1 Multi-Armed Bandit

Consider a casino owner who presents a gambler with a set of $N$ coins, of which $k < N$ are biased in favor of the gambler. In order for the casino owner to ensure that the regret of the gambler is high:

1. there should be few favorably biased coins, i.e. $k << N$
2. the biased coins should not be too favorable; otherwise, the gambler will win with little regret.
3. the biased coins should sufficiently favorable

Recall that given a bias $\epsilon$, the gambler must flip $T = O\left(\frac{N}{\epsilon^2}\right)$ coins. Therefore the optimal bias is given as

$$\epsilon = O\left(\sqrt{\frac{N}{T}}\right)$$

With this bias, the gambler has at least regret of

$$\epsilon = O\left(\sqrt{NT}\right)$$

The best known algorithm for finding the biased coin incurs a regret of $O\left(\sqrt{NT\log N}\right)$. The algorithm is as follows:

1. with probability $p$, select a random coin
2. with probability $1-p$, toss according to the best estimate
3. $p$ is the only hyperparameter and is selected as a function of $N$ and $T$

2 Differential Entropy

We can define entropy over a continuous random variable, much in the same way we did for discrete random variables. The differential entropy of a random variable $X$ is defined as

$$h(X) \triangleq -\int f_X(x) \ln f_X(x) dx$$

2.1 Differential Entropy of a Uniform Random Variable

Recall a uniform distribution $f$ over a random variable $X$ is defined as

$$f_X(x) = \begin{cases} \frac{1}{a} & x \in [0, a] \\ 0 & otherwise \end{cases}$$

The differential entropy of $X$ is


\[ h(X) = \int_0^a \frac{1}{a} \ln \frac{1}{a} \, dx \]
\[ = \frac{1}{a} \ln a \int_0^a \, dx \]
\[ = \frac{1}{a} \ln a = \ln a \]

2.2 Properties of Differential Entropy

**Theorem**: Given any two random variables \( X \) and \( Y = X + a \),

\[ h(Y) = h(X) \]  \hspace{1cm} (1)

**Proof**:

\[ h(Y) = -\int f_Y(y) \ln f_Y(y) \, dy \]
\[ = -\int f_X(y-a) \ln f_X(y-a) \, dy \]
\[ = -\int f_X(x) \ln f_X(x) \, dx = h(X) \]

The equivalence \( f_Y(y) = f_X(y-a) \) follow from:

\[ F_Y(y) = P(Y \leq y) \]
\[ = P(x + a \leq y) \]
\[ = P(x \leq y - a) \]
\[ = F_X(y-a) \]

2.3 Differential Entropy of a Normal Random Variable

Consider \( X \sim \mathcal{N}(\mu, \sigma^2) \). Recall the equation of the normal distribution is

\[ f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

As a result of theorem 1, we can safely drop \( \mu \) from the equation. That is, the entropy of \( \sim \mathcal{N}(\mu, \sigma^2) \) is equal to the entropy of \( \sim \mathcal{N}(0, \sigma^2) \), because \( \mu \) only shifts the distribution. The area under the normal distribution remains unchanged. So let

\[ f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \]

The differential entropy of \( X \) is
\[ h(X) = - \int f_X(x) \ln \left( \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right) dx \]
\[ = \int \frac{1}{2} f_X(x) \ln(2\pi\sigma^2) dx + \int \frac{f_X(x)x^2}{2\sigma^2} dx \]
\[ = \frac{1}{2} \ln(2\pi\sigma^2) \int f_X(x) dx + \frac{1}{2\sigma^2} \int f_X(x)x^2 dx \]

The first integral is equivalent to 1, by the integrate-to-one constraint on valid probability distributions. The second integral is exactly the second moment of the normal distribution, which is simply the variance \( \sigma^2 \). Then

\[ h(X) = \frac{1}{2} \ln(2\pi\sigma^2) + \frac{\sigma^2}{2\sigma^2} \]
\[ = \frac{1}{2} (\ln(2\pi\sigma^2) + 1) \]
\[ = \frac{1}{2} (\ln(2\pi e\sigma^2)) \]

### 2.4 Quantization of the Probability Density Function

We can generate a discrete probability distribution from a continuous probability distribution over \( X \) by partitioning the \( x \)-axis into \( \delta \)-sized intervals. Then for any point in interval \( i \), the probability is

\[ p(x_i) = \int_{i\Delta}^{(i+1)\Delta} f_X(x) dx \approx f_X(x_i) \Delta \]

Then the discrete entropy is

\[ H(\hat{X}_\Delta) = - \sum_i p(x_i) \ln p(x_i) \]

where \( \hat{X}_{\Delta \delta} \) is the discrete analog of the continuous random variable \( X \), defined by \( \Delta \).

\[ H(\hat{X}_\Delta) = - \sum_i f(x_i) \Delta \ln(p(x_i)\Delta) \]
\[ = - \sum_i f(x_i) \Delta \ln p(x_i) + \ln \Delta \]
\[ = - \sum_i f(x_i) \Delta \ln p(x_i) - \sum_i f(x_i) \Delta \ln \Delta \]

Notice that

\[ \sum_i f_X(x_i) \Delta = \sum_i \int_{i\Delta}^{(i+1)\Delta} f_X(x) dx = \int f_X(x) dx = 1 \]
The first term, therefore, is \( h(X) \) and the second term becomes \( \ln \Delta \), giving

\[
H(\tilde{X}_\Delta) = h(X) - \ln \Delta
\]

Therefore the entropy of \( n \)-bit quantization of the distribution over \( X \) is \( h(X) + n \).

2.5 Maximal Entropy with Fixed Variance

**Theorem:** A Gaussian random variable has the highest entropy of all random variables with fixed variance.

**Proof:** Consider an arbitrary distribution \( g(x) \) with variance \( \sigma^2 \). Because \( g \) is a valid probability distribution,

\[
\int g(x)dx = 1
\]

And by definition of the second moment,

\[
\int x^2 g(x)dx = \sigma^2
\]

We want to show that \( h_f(X) - h_g(X) \geq 0 \).

\[
h_f(X) - h_g(X) = -\int f(x) \ln f(x)dx + \int g(x) \ln g(x)
\]

\[
= \int g(x) \ln \frac{g(x)}{f(x)}dx + \int g(x) \ln f(x)dx - \int f(x) \ln f(x)dx
\]

Notice that \( \int g(x) \ln \frac{g(x)}{f(x)}dx \) is the divergence \( D(g||f) \geq 0 \). Thus

\[
h_f(X) - h_g(X) = D(g||f) + \int g(x) \ln f(x)dx - \int f(x) \ln f(x)dx
\]

For the second and third term, we have

\[
\int g(x) \ln f(x)dx - \int f(x) \ln f(x)dx \geq \int g(x) \left( \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{x^2}{2\sigma^2} \right)dx - \int f(x) \left( \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{x^2}{2\sigma^2} \right)dx
\]

\[
= -\frac{1}{2\sigma^2}\sigma^2 + \frac{1}{2\sigma^2}\sigma^2 = 0
\]

Therefore, \( h_f(X) - h_g(X) \geq D(g||f) \geq 0 \), thus \( h_f(X) \geq h_g(X) \). In other words, given fixed energy (variance), Gaussian random variables are the most uncertain. This has some significant implications in the domain of quantum mechanics.

This means that

\[
m_xh(w) = -\frac{1}{2} \ln(2\pi e\sigma^2)
\]

We can express \( \sigma^2 \) in terms of \( m_xh(w) \) as follows:
\[ \sigma^2 = \frac{1}{2\pi e} e^{\max 2h(x)} \]

And this can be expressed as an inequality (because it may be difficult to compute the maximum), as follows:

\[ \sigma^2 \geq \frac{1}{2\pi e} e^{2h(x)} \]

### 2.6 Mean Square Error

Given a random variable \( X \) and an estimate \( \hat{X} \), the mean square error (MSE) is defined as

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

Therefore the mean square error is lower bounded by the variance. How do we know this inequality holds?

\[ E[(X - a)^2] = E[X^2 - 2Xa + a^2] \]
\[ = E[X^2] - 2aE[X] + a^2 \]

Taking the derivative in terms of \( a \) gives

\[ \frac{d}{da} (-) = -2E[X] + 2a \]

This is 0 when \( a = E[X] \) and it can be easily shown (by examining the sign of the second derivative) that this is a local minimum. Altogether this implies that

\[ \text{MSE} \geq \frac{1}{2\pi e} e^{2h(x)} \]

### 2.7 Parameter Estimation

Consider a family of distributions \( f(x; \theta) \) with parameterization \( \theta \in \Theta \). The Gaussian distribution is an example of such a family with parameters \( \theta = \mu, \sigma \).

In the problem of parameter estimation, we want to find a function \( T \) which estimates \( \theta \) given samples \( X_1, X_2, ..., X_n \) which minimizes the estimation error

\[ E_\theta[T(X_1, X_2, ..., X_n) - \theta] \]

if for all \( \theta \), this is 0, then the estimator is called unbiased. Also, it is said that \( T_1 \) dominates \( T_2 \) if \( \forall \theta:\n\]

\[ E_\theta[(T_1(X_1, X_2, ..., X_n) - \theta)^2] \leq E_\theta[(T_2(X_1, X_2, ..., X_n) - \theta)^2] \]

In other words \( T_1 \) has a smaller mean square error. In the next class we will define the Fisher information \( J(\theta) \) and show how this relates to the mean square error of an estimator.