the corresponding noise powers are $\sigma_1^2, \sigma_2^2, \cdots, \sigma_K^2$
- And say, we can allocate a total of P power to all channels. The powers assigned to the channels are $P_1, P_2, \cdots, P_K$. So we need $\sum_{i=1}^K P_i \leq P$

Parallel Gaussian channels

- Consider that we have K parallel channels (K bands) and the corresponding noise powers are $\sigma_1^2, \sigma_2^2, \cdots, \sigma_K^2$
- And say, we can allocate a total of P power to all channels. The powers assigned to the channels are $P_1, P_2, \cdots, P_K$. So we need $\sum_{i=1}^K P_i \le P$
- Therefore, for the k-th channel, we can transmit $\frac{1}{2}\log\left(1+\frac{P_k}{\sigma_k^2}\right)$ bits per channel use

Parallel Gaussian channels

- Consider that we have K parallel channels (K bands) and the corresponding noise powers are $\sigma_1^2, \sigma_2^2, \cdots, \sigma_K^2$
- And say, we can allocate a total of P power to all channels. The powers assigned to the channels are $P_1, P_2, \cdots, P_K$. So we need $\sum_{i=1}^K P_i \leq P$
- Therefore, for the k-th channel, we can transmit $\frac{1}{2}\log\left(1+\frac{P_k}{\sigma_k^2}\right)$ bits per channel use
- So our goal is to assign $P_1, P_2, \cdots, P_K \geq 0$ ($\sum_{k=1}^K P_k \leq P$) such that the total capacity

$$\sum_{k=1}^K \frac{1}{2}\log\left(1+\frac{P_k}{\sigma_k^2}\right)$$

is maximize

KKT conditions

Let's list all the KKT conditions for the optimization problem

$$\max \sum_{k=1}^{K} \frac{1}{2} \log\left(1+\frac{P_k}{\sigma_k^2}\right) \qquad \text{such that}$$

$$P_1, \cdots, P_K \geq 0, \qquad \sum_{k=1}^{K} P_k \leq P$$

$$\frac{\partial}{\partial P_i}\left[\sum_{k=1}^{K} \frac{1}{2} \log\left(1+\frac{P_k}{\sigma_k^2}\right) + \sum_{k=1}^{K} \lambda_k P_k - \mu\left(\sum_{k=1}^{K} P_k - P\right)\right] = 0$$

KKT conditions

Let's list all the KKT conditions for the optimization problem

$$\max \sum_{k=1}^{K} \frac{1}{2} \log \left(1 + \frac{P_k}{\sigma_k^2}\right) \qquad \text{such that}$$

$$P_1, \cdots, P_K \geq 0, \qquad \sum_{k=1}^{K} P_k \leq P$$

$$\frac{\partial}{\partial P_i}\left[\sum_{k=1}^{K} \frac{1}{2} \log \left(1 + \frac{P_k}{\sigma_k^2}\right) + \sum_{k=1}^{K} \lambda_k P_k - \mu \left(\sum_{k=1}^{K} P_k - P\right)\right] = 0$$

$$\mu, \lambda_1, \cdots, \lambda_K \geq 0$$

KKT conditions

Let's list all the KKT conditions for the optimization problem

$$\max \sum_{k=1}^{K} \frac{1}{2} \log\left(1+\frac{P_k}{\sigma_k^2}\right) \qquad \text{such that}$$

$$P_1, \cdots, P_K \geq 0, \qquad \sum_{k=1}^{K} P_k \leq P$$

$$\frac{\partial}{\partial P_i}\left[\sum_{k=1}^{K} \frac{1}{2} \log\left(1+\frac{P_k}{\sigma_k^2}\right) + \sum_{k=1}^{K} \lambda_k P_k - \mu\left(\sum_{k=1}^{K} P_k - P\right)\right] = 0$$

$$\mu, \lambda_1, \cdots, \lambda_K \geq 0, P_1, \cdots, P_K \geq 0, \sum_{k=1}^{K} P_k \leq P$$

KKT conditions

Let's list all the KKT conditions for the optimization problem

$$\max \sum_{k=1}^{K} \frac{1}{2} \log\left(1+\frac{P_k}{\sigma_k^2}\right) \qquad \text{such that}$$

$$P_1, \cdots, P_K \geq 0, \qquad \sum_{k=1}^{K} P_k \leq P$$

$$\frac{\partial}{\partial P_i}\left[\sum_{k=1}^{K} \frac{1}{2} \log\left(1+\frac{P_k}{\sigma_k^2}\right) + \sum_{k=1}^{K} \lambda_k P_k - \mu\left(\sum_{k=1}^{K} P_k - P\right)\right] = 0$$

$$\mu, \lambda_1, \cdots, \lambda_K \geq 0, P_1, \cdots, P_K \geq 0, \sum_{k=1}^{K} P_k \leq P$$

$$\mu\left(\sum_{k=1}^{K} P_k - P\right) = 0, \qquad \lambda_k P_k = 0, \forall k$$

Capacity of parallel channels

$$\frac{\partial}{\partial P_i}\left[\sum_{k=1}^{K}\frac{1}{2}\log\left(1+\frac{P_k}{\sigma_k^2}\right)+\sum_{k=1}^{K}\lambda_k P_k-\mu\left(\sum_{k=1}^{K}P_k-P\right)\right]=0$$

Capacity of parallel channels

$$\frac{\partial}{\partial P_i}\left[\sum_{k=1}^{K}\frac{1}{2}\log\left(1+\frac{P_k}{\sigma_k^2}\right)+\sum_{k=1}^{K}\lambda_k P_k-\mu\left(\sum_{k=1}^{K}P_k-P\right)\right]=0$$

$$\Rightarrow\frac{1}{2}\frac{1}{P_i+\sigma_i^2}=\mu-\lambda_i$$

Capacity of parallel channels

$$\frac{\partial}{\partial P_i}\left[\sum_{k=1}^{K}\frac{1}{2}\log\left(1+\frac{P_k}{\sigma_k^2}\right)+\sum_{k=1}^{K}\lambda_k P_k-\mu\left(\sum_{k=1}^{K}P_k-P\right)\right]=0$$

$$\Rightarrow\frac{1}{2}\frac{1}{P_i+\sigma_i^2}=\mu-\lambda_i\Rightarrow P_i+\sigma_i^2=\frac{1}{2(\mu-\lambda_i)}$$

Capacity of parallel channels

$$\frac{\partial}{\partial P_i}\left[\sum_{k=1}^{K}\frac{1}{2}\log\left(1+\frac{P_k}{\sigma_k^2}\right)+\sum_{k=1}^{K}\lambda_k P_k-\mu\left(\sum_{k=1}^{K}P_k-P\right)\right]=0$$

$$\Rightarrow\frac{1}{2}\frac{1}{P_i+\sigma_i^2}=\mu-\lambda_i\Rightarrow P_i+\sigma_i^2=\frac{1}{2(\mu-\lambda_i)}$$

Since $\lambda_i P_i = 0$, for $P_i > 0$, we have $\lambda_i = 0$ and thus

$$P_i+\sigma_i^2=\frac{1}{2\mu}$$

Capacity of parallel channels

$$\frac{\partial}{\partial P_i}\left[\sum_{k=1}^{K}\frac{1}{2}\log\left(1+\frac{P_k}{\sigma_k^2}\right)+\sum_{k=1}^{K}\lambda_k P_k-\mu\left(\sum_{k=1}^{K}P_k-P\right)\right]=0$$

$$\Rightarrow\frac{1}{2}\frac{1}{P_i+\sigma_i^2}=\mu-\lambda_i\Rightarrow P_i+\sigma_i^2=\frac{1}{2(\mu-\lambda_i)}$$

Since $\lambda_i P_i=0$, for $P_i>0$, we have $\lambda_i=0$ and thus

$$P_i+\sigma_i^2=\frac{1}{2\mu}$$

This suggests that $\mu>0$ and thus $\sum_{k=1}^{K}P_k=P$

Capacity of parallel channels

$$\frac{\partial}{\partial P_i}\left[\sum_{k=1}^{K}\frac{1}{2}\log\left(1+\frac{P_k}{\sigma_k^2}\right)+\sum_{k=1}^{K}\lambda_k P_k-\mu\left(\sum_{k=1}^{K}P_k-P\right)\right]=0$$

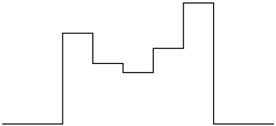
$$\Rightarrow\frac{1}{2}\frac{1}{P_i+\sigma_i^2}=\mu-\lambda_i\Rightarrow P_i+\sigma_i^2=\frac{1}{2(\mu-\lambda_i)}$$

Since $\lambda_i P_i = 0$, for $P_i > 0$, we have $\lambda_i = 0$ and thus

$$P_i+\sigma_i^2=\frac{1}{2\mu}=constant$$

This suggests that $\mu > 0$ and thus $\sum_{k=1}^{K} P_k = P$

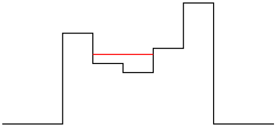
Water-filling interpretation



From $P_i + \sigma_i^2 = const$, power can be allocated intuitively as filling water to a pond (hence "water-filling")

Example

Water-filling interpretation

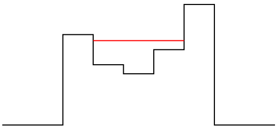


From $P_i + \sigma_i^2 = const$, power can be allocated intuitively as filling water to a pond (hence "water-filling")

Example

- $P_1 = 0, P_2 = 0.3, P_3 = 0.6, P_4 = 0, P_5 = 0$

Water-filling interpretation

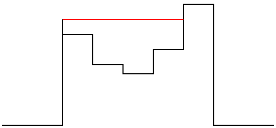


From $P_i + \sigma_i^2 = const$, power can be allocated intuitively as filling water to a pond (hence "water-filling")

Example

- $P_1 = 0, P_2 = 0.3, P_3 = 0.6, P_4 = 0, P_5 = 0$
- $P_1 = 0, P_2 = 0.8, P_3 = 1.1, P_4 = 0.3, P_5 = 0$

Water-filling interpretation

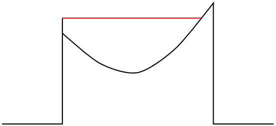


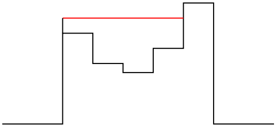
From $P_i + \sigma_i^2 = const$, power can be allocated intuitively as filling water to a pond (hence "water-filling")

Example

- $P_1 = 0, P_2 = 0.3, P_3 = 0.6, P_4 = 0, P_5 = 0$
- $P_1 = 0, P_2 = 0.8, P_3 = 1.1, P_4 = 0.3, P_5 = 0$
- $P_1 = 0.5, P_2 = 1.5, P_3 = 1.8, P_4 = 1, P_5 = 0$

Water-filling interpretation



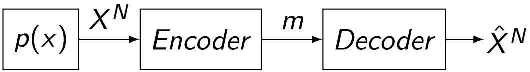


From $P_i + \sigma_i^2 = const$, power can be allocated intuitively as filling water to a pond (hence "water-filling")

Example

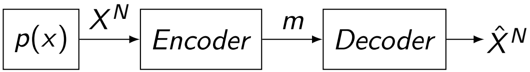
- $P_1 = 0, P_2 = 0.3, P_3 = 0.6, P_4 = 0, P_5 = 0$
- $P_1 = 0, P_2 = 0.8, P_3 = 1.1, P_4 = 0.3, P_5 = 0$
- $P_1 = 0.5, P_2 = 1.5, P_3 = 1.8, P_4 = 1, P_5 = 0$

Rate-distortion problem



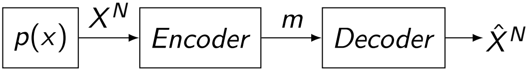
- We know that $H(X)$ bits are needed on average to represent each sample of a source X

Rate-distortion problem



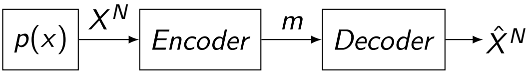
- We know that $H(X)$ bits are needed on average to represent each sample of a source X
- If X is continuous, there is no way to recover X precisely

Rate-distortion problem



- We know that $H(X)$ bits are needed on average to represent each sample of a source X
- If X is continuous, there is no way to recover X precisely
- Let say we are satisfied as long as we can recover X up to certain fidelity, how many bits are needed per sample?

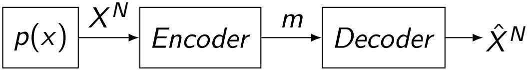
Rate-distortion problem



- We know that $H(X)$ bits are needed on average to represent each sample of a source X
- If X is continuous, there is no way to recover X precisely
- Let say we are satisfied as long as we can recover X up to certain fidelity, how many bits are needed per sample?
- There is an apparent rate (bits per sample) and distortion (fidelity) trade-off. We expect that needed rate is smaller if we allow a lower fidelity (higher distortion). What we are really interested in is a rate-distortion function

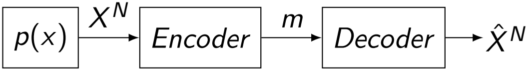
Rate-distortion function

$$m \in \{1, 2, \cdots, M\}$$



Rate-distortion function

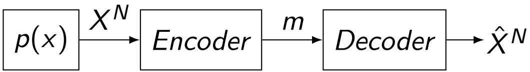
$$m \in \{1, 2, \cdots, M\}$$



$$R = \frac{\log M}{N}, \qquad D = E[d(\hat{X}^N, X^N)] = \frac{1}{N}\sum_{i=1}^{N} d(\hat{X}_i, X_i)$$

Rate-distortion function

$$m \in \{1, 2, \cdots, M\}$$

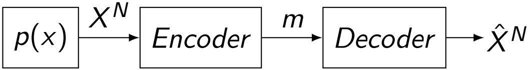


$$R = \frac{\log M}{N}, \qquad D = E[d(\hat{X}^N, X^N)] = \frac{1}{N}\sum_{i=1}^{N} d(\hat{X}_i, X_i)$$

- Maybe you can guess at this point. For given X and $\hat{X}$, the required rate is simply $I(X; \hat{X})$

Rate-distortion function

$$m \in \{1, 2, \cdots, M\}$$

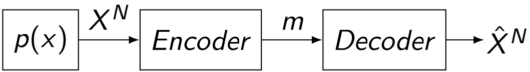


$$R = \frac{\log M}{N}, \qquad D = E[d(\hat{X}^N, X^N)] = \frac{1}{N}\sum_{i=1}^{N} d(\hat{X}_i, X_i)$$

- Maybe you can guess at this point. For given X and $\hat{X}$, the required rate is simply $I(X;\hat{X})$
- How is it related to the distortion though?

Rate-distortion function

$$m \in \{1, 2, \cdots, M\}$$

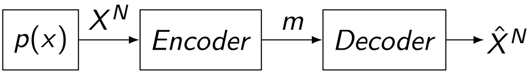


$$R = \frac{\log M}{N}, \qquad D = E[d(\hat{X}^N, X^N)] = \frac{1}{N}\sum_{i=1}^{N} d(\hat{X}_i, X_i)$$

- Maybe you can guess at this point. For given X and $\hat{X}$, the required rate is simply $I(X;\hat{X})$
- How is it related to the distortion though?
- Note that we have a freedom to pick $p(\hat{x}|x)$ such that $E[d(\hat{X}^N, X^N)]$ (less than or) equal to the desired $\mathcal{D}$

Rate-distortion function

$$m \in \{1, 2, \cdots, M\}$$



$$R = \frac{\log M}{N}, \qquad D = E[d(\hat{X}^N, X^N)] = \frac{1}{N}\sum_{i=1}^{N} d(\hat{X}_i, X_i)$$

- Maybe you can guess at this point. For given X and $\hat{X}$, the required rate is simply $I(X; \hat{X})$
- How is it related to the distortion though?
- Note that we have a freedom to pick $p(\hat{x}|x)$ such that $E[d(\hat{X}^N, X^N)]$ (less than or) equal to the desired $\mathcal{D}$
- Therefore given $\mathcal{D}$, the rate-distortion function is simply

$$R(\mathcal{D}) = min_{p(\hat{x}|x)} I(\hat{X}; X)$$

such that $E[d(\hat{X}^N, X^N)] \leq \mathcal{D}$

Binary symmetric source

- Let's try to compress outcome from a fair coin toss

Binary symmetric source

- Let's try to compress outcome from a fair coin toss
- We know that we need 1 bit to compress the outcome losslessly, what if we have only 0.5 bit per sample?

Binary symmetric source

- Let's try to compress outcome from a fair coin toss
- We know that we need 1 bit to compress the outcome losslessly, what if we have only 0.5 bit per sample?
- In this case, we can't losslessly recover the outcome. But how good will we do?

Binary symmetric source

- Let's try to compress outcome from a fair coin toss
- We know that we need 1 bit to compress the outcome losslessly, what if we have only 0.5 bit per sample?
- In this case, we can't losslessly recover the outcome. But how good will we do?
- We need to introduce a distortion measure first. Note that we have two types of errors: taking head as tail and taking tail as head. A natural measure will just weights both error equally

$$d(X=H,\hat{X}=T)=d(X=T,\hat{X}=H)=1$$
$$d(X=H,\hat{X}=H)=d(X=T,\hat{X}=T)=0$$

Binary symmetric source

- Let's try to compress outcome from a fair coin toss
- We know that we need 1 bit to compress the outcome losslessly, what if we have only 0.5 bit per sample?
- In this case, we can't losslessly recover the outcome. But how good will we do?
- We need to introduce a distortion measure first. Note that we have two types of errors: taking head as tail and taking tail as head. A natural measure will just weights both error equally

$$d(X=H,\hat{X}=T)=d(X=T,\hat{X}=H)=1$$
$$d(X=H,\hat{X}=H)=d(X=T,\hat{X}=T)=0$$

- If rate is >1 bit, we know that distortion is 0. How about rate is 0, what distortion suppose to be?

Binary symmetric source

- Let's try to compress outcome from a fair coin toss
- We know that we need 1 bit to compress the outcome losslessly, what if we have only 0.5 bit per sample?
- In this case, we can't losslessly recover the outcome. But how good will we do?
- We need to introduce a distortion measure first. Note that we have two types of errors: taking head as tail and taking tail as head. A natural measure will just weights both error equally

$$d(X=H,\hat{X}=T)=d(X=T,\hat{X}=H)=1$$
$$d(X=H,\hat{X}=H)=d(X=T,\hat{X}=T)=0$$

- If rate is >1 bit, we know that distortion is 0. How about rate is 0, what distortion suppose to be?
- If decoders know nothing, the best bet will be just always decode head (or tail). Then $D=E[d(X,H)]=0.5$

Binary symmetric source

For $0 < D < 0.5$, denote Z as the prediction error such that $X = \hat{X} + Z$.

Binary symmetric source

For $0 < D < 0.5$, denote Z as the prediction error such that $X = \hat{X} + Z$. Note that

$$Pr(Z = 1) = D$$

Binary symmetric source

For $0 < D < 0.5$, denote Z as the prediction error such that $X = \hat{X} + Z$. Note that

$$Pr(Z = 1) = D$$

$$R = min_{p(\hat{x}|x)} I(\hat{X}; X) = min_{p(\hat{x}|x)} H(X) - H(X|\hat{X})$$

Binary symmetric source

For $0 < D < 0.5$, denote Z as the prediction error such that $X = \hat{X} + Z$. Note that

$$Pr(Z = 1) = D$$

$$\begin{aligned} R &= min_{p(\hat{x}|x)} I(\hat{X}; X) = min_{p(\hat{x}|x)} H(X) - H(X|\hat{X}) \\ &= min_{p(\hat{x}|x)} H(X) - H(\hat{X} + Z|\hat{X}) \end{aligned}$$

Binary symmetric source

For $0 < D < 0.5$, denote Z as the prediction error such that $X = \hat{X} + Z$. Note that

$$Pr(Z = 1) = D$$

$$\begin{aligned} R &= min_{p(\hat{x}|x)} I(\hat{X}; X) = min_{p(\hat{x}|x)} H(X) - H(X|\hat{X}) \\ &= min_{p(\hat{x}|x)} H(X) - H(\hat{X} + Z|\hat{X}) \\ &= min_{p(\hat{x}|x)} H(X) - H(Z|\hat{X}) \end{aligned}$$

Binary symmetric source

For $0 < D < 0.5$, denote Z as the prediction error such that $X = \hat{X} + Z$. Note that

$$Pr(Z = 1) = D$$

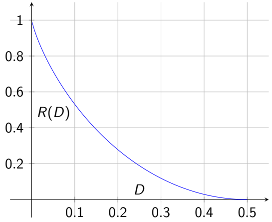
$$\begin{aligned} R &= min_{p(\hat{x}|x)} I(\hat{X}; X) = min_{p(\hat{x}|x)} H(X) - H(X|\hat{X}) \\ &= min_{p(\hat{x}|x)} H(X) - H(\hat{X} + Z|\hat{X}) \\ &= min_{p(\hat{x}|x)} H(X) - H(Z|\hat{X}) \\ &= min_{p(\hat{x}|x)} H(X) - H(Z) \end{aligned}$$

Binary symmetric source

For $0 < D < 0.5$, denote Z as the prediction error such that $X = \hat{X} + Z$. Note that

$$Pr(Z = 1) = D$$

$$\begin{aligned} R &= min_{p(\hat{x}|x)} I(\hat{X}; X) = min_{p(\hat{x}|x)} H(X) - H(X|\hat{X}) \\ &= min_{p(\hat{x}|x)} H(X) - H(\hat{X} + Z|\hat{X}) \\ &= min_{p(\hat{x}|x)} H(X) - H(Z|\hat{X}) \\ &= min_{p(\hat{x}|x)} H(X) - H(Z) \\ &= 1 - H(D) \end{aligned}$$



Gaussian source

- Consider $X \sim \mathcal{N}(0, \sigma_X^2)$. To determine the rate-distortion function, we need first to decide the distortion measure. An intuitive will be just the square error. That is,

$$d(\hat{X}, X) = (\hat{X} - X)^2$$

Gaussian source

- Consider $X \sim \mathcal{N}(0, \sigma_X^2)$. To determine the rate-distortion function, we need first to decide the distortion measure. An intuitive will be just the square error. That is,

$$d(\hat{X}, X) = (\hat{X} - X)^2$$

- Given $E[d(\hat{X}, X)] = D$, what is the minimum rate required?

Gaussian source

- Consider $X \sim \mathcal{N}(0, \sigma_X^2)$. To determine the rate-distortion function, we need first to decide the distortion measure. An intuitive will be just the square error. That is,

$$d(\hat{X}, X) = (\hat{X} - X)^2$$

- Given $E[d(\hat{X}, X)] = D$, what is the minimum rate required?
- Like before, let us denote $Z = X - \hat{X}$ as the prediction error. Note that $Var(Z) = D$

Gaussian source

- Consider $X \sim \mathcal{N}(0, \sigma_X^2)$. To determine the rate-distortion function, we need first to decide the distortion measure. An intuitive will be just the square error. That is,

$$d(\hat{X}, X) = (\hat{X} - X)^2$$

- Given $E[d(\hat{X}, X)] = D$, what is the minimum rate required?
- Like before, let us denote $Z = X - \hat{X}$ as the prediction error. Note that $Var(Z) = D$

$$R(D) = min_{p(\hat{x}|x)} I(\hat{X}; X) = min_{p(\hat{x}|x)} h(X) - h(X|\hat{X})$$

Gaussian source

- Consider $X \sim \mathcal{N}(0, \sigma_X^2)$. To determine the rate-distortion function, we need first to decide the distortion measure. An intuitive will be just the square error. That is,

$$d(\hat{X}, X) = (\hat{X} - X)^2$$

- Given $E[d(\hat{X}, X)] = D$, what is the minimum rate required?
- Like before, let us denote $Z = X - \hat{X}$ as the prediction error. Note that $Var(Z) = D$

$$\begin{aligned} R(D) &= min_{p(\hat{x}|x)} I(\hat{X}; X) = min_{p(\hat{x}|x)} h(X) - h(X|\hat{X}) \\ &= min_{p(\hat{x}|x)} h(X) - h(Z + \hat{X}|\hat{X}) \end{aligned}$$

Gaussian source

- Consider $X \sim \mathcal{N}(0, \sigma_X^2)$. To determine the rate-distortion function, we need first to decide the distortion measure. An intuitive will be just the square error. That is,

$$d(\hat{X}, X) = (\hat{X} - X)^2$$

- Given $E[d(\hat{X}, X)] = D$, what is the minimum rate required?
- Like before, let us denote $Z = X - \hat{X}$ as the prediction error. Note that $Var(Z) = D$

$$\begin{aligned} R(D) &= min_{p(\hat{x}|x)} I(\hat{X}; X) = min_{p(\hat{x}|x)} h(X) - h(X|\hat{X}) \\ &= min_{p(\hat{x}|x)} h(X) - h(Z + \hat{X}|\hat{X}) \\ &= min_{p(\hat{x}|x)} h(X) - h(Z|\hat{X}) \end{aligned}$$

Gaussian source

- Consider $X \sim \mathcal{N}(0, \sigma_X^2)$. To determine the rate-distortion function, we need first to decide the distortion measure. An intuitive will be just the square error. That is,

$$d(\hat{X}, X) = (\hat{X} - X)^2$$

- Given $E[d(\hat{X}, X)] = D$, what is the minimum rate required?
- Like before, let us denote $Z = X - \hat{X}$ as the prediction error. Note that $Var(Z) = D$

$$\begin{aligned} R(D) &= min_{p(\hat{x}|x)} I(\hat{X}; X) = min_{p(\hat{x}|x)} h(X) - h(X|\hat{X}) \\ &= min_{p(\hat{x}|x)} h(X) - h(Z + \hat{X}|\hat{X}) \\ &= min_{p(\hat{x}|x)} h(X) - h(Z|\hat{X}) \\ &= min_{p(\hat{x}|x)} h(X) - h(Z) \end{aligned}$$

Gaussian source

- Consider $X \sim \mathcal{N}(0, \sigma_X^2)$. To determine the rate-distortion function, we need first to decide the distortion measure. An intuitive will be just the square error. That is,

$$d(\hat{X}, X) = (\hat{X} - X)^2$$

- Given $E[d(\hat{X}, X)] = D$, what is the minimum rate required?
- Like before, let us denote $Z = X - \hat{X}$ as the prediction error. Note that $Var(Z) = D$

$$\begin{aligned} R(D) &= min_{p(\hat{x}|x)} I(\hat{X}; X) = min_{p(\hat{x}|x)} h(X) - h(X|\hat{X}) \\ &= min_{p(\hat{x}|x)} h(X) - h(Z + \hat{X}|\hat{X}) \\ &= min_{p(\hat{x}|x)} h(X) - h(Z|\hat{X}) \\ &= min_{p(\hat{x}|x)} h(X) - h(Z) \\ &= \log \frac{\sigma_X^2}{D} \end{aligned}$$

Forward proof

Forward statement

Given distortion constraint $\mathcal{D}$, we can find scheme such that the require rate is no bigger than

$$R(\mathcal{D}) = \min_{p(\hat{x}|x)} I(X;\hat{X}),$$

where the $\hat{X}$ introduced by $p(\hat{x}|x)$ should satisfy $E[d(X,\hat{X})] \leq \mathcal{D}$

Forward proof

Forward statement

Given distortion constraint $\mathcal{D}$, we can find scheme such that the require rate is no bigger than

$$R(\mathcal{D}) = \min_{p(\hat{x}|x)} I(X;\hat{X}),$$

where the $\hat{X}$ introduced by $p(\hat{x}|x)$ should satisfy $E[d(X,\hat{X})] \leq \mathcal{D}$

Code book construction

Let say $p^*(\hat{x}|x)$ is the distribution that achieve the rate-distortion optimiation problem. Randomly construct 2^{NR} codewords as follows

Forward proof

Forward statement

Given distortion constraint $\mathcal{D}$, we can find scheme such that the require rate is no bigger than

$$R(\mathcal{D}) = \min_{p(\hat{x}|x)} I(X;\hat{X}),$$

where the $\hat{X}$ introduced by $p(\hat{x}|x)$ should satisfy $E[d(X,\hat{X})] \leq \mathcal{D}$

Code book construction

Let say $p^*(\hat{x}|x)$ is the distribution that achieve the rate-distortion optimiation problem. Randomly construct 2^{NR} codewords as follows

- Sample X from the source and pass X into $p^*(\hat{x}|x)$ to obtain $\hat{X}$

Forward proof

Forward statement

Given distortion constraint $\mathcal{D}$, we can find scheme such that the require rate is no bigger than

$$R(\mathcal{D}) = \min_{p(\hat{x}|x)} I(X;\hat{X}),$$

where the $\hat{X}$ introduced by $p(\hat{x}|x)$ should satisfy $E[d(X,\hat{X})] \leq \mathcal{D}$

Code book construction

Let say $p^*(\hat{x}|x)$ is the distribution that achieve the rate-distortion optimiation problem. Randomly construct 2^{NR} codewords as follows

- Sample X from the source and pass X into $p^*(\hat{x}|x)$ to obtain $\hat{X}$
- Repeat this N time to get a length-N codeword
- Store the i-th codeword as $\mathbf{C}(i)$

Forward proof

Forward statement

Given distortion constraint $\mathcal{D}$, we can find scheme such that the require rate is no bigger than

$$R(\mathcal{D}) = \min_{p(\hat{x}|x)} I(X;\hat{X}),$$

where the $\hat{X}$ introduced by $p(\hat{x}|x)$ should satisfy $E[d(X,\hat{X})] \leq \mathcal{D}$

Code book construction

Let say $p^*(\hat{x}|x)$ is the distribution that achieve the rate-distortion optimiation problem. Randomly construct 2^{NR} codewords as follows

- Sample X from the source and pass X into $p^*(\hat{x}|x)$ to obtain $\hat{X}$
- Repeat this N time to get a length-N codeword
- Store the i-th codeword as $\mathbf{C}(i)$

Note that the code rate is $\frac{\log 2^{NR}}{N} = R$ as desired

Covering lemma and distortion typical sequences

We say joint typical sequences x^N and $\hat{x}^N$ are distortion typical $((x^N, \hat{x}^N) \in \mathcal{A}^N_{d,\epsilon})$ if $|d(x^N, \hat{x}^N) - E[d(X, \hat{X})]| \leq \epsilon$

Covering lemma and distortion typical sequences

We say joint typical sequences x^N and $\hat{x}^N$ are distortion typical $((x^N, \hat{x}^N) \in \mathcal{A}^N_{d,\epsilon})$ if $|d(x^N, \hat{x}^N) - E[d(X, \hat{X})]| \leq \epsilon$

- By LLN, every pair of sequences sampled from the joint source will virtually be distortion typical

Covering lemma and distortion typical sequences

We say joint typical sequences x^N and $\hat{x}^N$ are distortion typical $((x^N, \hat{x}^N) \in \mathcal{A}^N_{d,\epsilon})$ if $|d(x^N, \hat{x}^N) - E[d(X, \hat{X})]| \leq \epsilon$

- By LLN, every pair of sequences sampled from the joint source will virtually be distortion typical
- Consequently, $(1-\delta)2^{N(H(X,\hat{X})-\epsilon)} \leq |\mathcal{A}^N_{d,\epsilon}| \leq 2^{N(H(X,\hat{X})+\epsilon)}$ as before

Covering lemma and distortion typical sequences

We say joint typical sequences x^N and $\hat{x}^N$ are distortion typical $((x^N, \hat{x}^N) \in \mathcal{A}^N_{d,\epsilon})$ if $|d(x^N, \hat{x}^N) - E[d(X, \hat{X})]| \leq \epsilon$

- By LLN, every pair of sequences sampled from the joint source will virtually be distortion typical
- Consequently, $(1-\delta)2^{N(H(X,\hat{X})-\epsilon)} \leq |\mathcal{A}^N_{d,\epsilon}| \leq 2^{N(H(X,\hat{X})+\epsilon)}$ as before
- For two independently drawn sequences $\hat{X}^N$ and X^N, the probability for them to be distortion typical will be just the same as before. In particular, $(1-\delta)2^{-N(I(X;\hat{X})-3\epsilon)} \leq Pr((X^N, \hat{X}^N) \in \mathcal{A}^N_{d,\epsilon}(X, \hat{X}))$

Covering lemma for distortion typical sequences

Covering lemma for distortion typical sequences

$$Pr((X^N(m), \hat{X}^N) \notin \mathcal{A}_{d,\epsilon}^{(N)}(X, \hat{X}) \text{ for all } m)$$

Covering lemma for distortion typical sequences

$$
\begin{aligned}
&Pr((X^N(m), \hat{X}^N) \notin \mathcal{A}_{d,\epsilon}^{(N)}(X, \hat{X}) \text{ for all } m) \\
&= \prod_{m=1}^{M} Pr((X^N(m), \hat{X}^N) \notin \mathcal{A}_{d,\epsilon}^{(N)}(\hat{X}, X))
\end{aligned}
$$

Covering lemma for distortion typical sequences

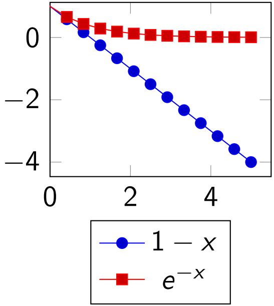
$$
\begin{aligned}
& Pr((X^N(m), \hat{X}^N) \notin \mathcal{A}_{d,\epsilon}^{(N)}(X, \hat{X}) \text{ for all } m) \\
&= \prod_{m=1}^{M} Pr((X^N(m), \hat{X}^N) \notin \mathcal{A}_{d,\epsilon}^{(N)}(\hat{X}, X)) \\
&= \prod_{m=1}^{M} \left[1 - Pr((X^N(m), \hat{X}^N) \in \mathcal{A}_{d,\epsilon}^{(N)}(\hat{X}, X))\right]
\end{aligned}
$$

Covering lemma for distortion typical sequences

$$
\begin{aligned}
&Pr((X^N(m), \hat{X}^N) \notin \mathcal{A}_{d,\epsilon}^{(N)}(X, \hat{X}) \text{ for all } m)\\
=&\prod_{m=1}^{M} Pr((X^N(m), \hat{X}^N) \notin \mathcal{A}_{d,\epsilon}^{(N)}(\hat{X}, X))\\
=&\prod_{m=1}^{M} \left[1 - Pr((X^N(m), \hat{X}^N) \in \mathcal{A}_{d,\epsilon}^{(N)}(\hat{X}, X))\right]\\
\leq&(1-(1-\delta)2^{-N(I(\hat{X};X)+3\epsilon)})^M
\end{aligned}
$$

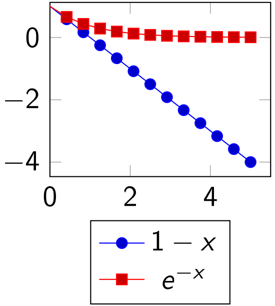
Covering lemma for distortion typical sequences

$$
\begin{aligned}
& Pr((X^N(m), \hat{X}^N) \notin \mathcal{A}_{d,\epsilon}^{(N)}(X, \hat{X}) \text{ for all } m) \\
= & \prod_{m=1}^{M} Pr((X^N(m), \hat{X}^N) \notin \mathcal{A}_{d,\epsilon}^{(N)}(\hat{X}, X)) \\
= & \prod_{m=1}^{M} \left[1 - Pr((X^N(m), \hat{X}^N) \in \mathcal{A}_{d,\epsilon}^{(N)}(\hat{X}, X))\right] \\
\leq & (1 - (1-\delta)2^{-N(I(\hat{X};X)+3\epsilon)})^M \\
\leq & \exp(-M(1-\delta)2^{-N(I(\hat{X};X)+3\epsilon)})
\end{aligned}
$$



Covering lemma for distortion typical sequences

$$
\begin{aligned}
&Pr((X^N(m),\hat{X}^N)\notin \mathcal{A}_{d,\epsilon}^{(N)}(X,\hat{X}) \text{ for all } m)\\
=&\prod_{m=1}^{M} Pr((X^N(m),\hat{X}^N)\notin \mathcal{A}_{d,\epsilon}^{(N)}(\hat{X},X))\\
=&\prod_{m=1}^{M}\left[1-Pr((X^N(m),\hat{X}^N)\in \mathcal{A}_{d,\epsilon}^{(N)}(\hat{X},X))\right]\\
\le&(1-(1-\delta)2^{-N(I(\hat{X};X)+3\epsilon)})^M\\
\le&\exp(-M(1-\delta)2^{-N(I(\hat{X};X)+3\epsilon)})\\
\le&\exp(-(1-\delta)2^{-N(I(\hat{X};X)-R+3\epsilon)})\to 0 \text{ as } N\to\infty \text{ and } R>I(X;\hat{X})
\end{aligned}
$$



Forward proof

Encoding

Given input X^N, find out of the codewords the one that is jointly typical with X^N. And say, if the codeword is $\mathbf{C}(i)$, output index i to the decoder

Forward proof

Encoding

Given input X^N, find out of the codewords the one that is jointly typical with X^N. And say, if the codeword is $\mathbf{C}(i)$, output index i to the decoder

Decoding

Upon receiving the index i, simply output $\mathbf{C}(i)$

Forward proof

Encoding

Given input X^N, find out of the codewords the one that is jointly typical with X^N. And say, if the codeword is $\mathbf{C}(i)$, output index i to the decoder

Decoding

Upon receiving the index i, simply output $\mathbf{C}(i)$

Performance analysis

- First of all, the only point of failure lies on encoding, that is when the encoder cannot find a codeword jointly typical with X^N

Forward proof

Encoding

Given input X^N, find out of the codewords the one that is jointly typical with X^N. And say, if the codeword is $\mathbf{C}(i)$, output index i to the decoder

Decoding

Upon receiving the index i, simply output $\mathbf{C}(i)$

Performance analysis

- First of all, the only point of failure lies on encoding, that is when the encoder cannot find a codeword jointly typical with X^N
- By covering Lemma, encoding failure is neglible as long as $R > I(X; \hat{X})$

Forward proof

Encoding

Given input X^N, find out of the codewords the one that is jointly typical with X^N. And say, if the codeword is $\mathbf{C}(i)$, output index i to the decoder

Decoding

Upon receiving the index i, simply output $\mathbf{C}(i)$

Performance analysis

- First of all, the only point of failure lies on encoding, that is when the encoder cannot find a codeword jointly typical with X^N
- By covering Lemma, encoding failure is neglible as long as $R > I(X;\hat{X})$
- If encoding is successful, $\mathbf{C}(i)$ and X^N should be distortion typical. Therefore, $E[d(\mathbf{C}(i);X^N)] \sim E[d(\hat{X},X)] \leq \mathcal{D}$ as desired

Converse proof

Converse statement

If rate is smaller than $R(\mathcal{D})$, distortion will be larger than $\mathcal{D}$

Converse proof

Converse statement

If rate is smaller than $R(\mathcal{D})$, distortion will be larger than $\mathcal{D}$

Alternative statement

If distortion is less than or equal to $\mathcal{D}$, the rate must be larger than $R(\mathcal{D})$

Converse proof

Converse statement

If rate is smaller than $R(\mathcal{D})$, distortion will be larger than $\mathcal{D}$

Alternative statement

If distortion is less than or equal to $\mathcal{D}$, the rate must be larger than $R(\mathcal{D})$

In the proof, we need to use the convex property of $R(\mathcal{D})$. That is,

$$R(a\mathcal{D}_1 + (1-a)\mathcal{D}_2) \geq aR(\mathcal{D}_1) + (1-a)R(\mathcal{D}_2)$$

So we will digress a little bit to show this convex property first

Log-sum inequality

Log-sum inequality

For any $a_1, \cdots, a_n \geq 0$ and $b_1, \cdots, b_n \geq 0$, we have

$$\sum_i a_i \log_2 \frac{a_i}{b_i} \geq \sum_i a_i \log_2 \frac{\sum_i a_i}{\sum_i b_i}.$$

Log-sum inequality

Log-sum inequality

For any $a_1, \cdots, a_n \geq 0$ and $b_1, \cdots, b_n \geq 0$, we have

$$\sum_i a_i \log_2 \frac{a_i}{b_i} \geq \sum_i a_i \log_2 \frac{\sum_i a_i}{\sum_i b_i}.$$

Proof

We can define two distributions $p(x)$ and $q(x)$ with $p(x_i) = \frac{a_i}{\sum_i a_i}$ and $q(x_i) = \frac{b_i}{\sum_i b_i}$. Since $p(x)$ and $q(x)$ are both non-negative and sum up to 1, they are indeed valid probability mass functions.

Log-sum inequality

Log-sum inequality

For any $a_1, \cdots, a_n \geq 0$ and $b_1, \cdots, b_n \geq 0$, we have

$$\sum_i a_i \log_2 \frac{a_i}{b_i} \geq \sum_i a_i \log_2 \frac{\sum_i a_i}{\sum_i b_i}.$$

Proof

We can define two distributions $p(x)$ and $q(x)$ with $p(x_i) = \frac{a_i}{\sum_i a_i}$ and $q(x_i) = \frac{b_i}{\sum_i b_i}$. Since $p(x)$ and $q(x)$ are both non-negative and sum up to 1, they are indeed valid probability mass functions. Then, we have

Log-sum inequality

Log-sum inequality

For any $a_1, \cdots, a_n \geq 0$ and $b_1, \cdots, b_n \geq 0$, we have

$$\sum_i a_i \log_2 \frac{a_i}{b_i} \geq \sum_i a_i \log_2 \frac{\sum_i a_i}{\sum_i b_i}.$$

Proof

We can define two distributions $p(x)$ and $q(x)$ with $p(x_i) = \frac{a_i}{\sum_i a_i}$ and $q(x_i) = \frac{b_i}{\sum_i b_i}$. Since $p(x)$ and $q(x)$ are both non-negative and sum up to 1, they are indeed valid probability mass functions. Then, we have

$$0 \leq KL(p(x) \| q(x)) = \sum_i p(x_i) \log_2 \frac{p(x_i)}{q(x_i)}$$

Log-sum inequality

Log-sum inequality

For any $a_1, \cdots, a_n \geq 0$ and $b_1, \cdots, b_n \geq 0$, we have

$$\sum_i a_i \log_2 \frac{a_i}{b_i} \geq \sum_i a_i \log_2 \frac{\sum_i a_i}{\sum_i b_i}.$$

Proof

We can define two distributions $p(x)$ and $q(x)$ with $p(x_i) = \frac{a_i}{\sum_i a_i}$ and $q(x_i) = \frac{b_i}{\sum_i b_i}$. Since $p(x)$ and $q(x)$ are both non-negative and sum up to 1, they are indeed valid probability mass functions. Then, we have

$$\begin{aligned} 0 \leq KL(p(x) \| q(x)) &= \sum_i p(x_i) \log_2 \frac{p(x_i)}{q(x_i)} \\ &= \sum_i \frac{a_i}{\sum_i a_i} \left(\log_2 \frac{a_i}{b_i} - \log_2 \frac{\sum_i a_i}{\sum_i b_i} \right) \end{aligned}$$

Convexity of KL-Divergence

For any four distributions $p_1(\cdot)$, $p_2(\cdot)$, $q_1(\cdot)$, and $q_2(\cdot)$, we have

$$\lambda_1 KL(p_1 \| q_1) + \lambda_2 KL(p_2 \| q_2) \geq KL(\lambda_1 p_1 + \lambda_2 p_2 \| \lambda_1 q_1 + \lambda_2 q_2),$$

where $\lambda_1, \lambda_2 \geq 0$ and $\lambda_1 + \lambda_2 = 1$

Convexity of KL-Divergence

For any four distributions $p_1(\cdot)$, $p_2(\cdot)$, $q_1(\cdot)$, and $q_2(\cdot)$, we have

$$\lambda_1 KL(p_1\|q_1) + \lambda_2 KL(p_2\|q_2) \geq KL(\lambda_1 p_1 + \lambda_2 p_2\|\lambda_1 q_1 + \lambda_2 q_2),$$

where $\lambda_1, \lambda_2 \geq 0$ and $\lambda_1 + \lambda_2 = 1$

Proof

$$\begin{aligned}&\lambda_1 KL(p_1\|q_1) + \lambda_2 KL(p_2\|q_2)\\ =&\lambda_1 \sum_{x\in\mathcal{X}} p_1(x)\log\frac{p_1(x)}{q_1(x)} + \lambda_2 \sum_{x\in\mathcal{X}} p_2(x)\log\frac{p_2(x)}{q_2(x)}\end{aligned}$$

Convexity of KL-Divergence

For any four distributions $p_1(\cdot)$, $p_2(\cdot)$, $q_1(\cdot)$, and $q_2(\cdot)$, we have

$$\lambda_1 KL(p_1\|q_1) + \lambda_2 KL(p_2\|q_2) \geq KL(\lambda_1 p_1 + \lambda_2 p_2 \| \lambda_1 q_1 + \lambda_2 q_2),$$

where $\lambda_1, \lambda_2 \geq 0$ and $\lambda_1 + \lambda_2 = 1$

Proof

$$\begin{aligned}
&\lambda_1 KL(p_1\|q_1) + \lambda_2 KL(p_2\|q_2) \\
=&\lambda_1 \sum_{x\in\mathcal{X}} p_1(x)\log\frac{p_1(x)}{q_1(x)} + \lambda_2 \sum_{x\in\mathcal{X}} p_2(x)\log\frac{p_2(x)}{q_2(x)} \\
=&\sum_{x\in\mathcal{X}} \lambda_1 p_1(x)\log\frac{\lambda_1 p_1(x)}{\lambda_1 q_1(x)} + \lambda_2 p_2(x)\log\frac{\lambda_2 p_2(x)}{\lambda_2 q_2(x)}
\end{aligned}$$

Convexity of KL-Divergence

For any four distributions $p_1(\cdot)$, $p_2(\cdot)$, $q_1(\cdot)$, and $q_2(\cdot)$, we have

$$\lambda_1 KL(p_1\|q_1) + \lambda_2 KL(p_2\|q_2) \geq KL(\lambda_1 p_1 + \lambda_2 p_2 \| \lambda_1 q_1 + \lambda_2 q_2),$$

where $\lambda_1, \lambda_2 \geq 0$ and $\lambda_1 + \lambda_2 = 1$

Proof

$$\begin{aligned}
&\lambda_1 KL(p_1\|q_1) + \lambda_2 KL(p_2\|q_2) \\
=&\lambda_1 \sum_{x\in\mathcal{X}} p_1(x)\log\frac{p_1(x)}{q_1(x)} + \lambda_2 \sum_{x\in\mathcal{X}} p_2(x)\log\frac{p_2(x)}{q_2(x)} \\
=&\sum_{x\in\mathcal{X}} \lambda_1 p_1(x)\log\frac{\lambda_1 p_1(x)}{\lambda_1 q_1(x)} + \lambda_2 p_2(x)\log\frac{\lambda_2 p_2(x)}{\lambda_2 q_2(x)} \\
\geq&\sum_{x\in\mathcal{X}} (\lambda_1 p_1(x) + \lambda_2 p_2(x))\log\frac{\lambda_1 p_1(x)+\lambda_2 p_2(x)}{\lambda_1 q_1(x)+\lambda_2 q_2(x)} \quad \text{(by log-sum inequality)}
\end{aligned}$$

Convexity of KL-Divergence

For any four distributions $p_1(\cdot)$, $p_2(\cdot)$, $q_1(\cdot)$, and $q_2(\cdot)$, we have

$$\lambda_1 KL(p_1\|q_1) + \lambda_2 KL(p_2\|q_2) \geq KL(\lambda_1 p_1 + \lambda_2 p_2\|\lambda_1 q_1 + \lambda_2 q_2),$$

where $\lambda_1, \lambda_2 \geq 0$ and $\lambda_1 + \lambda_2 = 1$

Proof

$$\begin{aligned}
&\lambda_1 KL(p_1\|q_1) + \lambda_2 KL(p_2\|q_2)\\
=&\lambda_1 \sum_{x\in\mathcal{X}} p_1(x)\log\frac{p_1(x)}{q_1(x)} + \lambda_2 \sum_{x\in\mathcal{X}} p_2(x)\log\frac{p_2(x)}{q_2(x)}\\
=&\sum_{x\in\mathcal{X}} \lambda_1 p_1(x)\log\frac{\lambda_1 p_1(x)}{\lambda_1 q_1(x)} + \lambda_2 p_2(x)\log\frac{\lambda_2 p_2(x)}{\lambda_2 q_2(x)}\\
\geq&\sum_{x\in\mathcal{X}} (\lambda_1 p_1(x) + \lambda_2 p_2(x))\log\frac{\lambda_1 p_1(x) + \lambda_2 p_2(x)}{\lambda_1 q_1(x) + \lambda_2 q_2(x)} \quad \text{(by log-sum inequality)}\\
=&KL(\lambda_1 p_1 + \lambda_2 p_2\|\lambda_1 q_1 + \lambda_2 q_2)
\end{aligned}$$

Convexity of $I(X;Y)$ with respect to $p(y|x)$

For any random variables X and Y, $I(X;Y)$ is a convex function of $p(y|x)$ for a fixed $p(x)$

Convexity of $I(X;Y)$ with respect to $p(y|x)$

For any random variables X and Y, $I(X;Y)$ is a convex function of $p(y|x)$ for a fixed $p(x)$

Remark

$I(X;Y)$ is concave with respect to $p(x)$ for fixed $p(y|x)$ though. A proof is given in Cover and Thomas and will be omitted here

Convexity of $I(X;Y)$ with respect to $p(y|x)$

For any random variables X and Y, $I(X;Y)$ is a convex function of $p(y|x)$ for a fixed $p(x)$

Remark

$I(X;Y)$ is concave with respect to $p(x)$ for fixed $p(y|x)$ though. A proof is given in Cover and Thomas and will be omitted here

Proof

Let us write

$$\begin{aligned} I(X;Y) &= KL(p(x,y)\|p(x)p(y)) \\ &= KL\left(p(x)p(y|x)\middle\|p(x)\sum_x p(x)p(y|x)\right) \triangleq f(p(y|x)) \end{aligned}$$

Convexity of $I(X;Y)$ with respect to $p(y|x)$

For any random variables X and Y, $I(X;Y)$ is a convex function of $p(y|x)$ for a fixed $p(x)$

Remark

$I(X;Y)$ is concave with respect to $p(x)$ for fixed $p(y|x)$ though. A proof is given in Cover and Thomas and will be omitted here

Proof

Let us write

$$\begin{aligned} I(X;Y) &= KL(p(x,y)\|p(x)p(y)) \\ &= KL\left(p(x)p(y|x)\middle\|p(x)\sum_x p(x)p(y|x)\right) \triangleq f(p(y|x)) \end{aligned}$$

We want to show

$$\lambda f(p_1(y|x)) + (1-\lambda)f(p_2(y|x)) \geq f(\lambda p_1(y|x) + (1-\lambda)p_2(y|x))$$

Proof

Continue from previous slide, we have

$$\begin{aligned}
&\lambda f(p_1(y|x)) + (1-\lambda) f(p_2(y|x)) \\
=&\lambda KL\left(p(x)p_1(y|x)\middle\| p(x)\sum_x p(x)p_1(y|x)\right) \\
&\quad + (1-\lambda) KL\left(p(x)p_2(y|x)\middle\| p(x)\sum_x p(x)p_2(y|x)\right)
\end{aligned}$$

Proof

Continue from previous slide, we have

$$
\begin{aligned}
&\lambda f(p_1(y|x)) + (1-\lambda) f(p_2(y|x)) \\
=&\lambda KL\left(p(x)p_1(y|x) \middle\| p(x)\sum_x p(x)p_1(y|x)\right) \\
&\quad + (1-\lambda) KL\left(p(x)p_2(y|x) \middle\| p(x)\sum_x p(x)p_2(y|x)\right) \\
\geq& KL\left(\lambda p(x)p_1(y|x) + (1-\lambda)p(x)p_2(y|x) \middle\| \lambda p(x)\sum_x p(x)p_1(y|x)\right. \\
&\quad \left. + (1-\lambda)p(x)\sum_x p(x)p_2(y|x)\right)
\end{aligned}
$$

Proof

Continue from previous slide, we have

$$
\begin{aligned}
&\lambda f(p_1(y|x)) + (1-\lambda) f(p_2(y|x)) \\
=&\lambda KL\left(p(x)p_1(y|x)\middle\| p(x)\sum_x p(x)p_1(y|x)\right) \\
&\quad + (1-\lambda) KL\left(p(x)p_2(y|x)\middle\| p(x)\sum_x p(x)p_2(y|x)\right) \\
\geq& KL\left(\lambda p(x)p_1(y|x) + (1-\lambda)p(x)p_2(y|x)\middle\| \lambda p(x)\sum_x p(x)p_1(y|x)\right. \\
&\quad \left. + (1-\lambda)p(x)\sum_x p(x)p_2(y|x)\right) \\
=& KL\left(p(x)[\lambda p_1(y|x) + (1-\lambda)p_2(y|x)]\middle\| p(x)\sum_x p(x)[\lambda p_1(y|x) + (1-\lambda)p_2(y|x)]\right)
\end{aligned}
$$

Proof

Continue from previous slide, we have

$$
\begin{aligned}
&\lambda f(p_1(y|x)) + (1-\lambda) f(p_2(y|x)) \\
=&\lambda KL\left(p(x)p_1(y|x)\middle\| p(x)\sum_x p(x)p_1(y|x)\right) \\
&\quad + (1-\lambda) KL\left(p(x)p_2(y|x)\middle\| p(x)\sum_x p(x)p_2(y|x)\right) \\
\geq& KL\left(\lambda p(x)p_1(y|x) + (1-\lambda)p(x)p_2(y|x)\middle\| \lambda p(x)\sum_x p(x)p_1(y|x)\right. \\
&\quad \left. + (1-\lambda)p(x)\sum_x p(x)p_2(y|x)\right) \\
=& KL\left(p(x)[\lambda p_1(y|x) + (1-\lambda)p_2(y|x)]\middle\| p(x)\sum_x p(x)[\lambda p_1(y|x) + (1-\lambda)p_2(y|x)]\right) \\
=& f(\lambda p_1(y|x) + (1-\lambda)p_2(y|x))
\end{aligned}
$$

Convexity of $R(\mathcal{D})$

Recall that $R(\mathcal{D}) = \min_{p(\hat{x}|x)} I(\hat{X};X)$ with $E[d(X,\hat{X})] \leq \mathcal{D}$
We want to show that

$$R(\lambda\mathcal{D}_1 + (1-\lambda)\mathcal{D}_2) \leq \lambda R(\mathcal{D}_1) + (1-\lambda)R(\mathcal{D}_2)$$

Convexity of $R(\mathcal{D})$

Recall that $R(\mathcal{D}) = \min_{p(\hat{x}|x)} I(\hat{X};X)$ with $E[d(X,\hat{X})] \leq \mathcal{D}$
We want to show that

$$R(\lambda\mathcal{D}_1 + (1-\lambda)\mathcal{D}_2) \leq \lambda R(\mathcal{D}_1) + (1-\lambda)R(\mathcal{D}_2)$$

Proof

Let $p_1^*(\hat{x}|x)$ and $p_2^*(\hat{x}|x)$ be the distributions that optimize $R(\mathcal{D}_1)$ and $R(\mathcal{D}_2)$. Let's try to time share between the two distributions.

Convexity of $R(\mathcal{D})$

Recall that $R(\mathcal{D}) = \min_{p(\hat{x}|x)} I(\hat{X}; X)$ with $E[d(X, \hat{X})] \leq \mathcal{D}$
We want to show that

$$R(\lambda\mathcal{D}_1 + (1-\lambda)\mathcal{D}_2) \leq \lambda R(\mathcal{D}_1) + (1-\lambda)R(\mathcal{D}_2)$$

Proof

Let $p_1^*(\hat{x}|x)$ and $p_2^*(\hat{x}|x)$ be the distributions that optimize $R(\mathcal{D}_1)$ and $R(\mathcal{D}_2)$. Let's try to time share between the two distributions. That is, using $p_1^*(\hat{x}|x)$ with λ fraction of time and $p_2^*(\hat{x}|x)$ with $(1-\lambda)$ fraction of time.

Convexity of $R(\mathcal{D})$

Recall that $R(\mathcal{D}) = \min_{p(\hat{x}|x)} I(\hat{X}; X)$ with $E[d(X, \hat{X})] \leq \mathcal{D}$
We want to show that

$$R(\lambda \mathcal{D}_1 + (1-\lambda)\mathcal{D}_2) \leq \lambda R(\mathcal{D}_1) + (1-\lambda) R(\mathcal{D}_2)$$

Proof

Let $p_1^*(\hat{x}|x)$ and $p_2^*(\hat{x}|x)$ be the distributions that optimize $R(\mathcal{D}_1)$ and $R(\mathcal{D}_2)$. Let's try to time share between the two distributions. That is, using $p_1^*(\hat{x}|x)$ with λ fraction of time and $p_2^*(\hat{x}|x)$ with $(1-\lambda)$ fraction of time. The resulting distortion will be $\lambda \mathcal{D}_1 + (1-\lambda)\mathcal{D}_2$.

Convexity of $R(\mathcal{D})$

Recall that $R(\mathcal{D}) = \min_{p(\hat{x}|x)} I(\hat{X};X)$ with $E[d(X,\hat{X})] \leq \mathcal{D}$
We want to show that

$$R(\lambda\mathcal{D}_1 + (1-\lambda)\mathcal{D}_2) \leq \lambda R(\mathcal{D}_1) + (1-\lambda)R(\mathcal{D}_2)$$

Proof

Let $p_1^*(\hat{x}|x)$ and $p_2^*(\hat{x}|x)$ be the distributions that optimize $R(\mathcal{D}_1)$ and $R(\mathcal{D}_2)$. Let's try to time share between the two distributions. That is, using $p_1^*(\hat{x}|x)$ with λ fraction of time and $p_2^*(\hat{x}|x)$ with $(1-\lambda)$ fraction of time. The resulting distortion will be $\lambda\mathcal{D}_1 + (1-\lambda)\mathcal{D}_2$. Therefore,

$$\lambda R(\mathcal{D}_1) + (1-\lambda)R(\mathcal{D}_2) = \lambda I(\hat{X}_1;X) + (1-\lambda)I(\hat{X}_2;X)$$

Convexity of $R(\mathcal{D})$

Recall that $R(\mathcal{D}) = \min_{p(\hat{x}|x)} I(\hat{X};X)$ with $E[d(X,\hat{X})] \leq \mathcal{D}$
We want to show that

$$R(\lambda\mathcal{D}_1 + (1-\lambda)\mathcal{D}_2) \leq \lambda R(\mathcal{D}_1) + (1-\lambda)R(\mathcal{D}_2)$$

Proof

Let $p_1^*(\hat{x}|x)$ and $p_2^*(\hat{x}|x)$ be the distributions that optimize $R(\mathcal{D}_1)$ and $R(\mathcal{D}_2)$. Let's try to time share between the two distributions. That is, using $p_1^*(\hat{x}|x)$ with λ fraction of time and $p_2^*(\hat{x}|x)$ with $(1-\lambda)$ fraction of time. The resulting distortion will be $\lambda\mathcal{D}_1 + (1-\lambda)\mathcal{D}_2$. Therefore,

$$\lambda R(\mathcal{D}_1) + (1-\lambda)R(\mathcal{D}_2) = \lambda I(\hat{X}_1;X) + (1-\lambda)I(\hat{X}_2;X)$$

$$= \lambda f(p_1^*(\hat{x}|x)) + (1-\lambda)f(p_2^*(\hat{x}|x))$$

Convexity of $R(\mathcal{D})$

Recall that $R(\mathcal{D}) = \min_{p(\hat{x}|x)} I(\hat{X};X)$ with $E[d(X,\hat{X})] \leq \mathcal{D}$
We want to show that

$$R(\lambda\mathcal{D}_1 + (1-\lambda)\mathcal{D}_2) \leq \lambda R(\mathcal{D}_1) + (1-\lambda)R(\mathcal{D}_2)$$

Proof

Let $p_1^*(\hat{x}|x)$ and $p_2^*(\hat{x}|x)$ be the distributions that optimize $R(\mathcal{D}_1)$ and $R(\mathcal{D}_2)$. Let's try to time share between the two distributions. That is, using $p_1^*(\hat{x}|x)$ with λ fraction of time and $p_2^*(\hat{x}|x)$ with $(1-\lambda)$ fraction of time. The resulting distortion will be $\lambda\mathcal{D}_1 + (1-\lambda)\mathcal{D}_2$. Therefore,

$$\lambda R(\mathcal{D}_1) + (1-\lambda)R(\mathcal{D}_2) = \lambda I(\hat{X}_1;X) + (1-\lambda)I(\hat{X}_2;X)$$

$$=\lambda f(p_1^*(\hat{x}|x)) + (1-\lambda)f(p_2^*(\hat{x}|x)) \geq f(\lambda p_1^*(\hat{x}|x) + (1-\lambda)p_2^*(\hat{x}|x))$$

Convexity of $R(\mathcal{D})$

Recall that $R(\mathcal{D}) = \min_{p(\hat{x}|x)} I(\hat{X};X)$ with $E[d(X,\hat{X})] \le \mathcal{D}$
We want to show that

$$R(\lambda\mathcal{D}_1 + (1-\lambda)\mathcal{D}_2) \le \lambda R(\mathcal{D}_1) + (1-\lambda)R(\mathcal{D}_2)$$

Proof

Let $p_1^*(\hat{x}|x)$ and $p_2^*(\hat{x}|x)$ be the distributions that optimize $R(\mathcal{D}_1)$ and $R(\mathcal{D}_2)$. Let's try to time share between the two distributions. That is, using $p_1^*(\hat{x}|x)$ with λ fraction of time and $p_2^*(\hat{x}|x)$ with $(1-\lambda)$ fraction of time. The resulting distortion will be $\lambda\mathcal{D}_1 + (1-\lambda)\mathcal{D}_2$. Therefore,

$$\begin{aligned} &\lambda R(\mathcal{D}_1) + (1-\lambda)R(\mathcal{D}_2) = \lambda I(\hat{X}_1;X) + (1-\lambda)I(\hat{X}_2;X) \\ =&\lambda f(p_1^*(\hat{x}|x)) + (1-\lambda)f(p_2^*(\hat{x}|x)) \ge f(\lambda p_1^*(\hat{x}|x) + (1-\lambda)p_2^*(\hat{x}|x)) \\ =&I(\tilde{X};X) \end{aligned}$$

Convexity of $R(\mathcal{D})$

Recall that $R(\mathcal{D}) = \min_{p(\hat{x}|x)} I(\hat{X};X)$ with $E[d(X,\hat{X})] \leq \mathcal{D}$
We want to show that

$$R(\lambda\mathcal{D}_1 + (1-\lambda)\mathcal{D}_2) \leq \lambda R(\mathcal{D}_1) + (1-\lambda)R(\mathcal{D}_2)$$

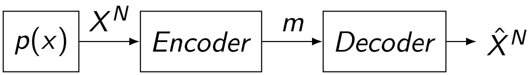
Proof

Let $p_1^*(\hat{x}|x)$ and $p_2^*(\hat{x}|x)$ be the distributions that optimize $R(\mathcal{D}_1)$ and $R(\mathcal{D}_2)$. Let's try to time share between the two distributions. That is, using $p_1^*(\hat{x}|x)$ with λ fraction of time and $p_2^*(\hat{x}|x)$ with $(1-\lambda)$ fraction of time. The resulting distortion will be $\lambda\mathcal{D}_1 + (1-\lambda)\mathcal{D}_2$. Therefore,

$$\begin{aligned}
&\lambda R(\mathcal{D}_1) + (1-\lambda)R(\mathcal{D}_2) = \lambda I(\hat{X}_1;X) + (1-\lambda)I(\hat{X}_2;X) \\
=&\lambda f(p_1^*(\hat{x}|x)) + (1-\lambda)f(p_2^*(\hat{x}|x)) \geq f(\lambda p_1^*(\hat{x}|x) + (1-\lambda)p_2^*(\hat{x}|x)) \\
=&I(\tilde{X};X) \geq R(\lambda\mathcal{D}_1 + (1-\lambda)\mathcal{D}_2),
\end{aligned}$$

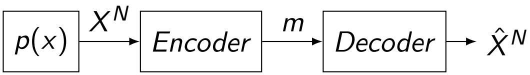
where $\tilde{X} = \begin{cases} \hat{X}_1 & \text{with } \lambda \text{ fraction of time} \\ \hat{X}_2 & \text{with } (1-\lambda) \text{ fraction of time} \end{cases}$

Converse proof



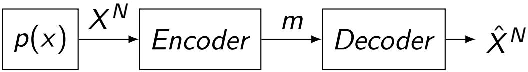
$NR \geq H(M)$

Converse proof



$$NR \geq H(M) \geq H(M) - H(M|X^N) = I(M; X^N)$$

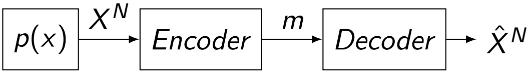
Converse proof



$$NR \geq H(M) \geq H(M) - H(M|X^N) = I(M; X^N) \geq I(\hat{X}^N; X^N)$$

$$= H(X^N) - H(X^N|\hat{X}^N)$$

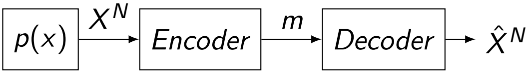
Converse proof



$$NR \geq H(M) \geq H(M) - H(M|X^N) = I(M;X^N) \geq I(\hat{X}^N;X^N)$$

$$= H(X^N) - H(X^N|\hat{X}^N) = \sum_{i=1}^{N} H(X_i) - \sum_{i=1}^{N} H(X_i|\hat{X}^N, X^{i-1})$$

Converse proof

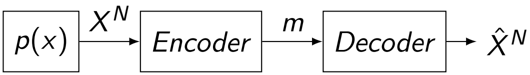


$$NR \geq H(M) \geq H(M) - H(M|X^N) = I(M; X^N) \geq I(\hat{X}^N; X^N)$$

$$= H(X^N) - H(X^N|\hat{X}^N) = \sum_{i=1}^{N} H(X_i) - \sum_{i=1}^{N} H(X_i|\hat{X}^N, X^{i-1})$$

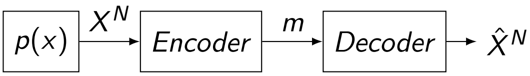
$$\geq \sum_{i=1}^{N} H(X_i) - \sum_{i=1}^{N} H(X_i|\hat{X}_i) = \sum_{i=1}^{N} I(X_i; \hat{X}_i)$$

Converse proof



$$
\begin{aligned}
NR &\geq H(M) \geq H(M) - H(M|X^N) = I(M;X^N) \geq I(\hat{X}^N;X^N) \\
&= H(X^N) - H(X^N|\hat{X}^N) = \sum_{i=1}^N H(X_i) - \sum_{i=1}^N H(X_i|\hat{X}^N, X^{i-1}) \\
&\geq \sum_{i=1}^N H(X_i) - \sum_{i=1}^N H(X_i|\hat{X}_i) = \sum_{i=1}^N I(X_i;\hat{X}_i) \\
&\geq \sum_{i=1}^N R(E[d(X_i,\hat{X}_i)]) = N\left(\frac{1}{N}\sum_{i=1}^N R(E[d(X_i;\hat{X}_i)])\right)
\end{aligned}
$$

Converse proof



$$NR \geq H(M) \geq H(M) - H(M|X^N) = I(M;X^N) \geq I(\hat{X}^N;X^N)$$

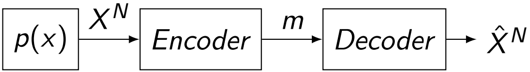
$$= H(X^N) - H(X^N|\hat{X}^N) = \sum_{i=1}^N H(X_i) - \sum_{i=1}^N H(X_i|\hat{X}^N, X^{i-1})$$

$$\geq \sum_{i=1}^N H(X_i) - \sum_{i=1}^N H(X_i|\hat{X}_i) = \sum_{i=1}^N I(X_i;\hat{X}_i)$$

$$\geq \sum_{i=1}^N R(E[d(X_i,\hat{X}_i)]) = N\left(\frac{1}{N}\sum_{i=1}^N R(E[d(X_i;\hat{X}_i)])\right)$$

$$\geq NR\left(\frac{1}{N}\sum_{i=1}^N E[d(X_i;\hat{X}_i)]\right)$$

Converse proof



$$NR \geq H(M) \geq H(M) - H(M|X^N) = I(M; X^N) \geq I(\hat{X}^N; X^N)$$

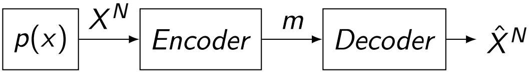
$$= H(X^N) - H(X^N|\hat{X}^N) = \sum_{i=1}^{N} H(X_i) - \sum_{i=1}^{N} H(X_i|\hat{X}^N, X^{i-1})$$

$$\geq \sum_{i=1}^{N} H(X_i) - \sum_{i=1}^{N} H(X_i|\hat{X}_i) = \sum_{i=1}^{N} I(X_i; \hat{X}_i)$$

$$\geq \sum_{i=1}^{N} R(E[d(X_i, \hat{X}_i)]) = N\left(\frac{1}{N}\sum_{i=1}^{N} R(E[d(X_i; \hat{X}_i)])\right)$$

$$\geq NR\left(\frac{1}{N}\sum_{i=1}^{N} E[d(X_i; \hat{X}_i)]\right) = NR\left(E\left[\frac{1}{N}\sum_{i=1}^{N} d(X_i; \hat{X}_i)\right]\right)$$

Converse proof



$$
\begin{aligned}
NR &\geq H(M) \geq H(M) - H(M|X^N) = I(M;X^N) \geq I(\hat{X}^N;X^N) \\
&= H(X^N) - H(X^N|\hat{X}^N) = \sum_{i=1}^{N} H(X_i) - \sum_{i=1}^{N} H(X_i|\hat{X}^N, X^{i-1}) \\
&\geq \sum_{i=1}^{N} H(X_i) - \sum_{i=1}^{N} H(X_i|\hat{X}_i) = \sum_{i=1}^{N} I(X_i;\hat{X}_i) \\
&\geq \sum_{i=1}^{N} R(E[d(X_i,\hat{X}_i)]) = N\left(\frac{1}{N}\sum_{i=1}^{N} R(E[d(X_i;\hat{X}_i)])\right) \\
&\geq NR\left(\frac{1}{N}\sum_{i=1}^{N} E[d(X_i;\hat{X}_i)]\right) = NR\left(E\left[\frac{1}{N}\sum_{i=1}^{N} d(X_i;\hat{X}_i)\right]\right) \\
&= NR(E[d(X^N;\hat{X}^N)]) \geq NR(D)
\end{aligned}
$$