

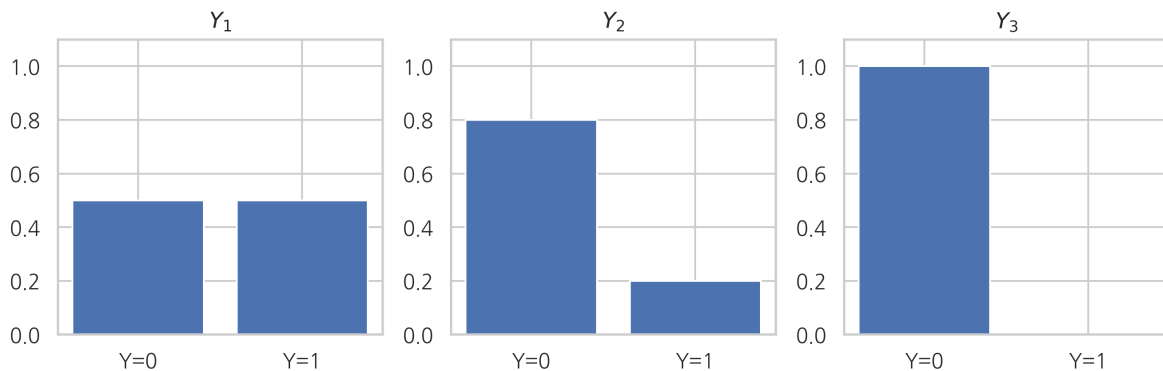
## 10.1 엔트로피

$Y = 0$  또는  $Y = 1$  인 두 가지 값을 가지는 확률변수의 확률분포가 다음과 같이 세 종류가 있다고 하자.

- 확률분포  $Y_1$  :  $P(Y = 0) = 0.5, P(Y = 1) = 0.5$
- 확률분포  $Y_2$  :  $P(Y = 0) = 0.8, P(Y = 1) = 0.2$
- 확률분포  $Y_3$  :  $P(Y = 0) = 1.0, P(Y = 1) = 0.0$

In [1]:

```
plt.figure(figsize=(9, 3))
plt.subplot(131)
plt.bar([0, 1], [0.5, 0.5])
plt.xticks([0, 1], ["Y=0", "Y=1"])
plt.ylim(0, 1.1)
plt.title("$Y_1$")
plt.subplot(132)
plt.bar([0, 1], [0.8, 0.2])
plt.xticks([0, 1], ["Y=0", "Y=1"])
plt.ylim(0, 1.1)
plt.title("$Y_2$")
plt.subplot(133)
plt.bar([0, 1], [1.0, 0.0])
plt.xticks([0, 1], ["Y=0", "Y=1"])
plt.ylim(0, 1.1)
plt.title("$Y_3$")
plt.tight_layout()
plt.show()
```



베이지안 관점에서 위 확률분포는 다음과 같은 정보를 나타낸다.

- 확률분포  $Y_1$  은  $y$  값에 대해 아무것도 모르는 상태
- 확률분포  $Y_2$  은  $y$  값이 0이라고 믿지만 아닐 가능성도 있다는 것을 아는 상태
- 확률분포  $Y_3$  은  $y$  값이 0이라고 100% 확신하는 상태

확률분포가 가지는 이러한 차이를 하나의 숫자로 나타낸 것이 바로 엔트로피다.

### 엔트로피의 정의

**엔트로피(entropy)**는 확률분포가 가지는 정보의 확신도 혹은 정보량을 수치로 표현한 것이다. 확률분포에서 특정한 값이 나올 확률이 높아지고 나머지 값의 확률은 낮아진다면 엔트로피가 작아진다. 반대로 여러가지 값이 나

올 확률이 대부분 비슷한 경우에는 엔트로피가 높아진다. 엔트로피는 확률분포의 모양이 어떤지를 나타내는 특성값 중 하나로 볼 수도 있다. 확률 또는 확률밀도가 특정값에 몰려있으면 엔트로피가 작다고 하고 반대로 여러가지 값에 골고루 퍼져 있다면 엔트로피가 크다고 한다.

확률분포의 엔트로피는 물리학의 엔트로피 용어를 빌려온 것이다. 물리학에서는 물질의 상태가 분산되는 정도를 엔트로피로 정의한다. 물체의 상태가 여러가지로 고루 분산되어 있으면 엔트로피가 높고 특정한 하나의 상태로 몰려있으면 엔트로피가 낮다.

수학적으로 엔트로피는 확률분포함수를 입력으로 받아 숫자를 출력하는 범함수(functional)로 정의한다.  $H[]$  기호로 표기한다.

확률변수  $Y$ 가 카테고리분포와 같은 이산확률변수이면 다음처럼 정의한다.

$$H[Y] = - \sum_{k=1}^K p(y_k) \log_2 p(y_k) \quad (10.1.1)$$

이 식에서  $K$ 는  $X$ 가 가질 수 있는 클래스의 수이고  $p(y)$ 는 확률질량함수다. 확률의 로그값이 항상 음수이므로 음수 기호를 붙여서 양수로 만들었다.

확률변수  $Y$ 가 정규분포와 같은 연속확률변수이면 다음처럼 정의한다.

$$H[Y] = - \int_{-\infty}^{\infty} p(y) \log_2 p(y) dy \quad (10.1.2)$$

이 식에서  $p(y)$ 는 확률밀도함수다.

로그의 밑(base)이 2로 정의된 것은 정보통신과 관련을 가지는 역사적인 이유 때문이다.

엔트로피 계산에서  $p(y) = 0$ 인 경우에는 로그값이 정의되지 않으므로 다음과 같은 극한값을 사용한다.

$$\lim_{p \rightarrow 0} p \log_2 p = 0 \quad (10.1.3)$$

이 값은 로피탈의 정리(L'Hôpital's rule)에서 구할 수 있다.

위에서 예를 든  $Y_1, Y_2, Y_3$  3 개의 이산확률분포에 대해 엔트로피를 구하면 다음과 같다.

$$H[Y_1] = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1 \quad (10.1.4)$$

$$H[Y_2] = -\frac{8}{10} \log_2 \frac{8}{10} - \frac{2}{10} \log_2 \frac{2}{10} \approx 0.72 \quad (10.1.5)$$

$$H[Y_3] = -1 \log_2 1 - 0 \log_2 0 = 0 \quad (10.1.6)$$

다음은 넘파이로 엔트로피를 계산한 결과다. 확률값이 0일 때는 가장 작은 값인 `eps` 를 대신 사용한다.

In [2]:

```
-0.5 * np.log2(0.5) - 0.5 * np.log2(0.5)
```

Out[2]:

1.0

In [3]:

```
-0.8 * np.log2(0.8) - 0.2 * np.log2(0.2)
```

Out[3]:

0.7219280948873623

In [4]:

```
eps = np.finfo(float).eps  
-1 * np.log2(1) - eps * np.log2(eps)
```

Out[4]:

1.1546319456101628e-14

### 연습 문제 10.1.1

베르누이분포에서 확률값  $P(Y = 1)$ 은 0부터 1까지의 값을 가질 수 있다. 각각의 값에 대해 엔트로피를 계산하여 가로축이  $P(Y = 1)$ 이고 세로축이  $H[Y]$ 인 그래프를 그려라.

### 연습 문제 10.1.2

다음 확률분포의 엔트로피를 계산하라.

$$(1) P(Y = 0) = \frac{1}{8}, P(Y = 1) = \frac{1}{8}, P(Y = 2) = \frac{1}{4}, P(Y = 3) = \frac{1}{2}$$

$$(2) P(Y = 0) = 1, P(Y = 1) = 0, P(Y = 2) = 0, P(Y = 3) = 0$$

$$(3) P(Y = 0) = \frac{1}{4}, P(Y = 1) = \frac{1}{4}, P(Y = 2) = \frac{1}{4}, P(Y = 3) = \frac{1}{4}$$

## 엔트로피의 성질

확률변수가 결정론적이면 확률분포에서 특정한 하나의 값이 나올 확률이 1이다. 이 때 엔트로피는 0이 되고 이 값은 엔트로피가 가질 수 있는 최솟값이다.

반대로 엔트로피의 최대값은 이산 확률변수의 클래스의 갯수에 따라 달라진다. 만약 이산확률분포가 가질 수 있는 값이  $2^K$  개면 엔트로피의 최대값은 각 값에 대한 확률이 모두 같은 값인  $\frac{1}{2^K}$ 이다. 엔트로피의 값은

$$H = -2^K \cdot \frac{1}{2^K} \log_2 \frac{1}{2^K} = K \quad (10.1.7)$$

이다.

## 엔트로피의 추정

이론적인 확률밀도함수가 없고 실제 데이터가 주어진 경우에는 데이터에서 확률질량함수를 추정한 후, 이를 기반으로 엔트로피를 계산한다.

예를 들어 데이터가 모두 80개가 있고 그 중  $Y = 0$  인 데이터가 40개,  $Y = 1$ 인 데이터가 40개 있는 경우는 엔트로피가 1이다.

$$P(y = 0) = \frac{40}{80} = \frac{1}{2} \quad (10.1.8)$$

$$P(y = 1) = \frac{40}{80} = \frac{1}{2} \quad (10.1.9)$$

$$H[Y] = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = \frac{1}{2} + \frac{1}{2} = 1 \quad (10.1.10)$$

사이파이의 stats 서브패키지는 엔트로피를 구하는 `entropy` 함수를 제공한다. `base` 인수값은 2가 되어야 한다.

In [5]:

```
p = [0.5, 0.5]
sp.stats.entropy(p, base=2)
```

Out[5]:

1.0

### 연습 문제 10.1.3

- (1) 데이터가 모두 60개가 있고 그 중  $Y = 0$ 인 데이터가 20개,  $Y = 1$ 인 데이터가 40개 있는 경우의 엔트로피를 계산하라.
- (2) 데이터가 모두 40개가 있고 그 중  $Y = 0$ 인 데이터가 30개,  $Y = 1$ 인 데이터가 10개 있는 경우의 엔트로피를 계산하라.
- (3) 데이터가 모두 20개가 있고 그 중  $Y = 0$ 인 데이터만 20개 있는 경우의 엔트로피를 계산하라.

## 가변길이 인코딩

엔트로피는 원래 통신 분야에서 데이터가 가지고 있는 정보량을 계산하기 위해 고안되었다. 예를 들어 4개의 글자 A, B, C, D로 씌여진 다음과 같은 문서가 있다고 하자.

In [6]:

```
N = 200
p = [1/2, 1/4, 1/8, 1/8]
doc0 = list("".join([int(N * p[i]) * c for i, c in enumerate("ABCD")]))
np.random.shuffle(doc0)
doc = "".join(doc0)
doc
```

Out[6]:

```
'BDABABACBABBAAACADAADAADAAAAAABBAABADAAAAABACAAACBBACDBAAACBCABBAABAAAAADDBABCBDB
BDDBAABBBADCAADAADCABADCAAAAACADBAABABCBAACAAABCDAAADCCGAAABABBDACACAAAAAABABBADAB
BABDBADBACAABDCAABAAABACCDABAABA'
```

이 문서를 0과 1로 이루어진 이진수로 변환해야 하면 보통 다음처럼 인코딩한다.

- A = "00"
- B = "01"
- C = "10"
- D = "11"

이렇게 인코딩을 하면 200 글자로 이루어진 문서는 이진수 400개가 된다.

In [7]:

```
encoder = {"A": "00", "B": "01", "C": "10", "D": "11"}
encoded_doc = "".join([encoder[c] for c in doc])
encoded_doc
```

Out[7]:

```
'011100010001001001000101000010001100001100000011000000000000101000010011000000000
00101001000000100100101001011010000001001100001010000100000000001111010001100111010
11111010000010101001110001000001100001110000100111000000000010001101000001000110010
000100000000110110000111110101000000001000101110010001000000000000001000101001100010
10001110100110100100000011110000000010000000100101011000100000100'
```

In [8]:

```
len(encoded_doc)
```

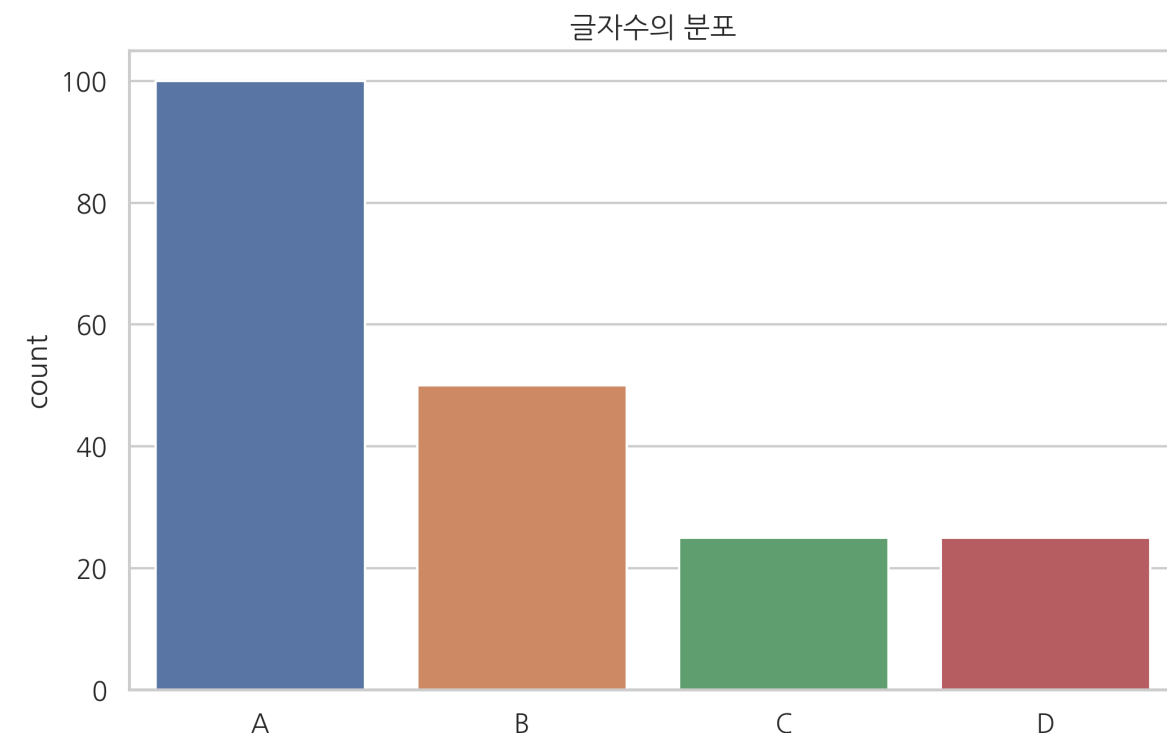
Out[8]:

400

그런데 이진수로 변환할 때 더 글자수를 줄일 수 있는 방법이 있다. 우선 위 글자의 분포를 조사하자.

In [9]:

```
sns.countplot(list(doc), order="ABCD")
plt.title("글자수의 분포")
plt.show()
```



글자수의 분포가 다음과 같다.

$$P(Y = A) = \frac{1}{2}, P(Y = B) = \frac{1}{4}, P(Y = C) = \frac{1}{8}, P(Y = D) = \frac{1}{8} \quad (10.1.11)$$

지프의 법칙(Zipf's law)에 따르면 이러한 분포는 현실의 글자 빈도수에서도 흔히 나타난다.

확률분포가 위와 같을 때는 다음처럼 인코딩하면 인코딩된 후의 이진수 수를 줄일 수 있다.

- A = "0"
- B = "10"
- C = "110"
- D = "111"

이 방법은 글자마다 인코딩하는 이진수의 숫자가 다르기 때문에 **가변길이 인코딩(variable length encoding)**이라고 한다. 가장 많이 출현하는 'A'는 두 글자가 아닌 한 글자이므로 인코딩 후의 이진수 수가 감소한다. 반대로 'C', 'D'는 이진수의 수가 3개로 많지만 글자의 빈도가 적어서 영향이 적다.

만약 문서의 분포가 위에서 가정한 분포와 정확하게 같다면 인코딩된 이진수의 숫자는 다음 계산에서 350개가 됨을 알 수 있다.

$$\left(200 \times \frac{1}{2}\right) \cdot 1 + \left(200 \times \frac{1}{4}\right) \cdot 2 + \left(200 \times \frac{1}{8}\right) \cdot 3 + \left(200 \times \frac{1}{8}\right) \cdot 3 = 350 \quad (10.1.12)$$

따라서 알파벳 한 글자를 인코딩하는데 필요한 평균 비트(bit)수는  $350 \div 200 = 1.75$ 이고 이 값은 확률변수의 엔트로피 값과 같다.

$$H = -\frac{1}{2}\log_2 \frac{1}{2} - \frac{1}{4}\log_2 \frac{1}{4} - \frac{2}{8}\log_2 \frac{1}{8} = 1.75 \quad (10.1.13)$$

In [10]:

```
vl_encoder = {"A": "0", "B": "10", "C": "110", "D": "111"}
vl_encoded_doc = "".join([vl_encoder[c] for c in doc])
vl_encoded_doc
```

Out[10]:

```
'10111010010011010010100011001110011100011100000010100010011100000101001100010011010
10011011110000110101100101000100000011111110010110101111010111111000101010011111001
100011100111110010011111000000110011110001001011010001100001011011100111111110110110
000100101011101100110000000100101001110101001011110011110011000101111100001000010011
011011101000100'
```

In [11]:

```
len(vl_encoded_doc)
```

Out[11]:

350

In [12]:

```
sp.stats.entropy([1/2, 1/4, 1/8, 1/8], base=2)
```

Out [12]:

1.75

#### 연습 문제 10.1.4

A, B, C, D, E, F, G, H의 8글자로 이루어진 문서가 있고 각각의 글자가 나올 확률이 다음과 같다고 가정하자.

$$\left\{ \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64} \right\} \quad (10.1.14)$$

이 문서를 위한 가변길이 인코딩 방식을 서술하고 한 글자를 인코딩하는데 필요한 평균 비트수를 계산하라.

### 지니불순도

엔트로피와 유사한 개념으로 **지니불순도(Gini impurity)**라는 것이 있다. 지니불순도는 엔트로피처럼 확률분포가 어느쪽에 치우쳐있는가를 재는 척도지만 로그를 사용하지 않으므로 계산량이 더 적어 엔트로피 대용으로 많이 사용된다. 경제학에서도 사용되지만 지니계수(Gini coefficient)와는 다른 개념이라는 점에 주의해야 한다.

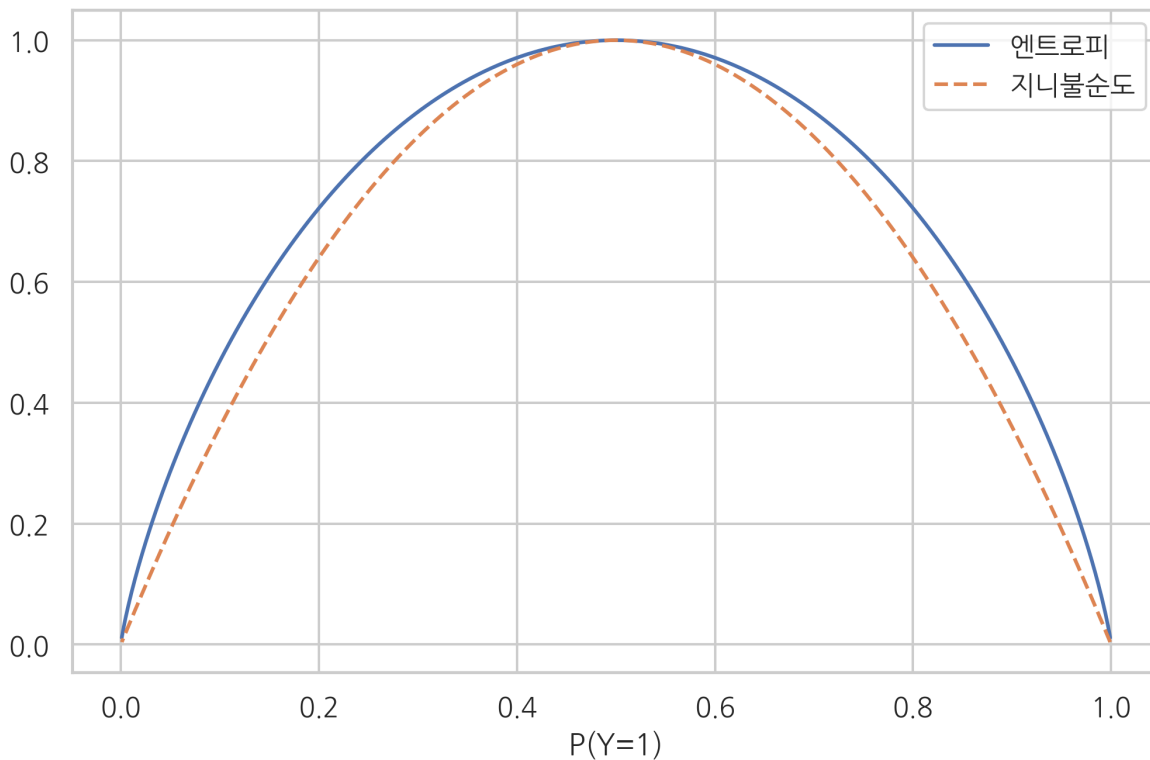
$$G[Y] = \sum_{k=1}^K P(y_k)(1 - P(y_k)) \quad (10.1.15)$$

다음 그림은 값이 두 개인 이산확률분포에서 지니불순도와 엔트로피를 비교한 결과다.

In [13]:

```
P0 = np.linspace(0.001, 1 - 0.001, 1000)
P1 = 1 - P0
H = - P0 * np.log2(P0) - P1 * np.log2(P1)
G = 2 * (P0 * (1 - P0) + P1 * (1 - P1))

plt.plot(P1, H, "-", label="엔트로피")
plt.plot(P1, G, "--", label="지니불순도")
plt.legend()
plt.xlabel("P(Y=1)")
plt.show()
```



## 엔트로피 최대화

기댓값 0, 분산  $\sigma^2$ 이 주어졌을 때 엔트로피  $H[p(x)]$ 를 가장 크게 만드는 확률밀도함수  $p(x)$ 는 정규분포가 된다. 이는 다음처럼 증명한다.

우선 확률밀도함수가 지켜야할 제한조건은 다음과 같다.

(1) 확률밀도함수의 총면적은 1

$$\int_{-\infty}^{\infty} p(x)dx = 1 \quad (10.1.16)$$

(2) 기댓값은 0

$$\int_{-\infty}^{\infty} xp(x)dx = 0 \quad (10.1.17)$$

(3) 분산은  $\sigma^2$



$$\int_{-\infty}^{\infty} x^2 p(x) dx = \sigma^2 \quad (10.1.18)$$

최대화할 목적범함수(objective functional)은 엔트로피다.

$$H[p(x)] = - \int_{-\infty}^{\infty} p(x) \log p(x) dx \quad (10.1.19)$$

라그랑주 승수법으로 제한조건을 추가하면 다음과 같아진다.

$$\begin{aligned} H[p(x)] &= - \int_{-\infty}^{\infty} p(x) \log p(x) dx + \lambda_1 \left( \int_{-\infty}^{\infty} p(x) dx - 1 \right) \\ &\quad + \lambda_2 \left( \int_{-\infty}^{\infty} x p(x) dx \right) + \lambda_3 \left( \int_{-\infty}^{\infty} x^2 p(x) dx - \sigma^2 \right) \\ &= \int_{-\infty}^{\infty} (-p(x) \log p(x) + \lambda_1 p(x) + \lambda_2 x p(x) + \lambda_3 x^2 p(x) - \lambda_1 - \lambda_3 \sigma^2) dx \end{aligned} \quad (10.1.20)$$

변분법에서 도함수는 다음과 같이 계산된다.

$$\frac{\delta H}{\delta p(x)} = -\log p(x) - 1 + \lambda_1 + \lambda_2 x + \lambda_3 x^2 = 0 \quad (10.1.21)$$

따라서 확률밀도함수의 형태는 다음과 같다.

$$p(x) = \exp(-1 + \lambda_1 + \lambda_2 x + \lambda_3 x^2) \quad (10.1.22)$$

적분을 통해 위 형태의 확률밀도함수의 면적, 기대값, 분산을 계산하고 주어진 제한조건을 만족하도록 연립방정식을 풀면 라그랑주 승수를 다음처럼 구할 수 있다. 이 과정은 이 책의 범위를 벗어나므로 생략한다.

$$\begin{aligned} \lambda_1 &= 1 - \frac{1}{2} \log 2\pi\sigma^2 \\ \lambda_2 &= 0 \\ \lambda_3 &= -\frac{1}{2\sigma^2} \end{aligned} \quad (10.1.23)$$

이 값을 대입하면 정규분포라는 것을 알 수 있다.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \quad (10.1.24)$$

따라서 정규분포는 기댓값과 표준편차를 알고있는 확률분포 중에서 가장 엔트로피가 크고 따라서 가장 정보가 적은 확률분포다. 정규분포는 베이즈 추정에서 있어서 사실상의 무정보 사전확률분포로 사용되는 경우가 많다.

## 연습 문제 10.1.5

정규분포의 확률밀도함수에 대한 엔트로피값이 다음과 같음을 증명하라.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \quad (10.1.25)$$

$$H[p(x)] = \frac{1}{2}(1 + \log 2\pi\sigma^2) \quad (10.1.26)$$

