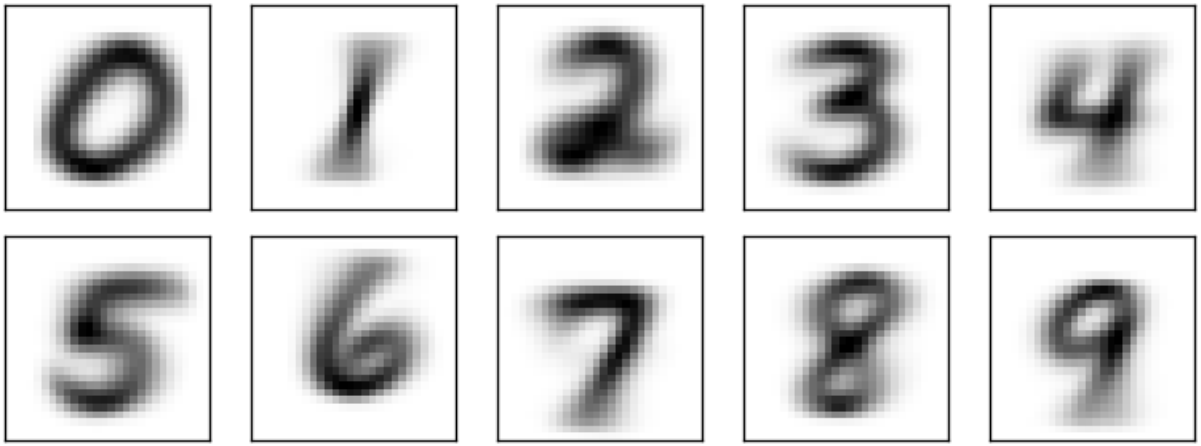
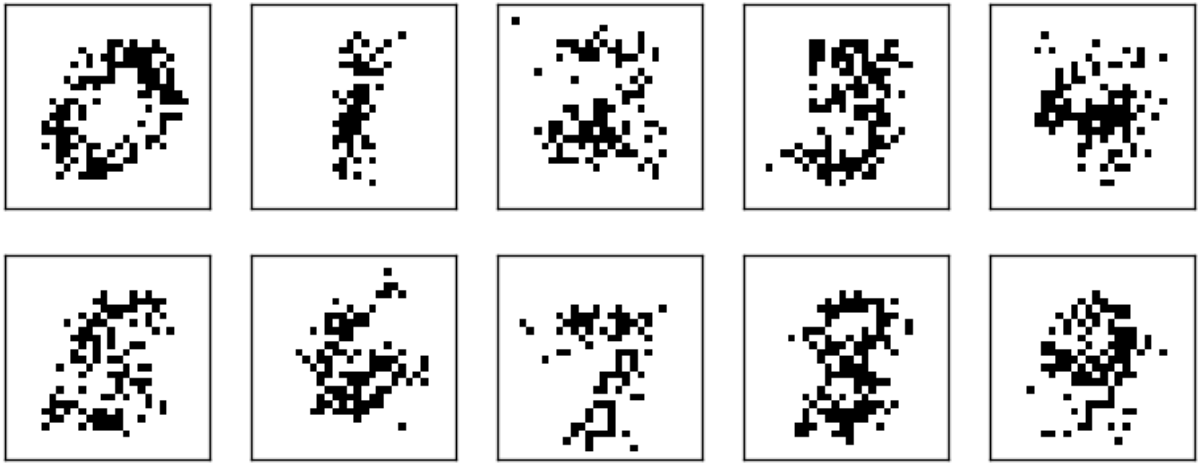
Entropy, Mutual Information

Naïve Bayes, MNIST

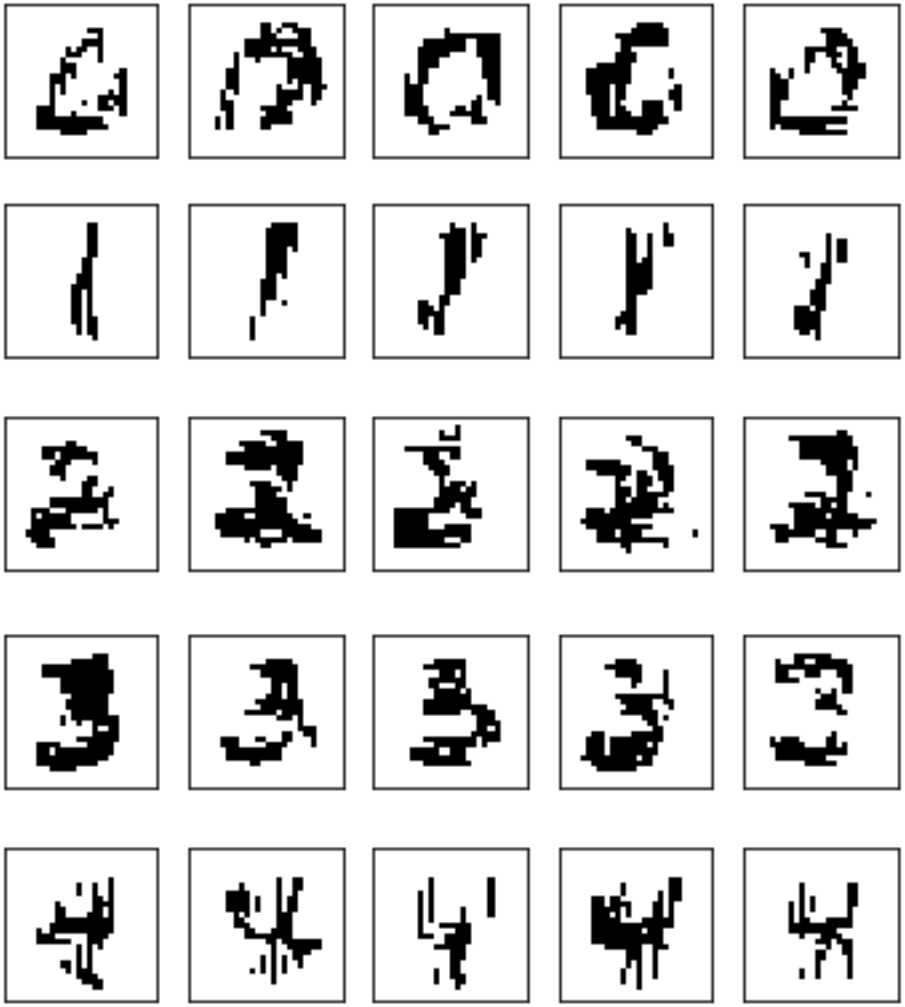
- average for each class

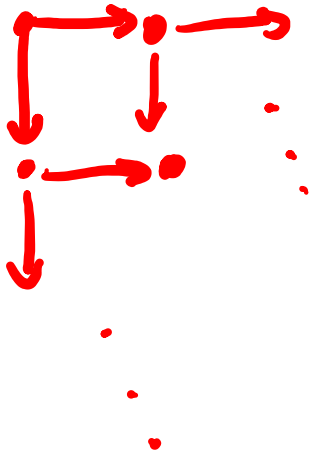


- sample images from our generative model



Bayes Net, Conditional Independence Assumption





- entropy
- mutual information
- convex functions
- Fano's inequality

Example 1

- entropy: how much information do we learn from realization of X *on average*?

$$H(X) = E\left[\log_2\left(\frac{1}{p(X)}\right)\right]$$

- flip a weighted coin with bias $p = 0.99$

$$H(X) = 0.99 \times \log\left(\frac{1}{0.99}\right) + 0.01 \times \log\left(\frac{1}{0.01}\right)$$

$$\approx 0.081 \text{ bits}$$

- flip a weighted coin with bias $p = 0.6$

$$H(X) = 0.6 \times \log\left(\frac{1}{0.6}\right) + 0.6 \times \log\left(\frac{1}{0.6}\right)$$

$$\approx 0.97 \text{ bits}$$

Binary Entropy

- discrete $X \in \mathcal{X}$, with $\mathcal{X}$ finite set

- binary outcomes – $|\mathcal{X}| = 2$

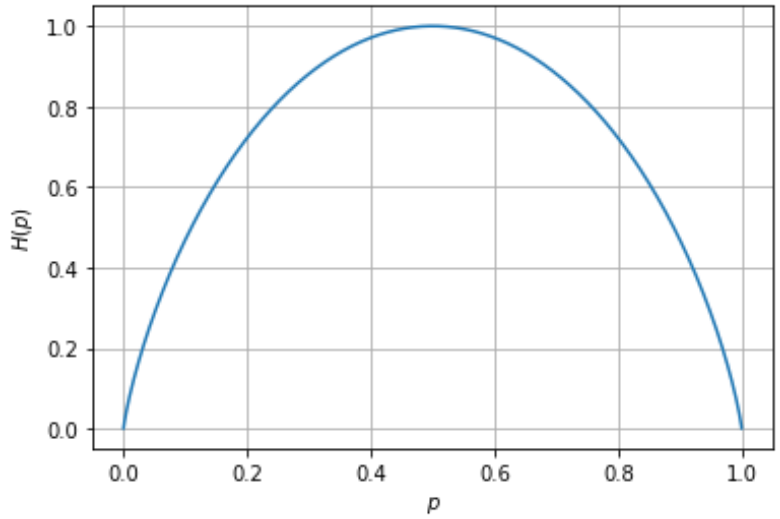
$$X = \begin{cases} 0 & \text{with probability } p \\ 1 & \text{with probability } (1-p) \end{cases}$$

$$X = \begin{cases} 7 & \text{with probability } p \\ 32 & \text{with probability } (1-p) \end{cases}$$

$$X = \begin{cases} \text{hippopotamus} & \text{with probability } p \\ \text{zebra} & \text{with probability } (1-p) \end{cases}$$

binary entropy function

$$H(X) = p \log_2 \left(\frac{1}{p}\right) + (1-p) \log_2 \left(\frac{1}{1-p}\right) =: h(p)$$



Entropy is a parameter of a distribution --- like the mean and variance: $E[X]$

$$\text{var}(X) = E[(X - E[X])^2]$$

$$H(X) = E\left[\log_2 \left(\frac{1}{p(X)}\right)\right]$$

Example 2

- game: find the value of a random variable X with yes/no questions

$$X = \begin{cases} \text{cat} & \text{with probability } 1/2 \\ \text{dog} & \text{with probability } 1/4 \\ \text{fish} & \text{with probability } 1/8 \\ \text{zebra} & \text{with probability } 1/8 \end{cases}$$

1. is it a cat?
2. is it a dog?
3. is it a zebra?

- strategy: ask informative questions
- what's the strategy with the fewest questions, on average?

$$H(X) = \frac{1}{2}\log_2 \frac{1}{\mathbb{P}(\text{cat})} + \frac{1}{4}\log_2 \frac{1}{\mathbb{P}(\text{dog})} + \frac{1}{8}\log_2 \frac{1}{\mathbb{P}(\text{fish})} + \frac{1}{8}\log_2 \frac{1}{\mathbb{P}(\text{zebra})}$$

$$H(X) = \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{1}{8}\log_2 8 + \frac{1}{8}\log_2 8 \quad = 1.75 \text{ bits}$$

- in general, best strategy has expected number of questions is between $H(X)$ and $H(X)+1$

Entropy

interpretations:

$$H(X) = E\left[\log_2\left(\frac{1}{p(X)}\right)\right]$$

- just another paramater/statistic of a random variable
- average amount of information we learn from a realization of a random variable
- amount of uncertainty in a random variable X (in bits)
- fundamental to encoding and compression (more later)

properties:

$$H(X) \geq 0$$

$$H(X) \leq \log_2(|\mathcal{X}|)$$

[Wikipedia, Geometric Distribution]

Parameters	$0 < p \leq 1$ success probability (real)	$0 < p \leq 1$ success probability (real)
Support	k trials where $k \in \{1, 2, 3, \ldots\}$	k failures where $k \in \{0, 1, 2, 3, \ldots\}$
PMF	$(1-p)^{k-1}p$	$(1-p)^k p$
CDF	$1-(1-p)^k$	$1-(1-p)^{k+1}$
Mean	$\frac{1}{p}$	$\frac{1-p}{p}$
Median	$\left\lceil \frac{-1}{\log_2(1-p)} \right\rceil$ (not unique if $-1/\log_2(1-p)$ is an integer)	$\left\lceil \frac{-1}{\log_2(1-p)} \right\rceil - 1$ (not unique if $-1/\log_2(1-p)$ is an integer)
Mode	1	0
Variance	$\frac{1-p}{p^2}$	$\frac{1-p}{p^2}$
Skewness	$\frac{2-p}{\sqrt{1-p}}$	$\frac{2-p}{\sqrt{1-p}}$
Ex. kurtosis	$6+\frac{p^2}{1-p}$	$6+\frac{p^2}{1-p}$
Entropy	$\frac{-(1-p)\log_2(1-p)-p\log_2 p}{p}$	$\frac{-(1-p)\log_2(1-p)-p\log_2 p}{p}$

Joint and Conditional Entropy


- definition - joint entropy:

$$H(X,Y) = \sum_x \sum_y p(x,y) \log_2 \left(\frac{1}{p(x,y)} \right)$$

$p(x,y)$		x		
		0.5	1	2
y	-1	$\frac{1}{8}$	$\frac{1}{8}$	0
	0	0	$\frac{1}{2}$	0
	1	0	$\frac{1}{8}$	$\frac{1}{8}$

- definition - conditional entropy:

$$H(X|Y) = \sum_y p(y) H(X|Y=y)$$

Joint and Conditional Entropy


$p(x,y)$		x		
		0.5	1	2
y	-1	$\frac{1}{8}$	$\frac{1}{8}$	0
	0	0	$\frac{1}{2}$	0
	1	0	$\frac{1}{8}$	$\frac{1}{8}$

- relating conditional entropy to joint (chain rule):

$$H(X,Y) = H(X) + H(Y|X)$$

- properties:

$$H(X) \geq H(X|Y)$$

$$H(X) = H(X|Y) \text{ if } X \perp Y$$

$$H(X,Y) = H(X) + H(Y) \text{ if } X \perp Y$$

$$H(X,Y) = H(X) \text{ if } Y = g(X)$$

Mutual Information

- mutual information:

$$I(Y;X) = H(Y) - H(Y|X)$$

- interpretation: the mutual information is the reduction in uncertainty of Y due to knowledge of X (in bits)

$p(x,y)$		x		
		0.5	1	2
y	-1	$\frac{1}{8}$	$\frac{1}{8}$	0
	0	0	$\frac{1}{2}$	0
	1	0	$\frac{1}{8}$	$\frac{1}{8}$

- properties:

$$I(Y;X) = H(Y) - H(Y|X)$$

$$I(Y;X) = H(X) - H(X|Y)$$

$$I(Y;X) = I(X;Y)$$

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

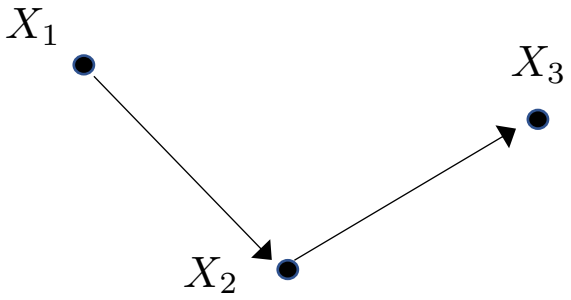
$$I(Y;X) = \sum_{x,y} p(x,y) \log \left(\frac{p(x,y)}{p(y)p(x)} \right)$$

- design principle of information gain algorithms: maximize mutual information

Example

1) flip one coin X_1.
2) if $X_1 = 1$, flip coin X_2, else set $X_2 = 0$.
3) set $X_3 = X_2$

X_1	X_2	X_3	$p(x_1, x_2, x_3)$
0	0	0	$\frac{1}{2}$
1	0	0	$\frac{1}{4}$
1	1	1	$\frac{1}{4}$



$$H(X_1, X_2, X_3) = \sum_{x_1,x_2,x_3} p(x_1, x_2, x_3) \log\left(\frac{1}{p(x_1,x_2,x_3)}\right)$$

$$= H(X_1, X_2)$$

$$= H(X_1) + H(X_2|X_1)$$

Chain Rule for Entropy

- chain rule for joint distributions:

$$p(\boldsymbol{x}) = p(x_1)p(x_2|x_1)p(x_3|x_1,x_2)p(x_4|x_1,x_2,x_3)\dots p(x_n|x_1,\dots,x_{n-1})$$

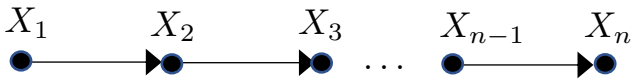
- chain rule for joint entropy:

$$H(X_1,X_2\dots,X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1},\dots,X_1)$$

- joint entropy, two random variables:

$$H(X_1,X_2) = H(X_1) + H(X_2|X_1)$$

Example: Markov Chain



- joint distribution:

$$p(\boldsymbol{x}) = \prod_j p(x_j|\text{parents of } x_j)$$
$$p(\boldsymbol{x}) = p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_3)\ldots p(x_n|x_{n-1})$$

- joint entropy:

$$H(X_1, X_2 \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1)$$
$$= \sum_{i=1}^{n} H(X_i|X_{i-1})$$

Proof of Upper Bound

properties:

$$H(X) \geq 0$$

$$H(X) \leq \log_2(|\mathcal{X}|)$$

proof:

$$H(X) = E\left[\log_2\left(\frac{1}{p(X)}\right)\right] \quad \leq \log_2\left(E\left[\frac{1}{p(X)}\right]\right)$$

1) Jensen's in-equality: for a convex function $f(\cdot)$ and a random variable X

$$E[f(X)] \geq f(E[X])$$

2) definition of expectation

$$E\left[\frac{1}{p(X)}\right] = |\mathcal{X}|$$

Convex Functions

- definition: a function $f(x)$ is convex if

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2) \qquad \text{for all } \lambda \in [0,1]$$

- interpretation: the line connecting two points on the surface of the function is *always* above the function

- equivalently, a function is convex if the second derivative is non-negative

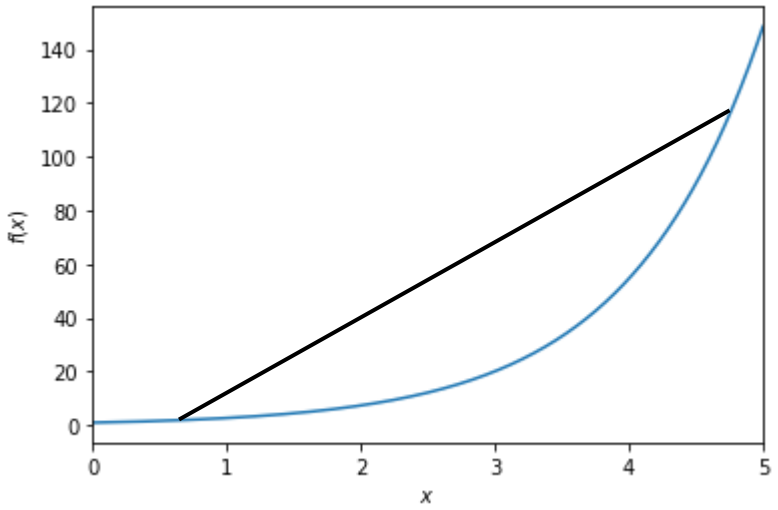
- examples:

$$f(x) = x^2$$

$$f(x) = e^x$$

$$f(x) = |x|$$

$$f(x) = \sin(x) \quad \text{for } \frac{\pi}{2} \leq x \leq \frac{3\pi}{2}$$



Convex Functions in Multiple Dimensions)

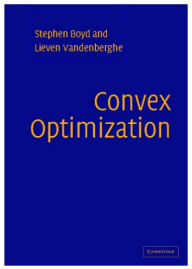
- definition: a function $f(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^n$ is convex if

$$f(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) \leq \lambda f(\boldsymbol{x}_1) + (1-\lambda) f(\boldsymbol{x}_2) \quad \text{for all } \lambda \in [0,1]$$

- interpretation: the line connecting two points on the surface of the function is *always* above the function
- equivalently, a function is convex if the Hessian, $\nabla^2 f(\boldsymbol{x})$ is positive semi-definite
- examples:

$$f(\boldsymbol{x}) = ||\boldsymbol{x}||_p \qquad p \geq 1$$

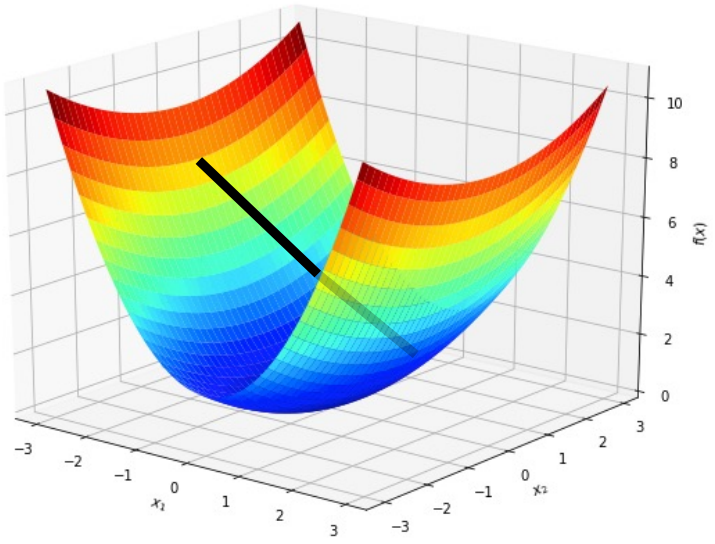
$$f(\boldsymbol{x}) = \max\{x_1, \ldots, x_n\}$$



- important because computers can solve basically any optimization problem

$$\min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$$

if $f(\boldsymbol{x})$ is a convex function over a convex set $\mathcal{X}$



Jensen's Inequality

Jensen's in-equality: for a convex function $f(\cdot)$ and a random variable X

$$E[f(X)] \geq f(E[X])$$

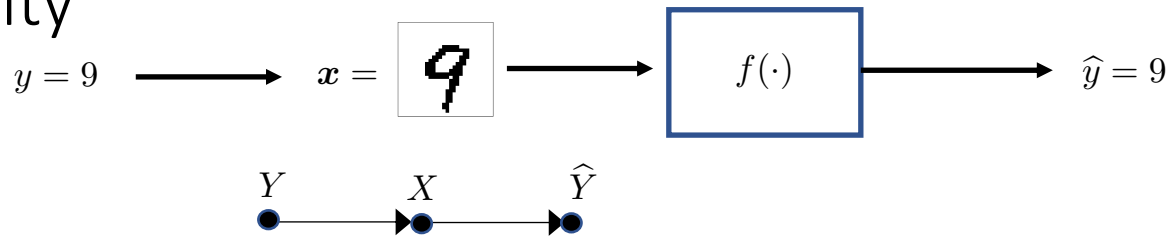
1) consider binary X

$$p(x_1)f(x_1) + p(x_2)f(x_2) \geq f(x_1 p(x_1) + x_2 p(x_2))$$

2) by induction

$$\begin{aligned} E[f(X)] &= p(x_1)f(x_1) + (1 - p(x_1)) \sum_{i \geq 2} \frac{p_i}{1 - p(x_1)} f(x_i) \\ &\geq p(x_1)f(x_1) + (1 - p(x_1)) \sum_{i \geq 2} f\left(\frac{p_i}{1 - p(x_1)} x_i\right) \\ &\geq f\left(p(x_1)x_1 + (1 - p(x_1)) \sum_{i \geq 2} \frac{p_i}{1 - p(x_1)} x_i\right) \quad = f\left(E[X]\right) \end{aligned}$$

Fano's Inequality



- for any $\widehat{y} = f(\boldsymbol{x})$ we have $y \to \boldsymbol{x} \to \widehat{y}$ and

$$\mathbb{P}(y \neq \widehat{y}) \geq \frac{H(y|\boldsymbol{x}) - 1}{\log(|\mathcal{Y}|)} \quad \text{where } |\mathcal{Y}| \text{ is the number of classes}$$

- proof: define $E = \mathbb{I}_{\{\widehat{Y} \neq Y\}}$

$$\begin{aligned} H(E, Y|\widehat{Y}) &= H(Y|\widehat{Y}) + H(E|Y, \widehat{Y}) \\ &= H(Y|\widehat{Y}) \\ &\geq H(Y|X) \end{aligned}$$

$$\begin{aligned} H(E, Y|\widehat{Y}) &= H(E|\widehat{Y}) + H(Y|E, \widehat{Y}) \\ &\leq 1 + H(Y|E, \widehat{Y}) \\ &\leq 1 + \mathbb{P}(E = 1) H(Y|E = 1, \widehat{Y}) + \mathbb{P}(E = 0) H(Y|E = 0, \widehat{Y}) \\ &\leq 1 + \mathbb{P}(\widehat{Y} \neq Y) H(Y|E = 1, \widehat{Y}) \\ &\leq 1 + \mathbb{P}(\widehat{Y} \neq Y) \log |\mathcal{Y}| \end{aligned}$$