# 10. Information theory

# 10.1. Shannon information and entropy

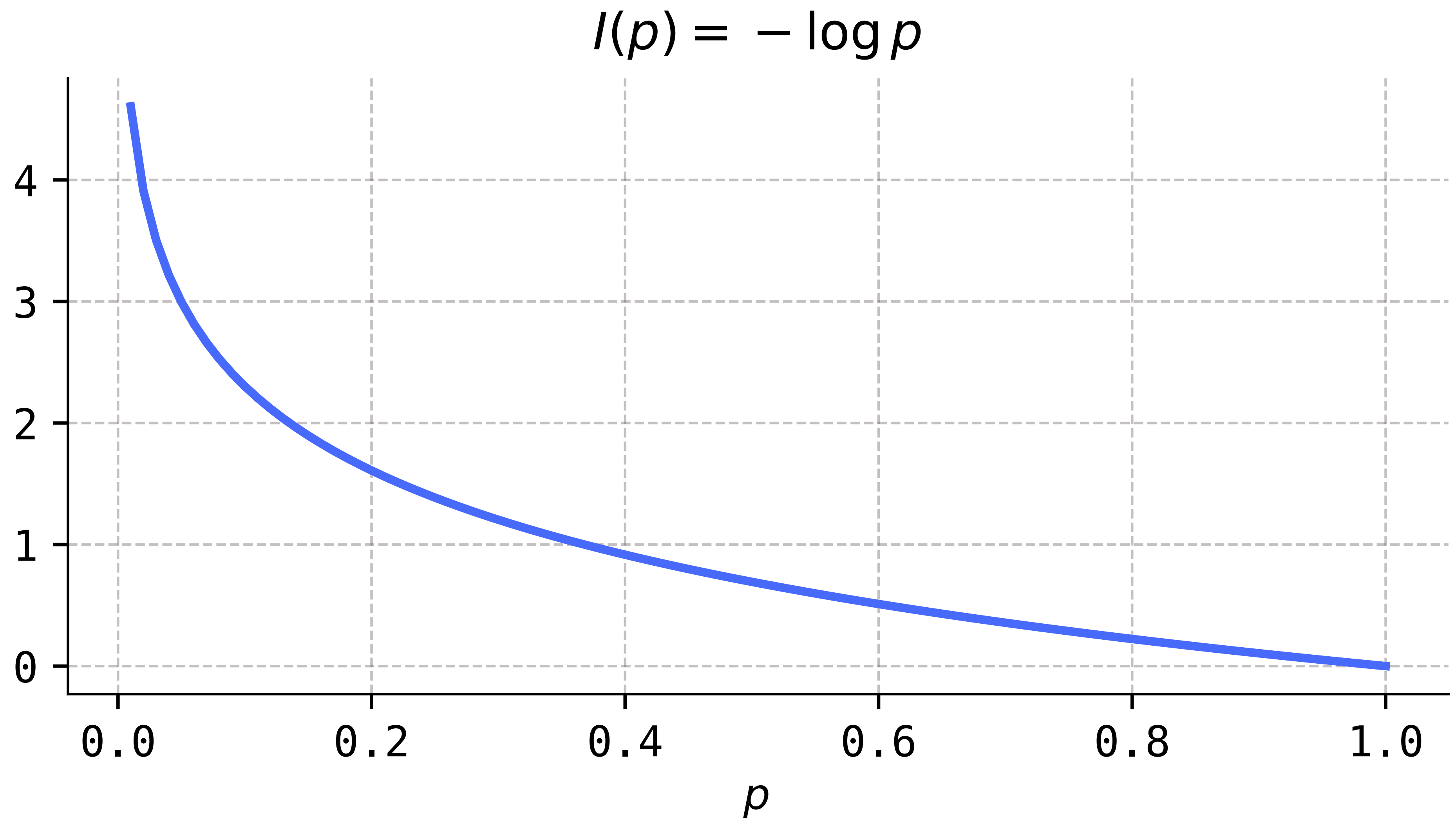
### Definition 10.1

Let $P$ be a probability measure on a finite sample space $S$ with mass function $p(s)$. The *(Shannon) information content* of the sample point $s \in S$, denoted $I_P(s)$, is defined to be

$$I_P(s) \stackrel{\text{def}}{=} -\log(p(s)).$$

The information content is also called the *surprisal*.

If the probability measure $P$ is clear from context, we will write $I(s)$ in place of $I_P(s)$. If $\mathbf{X}$ is a random vector with finite range and probability measure $P_{\mathbf{X}}$, we will write $I_{\mathbf{X}}(\mathbf{x})$ in place of $I_{P_{\mathbf{X}}}(\mathbf{x})$.


$I(p) = -\log p$
4
3
2
1
0
0.0
0.2
0.4
0.6
0.8
1.0
$p$

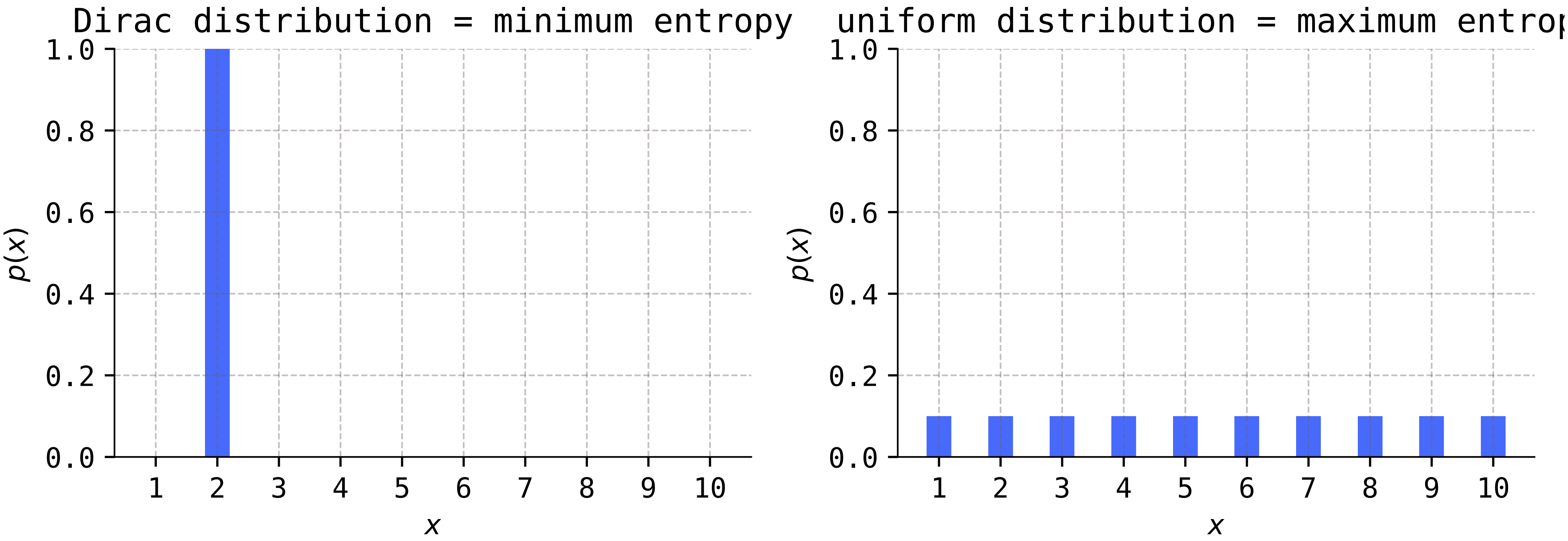
**Definition 10.2**

Let $P$ be a probability measure on a finite sample space $S$ with mass function $p(s)$. The *(Shannon) entropy* of $P$, denoted $H(P)$, is defined to be

$$H(P) \stackrel{\text{def}}{=} \sum_{s \in S} p(s) I_P(s).$$

The entropy is also called the *uncertainty*.

If $\mathbf{X}$ is a random vector with finite range and probability measure $P_{\mathbf{X}}$, we will write $H(\mathbf{X})$ in place of $H(P_{\mathbf{X}})$. If we write the vector in terms of its component random variables $\mathbf{X} = (X_1, \ldots, X_m)$, then we shall also write $H(X_1, \ldots, X_m)$ in place of $H(P_{\mathbf{X}})$ and call this the *joint entropy* of the random variables $X_1, \ldots, X_m$.


Dirac distribution = minimum entropy
uniform distribution = maximum entrop
1.0
0.8
0.6
0.4
0.2
0.0
$p(x)$
1 2 3 4 5 6 7 8 9 10
$x$

**Problem Prompt**

Do problems 1 and 2 on the worksheet.

**Definition 10.3**

Let $P$ and $Q$ be two probability measures on a finite sample space $S$ with mass functions $p(s)$ and $q(s)$. Suppose they satisfy the following condition:

- *Absolute continuity*. For all $s \in S$, if $q(s) = 0$, then $p(s) = 0$. Or equivalently, the support of $q(s)$ contains the support of $p(s)$.

Then the *cross entropy* from $P$ to $Q$, denoted $H_P(Q)$, is defined by

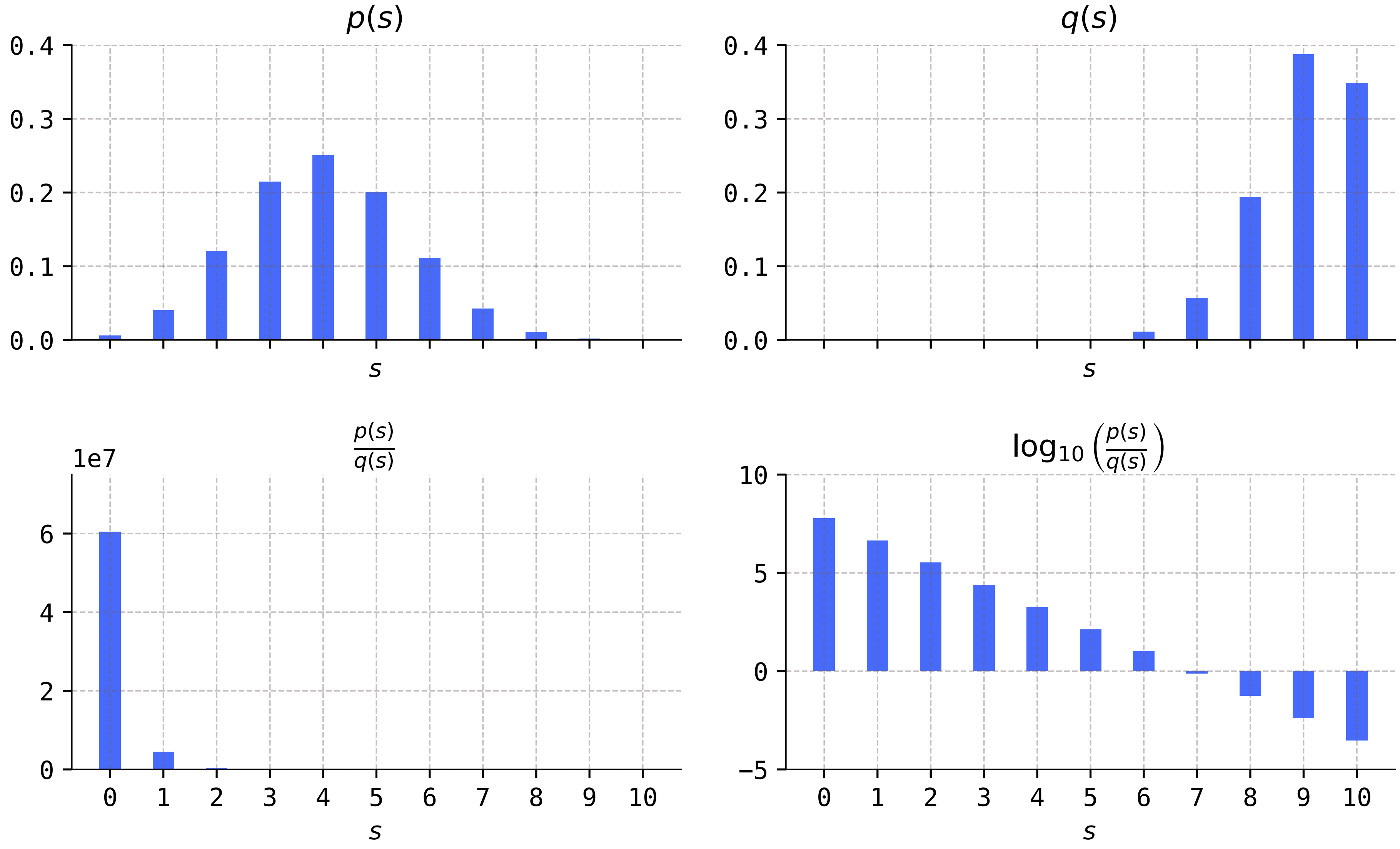
$$H_P(Q) \stackrel{\text{def}}{=} E_{s\sim p(s)}\left[I_Q(s)\right] = -\sum_{s\in S} p(s)\log(q(s)).$$

As usual, if $P_{\mathbf{X}}$ and $P_{\mathbf{Y}}$ are the probability measures of two random vectors $\mathbf{X}$ and $\mathbf{Y}$ with finite ranges, we will write $H_{\mathbf{Y}}(\mathbf{X})$ in place of $H_{P_{\mathbf{Y}}}(P_{\mathbf{X}})$.

**Problem Prompt**

Do problem 3 on the worksheet.

# 10.2. Kullback Leibler divergence



$p(s)$
$q(s)$
$\frac{p(s)}{q(s)}$
1e7
$\log_{10}\left(\frac{p(s)}{q(s)}\right)$
0.0
0.1
0.2
0.3
0.4
0
2
4
6
−5
5
10
0 1 2 3 4 5 6 7 8 9 10
$s$

**Definition 10.4**

Let $P$ and $Q$ be two probability measures on a finite sample space $S$ with mass functions $p(s)$ and $q(s)$. Suppose they satisfy the following condition:

- *Absolute continuity*. For all $s \in S$, if $q(s) = 0$, then $p(s) = 0$. Or equivalently, the support of $q(s)$ contains the support of $p(s)$.

Then the *Kullback-Leibler divergence* (or just *KL divergence*) from $P$ to $Q$, denoted $D(P \parallel Q)$, is the mean order of relative magnitude of $P$ to $Q$. Precisely, it is given by

$$D(P \parallel Q) \stackrel{\text{def}}{=} E_{s\sim p(s)}\left[\log\left(\frac{p(s)}{q(s)}\right)\right] = \sum_{s\in S} p(s)\log\left(\frac{p(s)}{q(s)}\right).$$

The KL divergence is also called the *relative entropy*.

As always, if $P_{\mathbf{X}}$ and $P_{\mathbf{Y}}$ are the probability measures of two random vectors $\mathbf{X}$ and $\mathbf{Y}$ with finite ranges, we will write $D(\mathbf{Y} \parallel \mathbf{X})$ in place of $D(P_{\mathbf{Y}} \parallel P_{\mathbf{X}})$.

**Problem Prompt**

To problem 4 on the worksheet.

**Theorem 10.1 (KL divergence and entropy)**

Let $P$ and $Q$ be two probability measures on a finite sample space $S$. Then

$$D(P \parallel Q) = H_P(Q) - H(P).$$

**Theorem 10.3 (Gibbs' inequality)**

Let $P$ and $Q$ be two probability measures on a finite probability space $S$ satisfying the absolute continuity condition in Definition 10.4. Then

$$D(P \parallel Q) \geq 0,$$

with equality if and only if $P = Q$.

**Corollary 10.1 (Uniform distributions maximize entropy)**

Let $P$ be a probability measures on a finite sample space $S$. Then

$$H(P) \leq \log |S|,$$

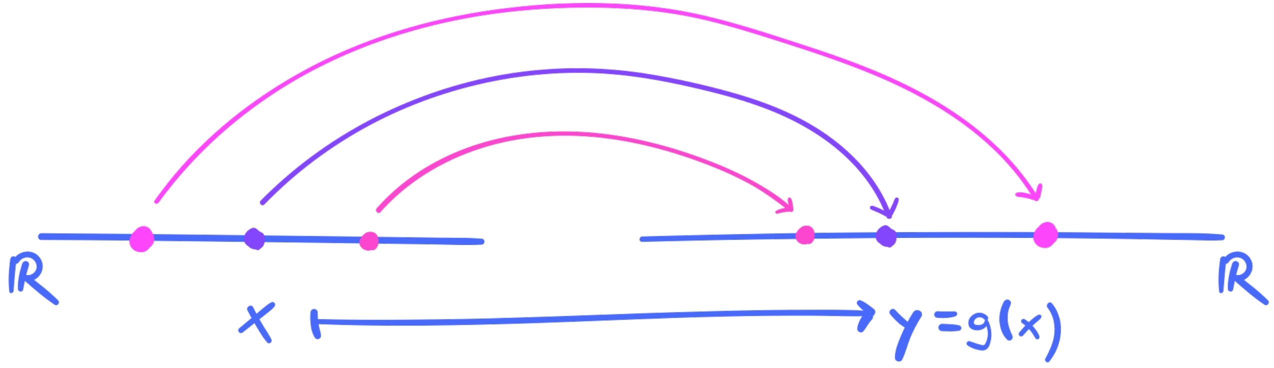
with equality if and only if $P$ is uniform.

**Problem Prompt**

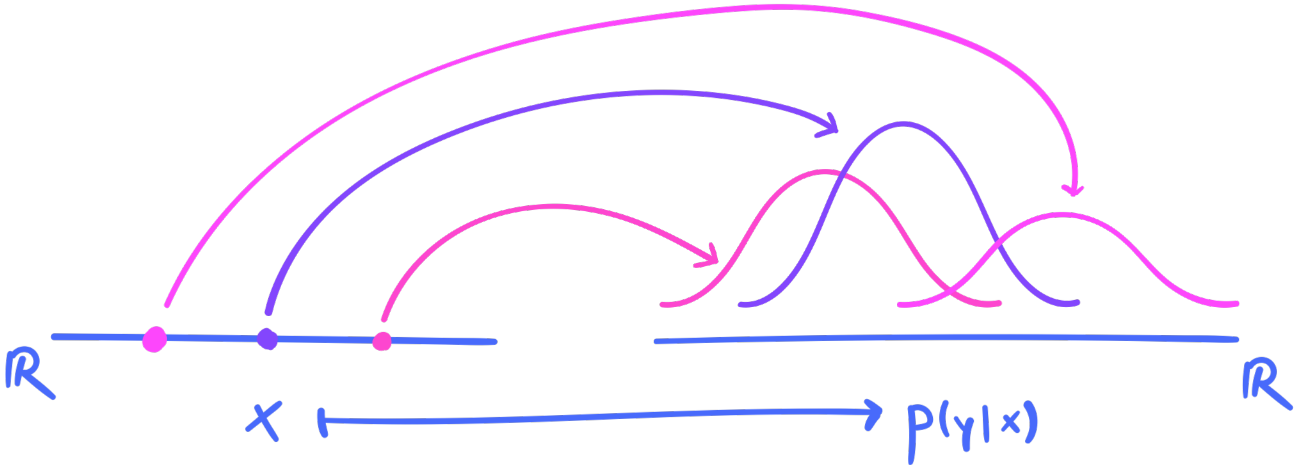
Do problem 5 on the worksheet.

# 10.3. Flow of information

## Deterministic link




## Stochastic link (Markov Kernel)



**Definition 10.5**

A *Markov kernel* is a mapping

$$\kappa : \{1, 2, \ldots, m\} \to \mathbb{R}^n$$

such that each vector $\kappa(i) \in \mathbb{R}^n$ is a probability vector (i.e., a vector with nonnegative entries that sum to 1). The $m \times n$ matrix

$$\mathbf{K} = \begin{bmatrix} \leftarrow & \kappa(1)^\intercal & \rightarrow \\ \vdots & \vdots & \vdots \\ \leftarrow & \kappa(m)^\intercal & \rightarrow \end{bmatrix}$$

is called the *transition matrix* of the Markov kernel.

**Definition 10.6**

A *communication channel* is a Markov kernel.

**Theorem 10.4 (Conditional distributions determine communication channels)**

Let $X$ and $Y$ be two random variables with finite ranges

$$\{x_1, \ldots, x_m\} \quad \text{and} \quad \{y_1, \ldots, y_n\}. \tag{10.8}$$

Then the matrix

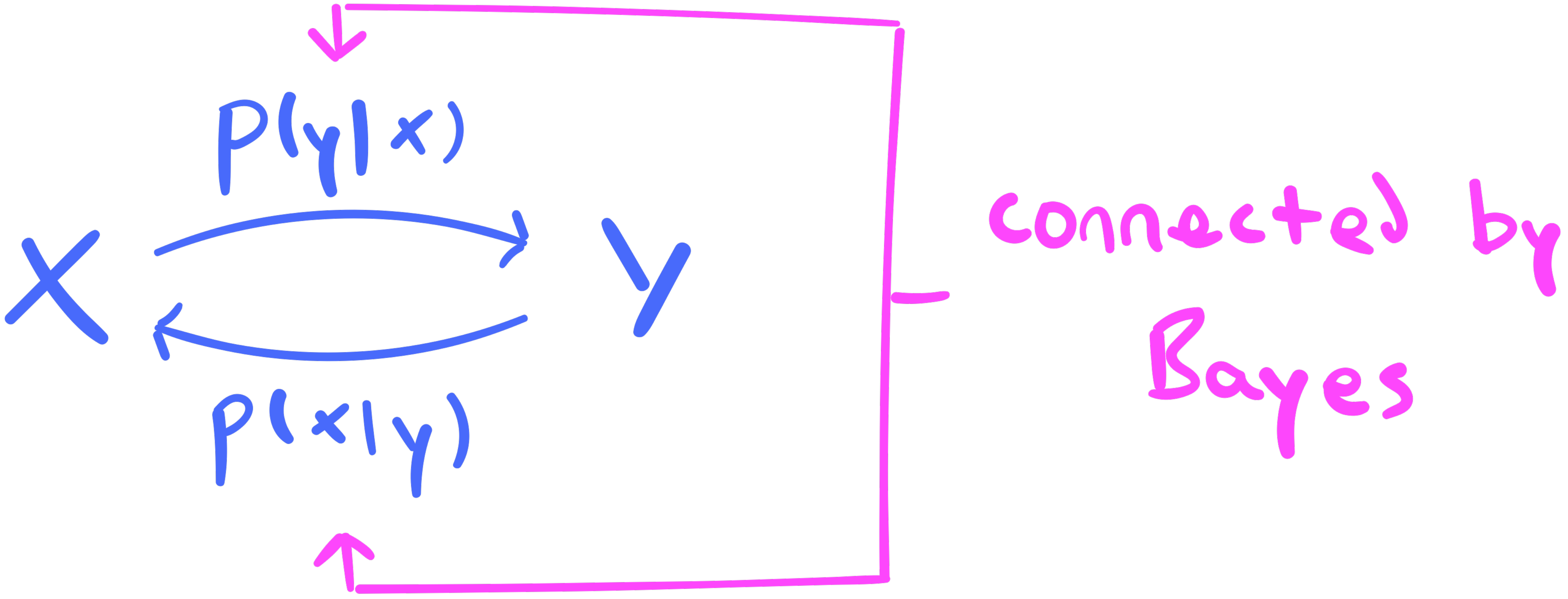
$$\mathbf{K} = [p(y_j|x_i)] = \begin{bmatrix} p(y_1|x_1) & \cdots & p(y_n|x_1) \\ \vdots & \ddots & \vdots \\ p(y_1|x_m) & \cdots & p(y_n|x_m) \end{bmatrix} \tag{10.9}$$

is the transition matrix of a Markov kernel.

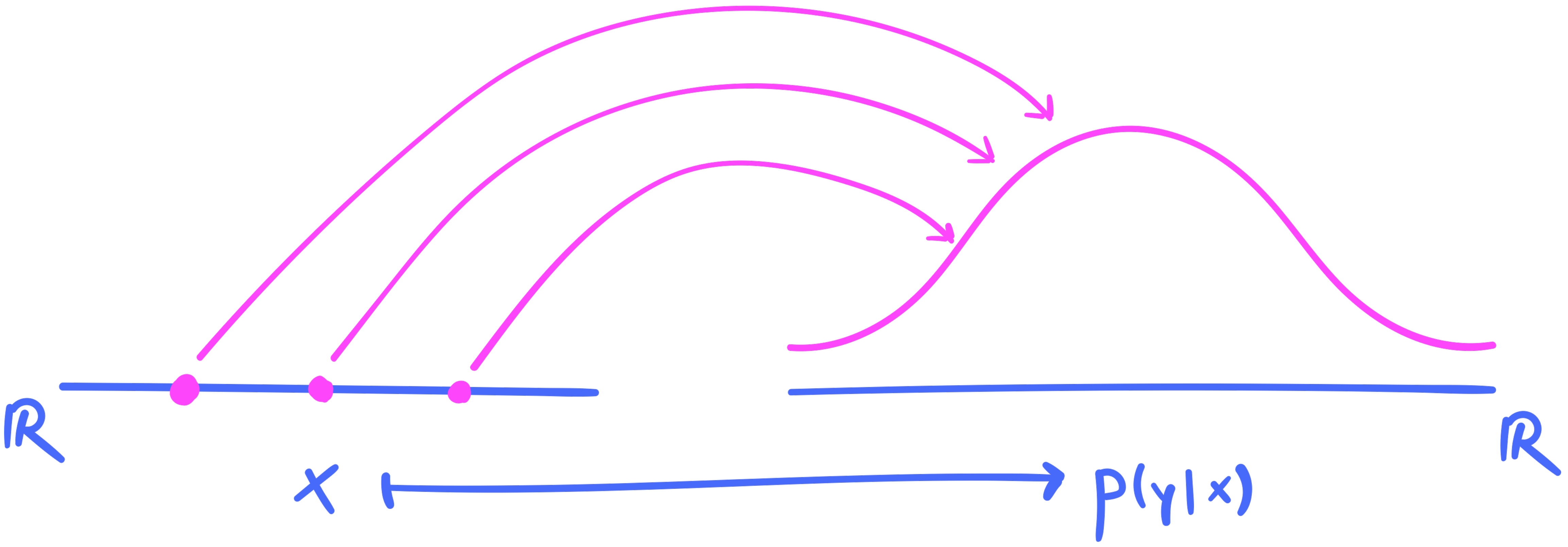
**Problem Prompt**

Do Problem 6 on the worksheet.

# Flow of information



Constant communication channel
= no information transfer



**Theorem 10.5 (Independence $=$ constant Markov kernels)**

Let $X$ and $Y$ be two random variables with finite ranges

$$\{x_1,\ldots,x_m\} \quad \text{and} \quad \{y_1,\ldots,y_n\}.$$

Then the induced communication channel

$$\kappa : \{1,2,\ldots,m\} \to \mathbb{R}^n, \quad \kappa(i)^\intercal = \begin{bmatrix} p(y_1|x_i) & \cdots & p(y_n|x_i) \end{bmatrix},$$

is constant if and only if $X$ and $Y$ are independent. In this case, $\kappa(i) = \boldsymbol{\pi}(Y)$ for each $i = 1,\ldots,m$.

### Definition 10.7

Let $X$ and $Y$ be two random variables with finite ranges. The *mutual information* shared between $X$ and $Y$, denoted $I(X, Y)$, is the KL divergence

$$I(X,Y) \stackrel{\text{def}}{=} D(P_{XY} \parallel P_X \otimes P_Y) = \sum_{(x,y)\in\mathbb{R}^2} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right).$$

**Problem Prompt**

Do Problem 7 on the worksheet.

**Theorem 10.6 (Independence $=$ zero mutual information)**

Let $X$ and $Y$ be two random variables with finite ranges

$$\{x_1, \ldots, x_m\} \quad \text{and} \quad \{y_1, \ldots, y_n\}.$$

Then the following statements are equivalent:

1. The induced communication channel

$$\kappa : \{1, 2, \ldots, m\} \to \mathbb{R}^n, \quad \kappa(i)^\intercal = \begin{bmatrix} p(y_1|x_i) & \cdots & p(y_n|x_i) \end{bmatrix},$$

   is constant.
2. The random variables $X$ and $Y$ are independent.
3. The mutual information $I(X, Y) = 0$.

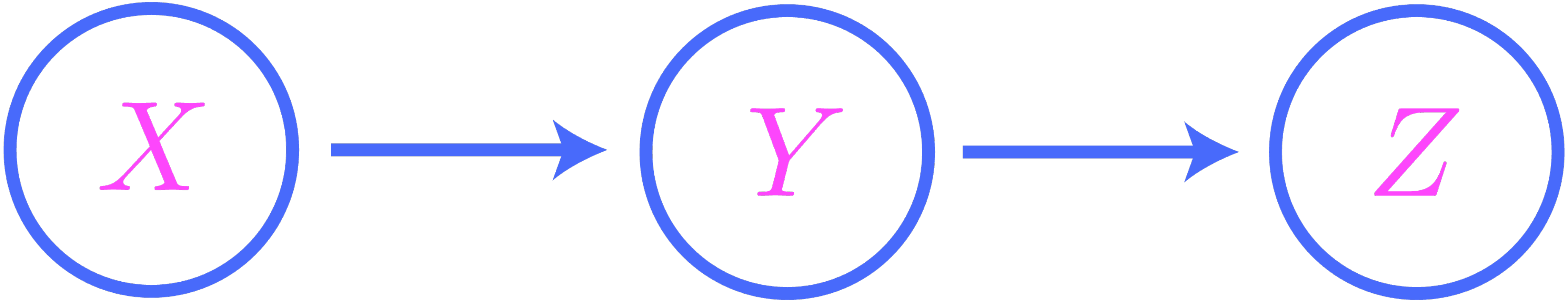
**Theorem 10.7 (Mutual information and entropy)**

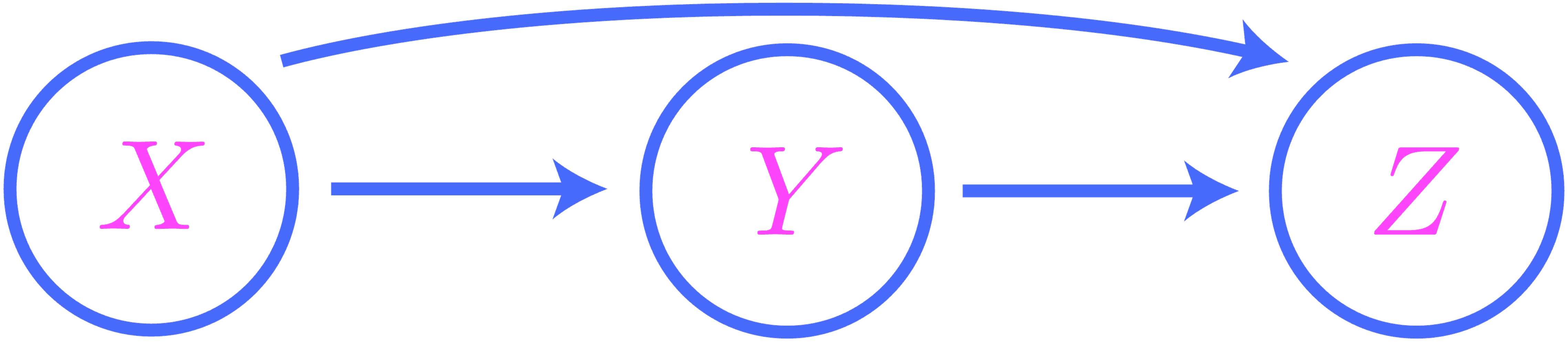
Let $X$ and $Y$ be two random variables with finite ranges. Then:

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

**Corollary 10.2 (Symmetry of mutual information)**

Let $X$ and $Y$ be random variables with finite ranges. Then $I(X, Y) = I(Y, X)$.


X
Y
Z


X
Y
Z

**Definition 10.8**

Let $X$, $Y$, and $Z$ be three random variables.

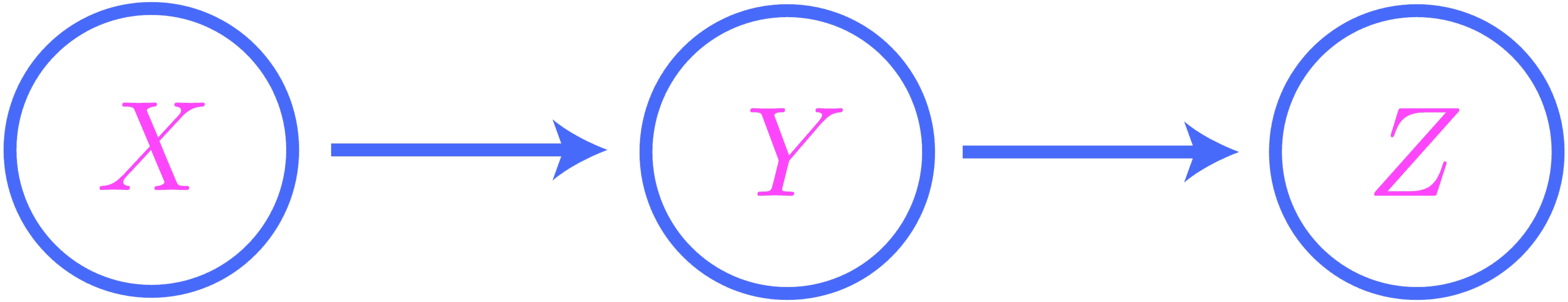
1. If the variables are jointly discrete, then we shall say $X$ and $Z$ are *conditionally independent* given $Y$ if

   $$p(x, z|y) = p(x|y)p(z|y)$$

   for all $x$, $y$, and $z$.
2. If the variables are jointly continuous, then we shall say $X$ and $Z$ are *conditionally independent* given $Y$ if

   $$f(x, z|y) = f(x|y)f(z|y)$$

   for all $x$, $y$, and $z$.


X
Y
Z

**Theorem 10.8 (Data Processing Inequality)**

Suppose $X$, $Y$, and $Z$ are three random variables with finite ranges, and suppose that $X$ and $Z$ are conditionally independent given $Y$. Then

$$I(X, Z) \leq I(X, Y), \tag{10.11}$$

with equality if and only if $X$ and $Y$ are independent given $Z$.