10. Information theory

10.1. Shannon information and entropy

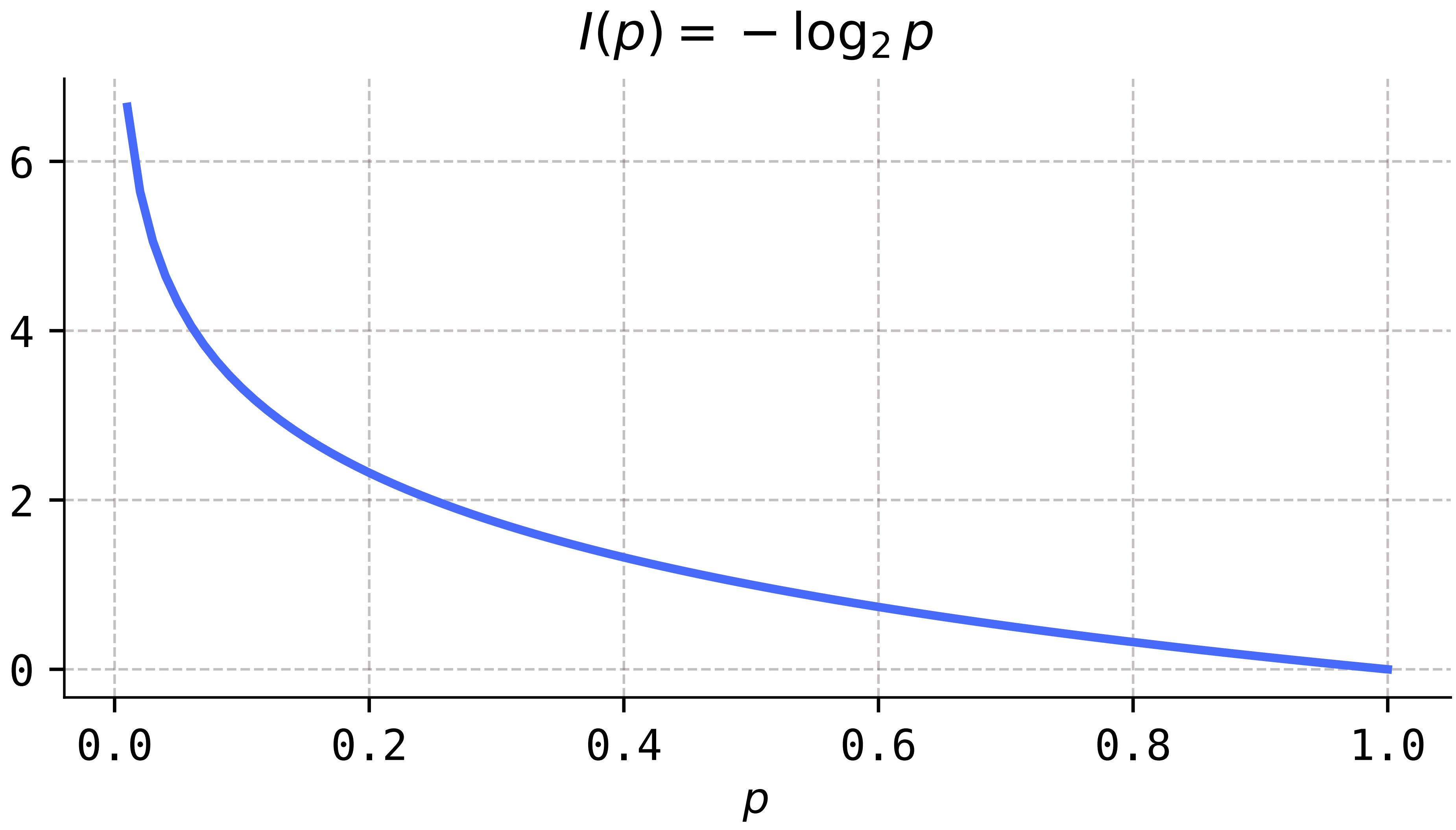
Definition 10.1

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

$$I_P(s) \stackrel{\text{def}}{=} -\log_2(p(s)).$$

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 $I_{P_{\mathbf{X}}}(\mathbf{x})$.


$I(p) = -\log_2 p$
6
4
2
0
0.0
0.2
0.4
0.6
0.8
1.0
p

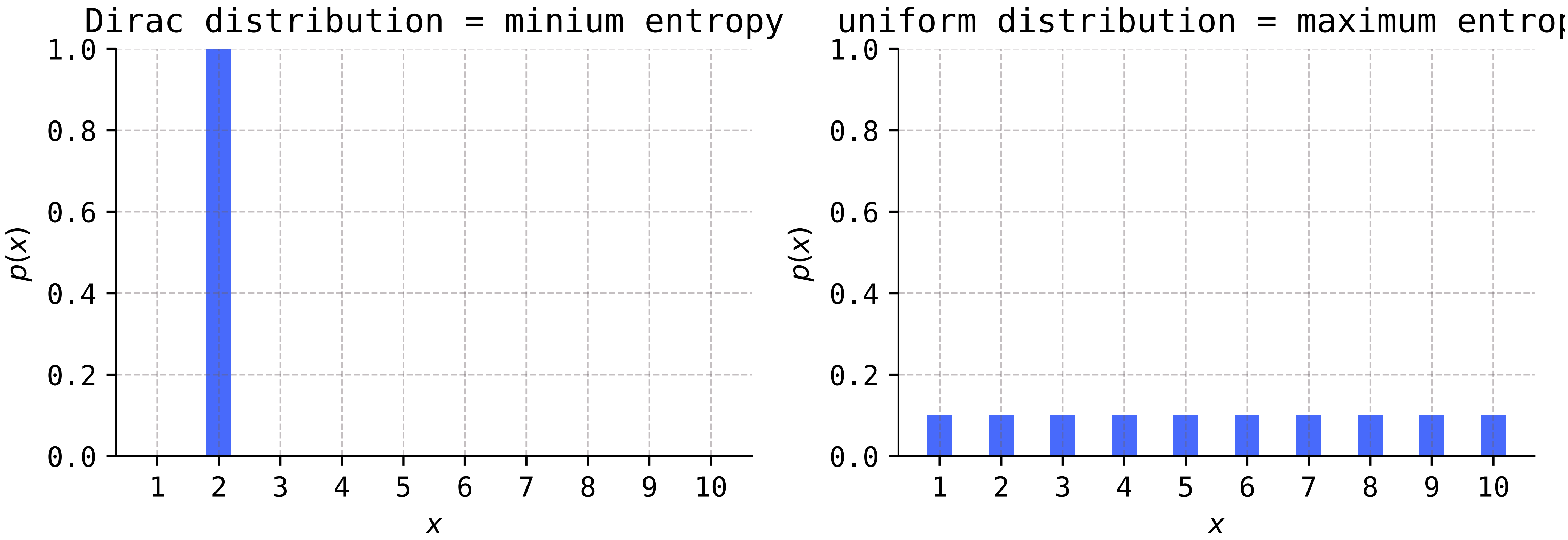
Definition 10.2

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 = minium entropy
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
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

$$H_P(Q) \stackrel{\text{def}}{=} E_{s\sim p(s)}\left[I_Q(s)\right] = -\sum_{s\in S} p(s)\log_2(q(s)).$$

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.