1 Gaussian computations

We now derive the expressions for entropy and KL-divergence of Gaussian distributions, which often come in handy.

1.1 Differential entropy

For a one-dimensional Gaussian $X \sim N(\mu, \sigma^2)$ we can calculate the differential entropy as

$$h(X) = \int p(x) \cdot \frac{1}{\ln 2} \cdot \left( \frac{(x - \mu)^2}{2\sigma^2} + \frac{1}{2} \ln(2\pi\sigma^2) \right) dx$$

$$= \frac{1}{\ln 2} \cdot \left( \frac{1}{2} + \frac{1}{2} \ln(2\pi\sigma^2) \right)$$

$$= \frac{1}{2} \cdot \log(2\pi \cdot e \cdot \sigma^2).$$

For the $n$-dimensional case, we first consider a Gaussian variable $X$ with mean 0 and covariance $I_n$, which means that we can think of $X = (X_1, \ldots, X_n)$ where each $X_i$ is a one-dimensional Gaussian with mean 0 and variance 1. Using the chain-rule for differential entropy (check that it holds) we get

$$h(X) = h(X_1) + \cdots + h(X_n) = \frac{n}{2} \cdot \log(2\pi \cdot e).$$

Before computing the entropy of a general Gaussian variable, it is helpful to consider the following rule for change of variables.

Exercise 1.1 (Change of variables). Let $X$ be a random variable over $\mathbb{R}^n$ with associated density function $p_X$. Using the Jacobian for change of variables in integrals, check that

1. If $c \in \mathbb{R}^n$ is a fixed vector, then the density function for $Y = X + c$ is given by $p_Y(y) = p_X(y - c)$.

2. If $A \in \mathbb{R}^{n \times n}$ is a nonsingular matrix, then the density function for $Y = AX$ is given by $p_Y(y) = \frac{p_X(A^{-1}y)}{|A|}$, where $|A|$ denotes $|\det(A)|$.
Using the above, we can derive how the differential entropy of a random variable changes due to translation and scaling.

**Proposition 1.2.** Let $X$ be a continuous random variable over $\mathbb{R}^n$. Let $c \in \mathbb{R}^n$ and let $A \in \mathbb{R}^{n \times n}$ be a non-singular matrix. Then

1. $h(X + c) = h(X)$.
2. $h(AX) = h(X) + \log |A|$.

**Proof:** Let $p_X$ be the density function for $X$. For $Y = X + c$, we have

\[
h(Y) = \int_{\mathbb{R}^n} p_Y(y) \cdot \log \left( \frac{1}{p_Y(y)} \right) \, dy
\]

\[
= \int_{\mathbb{R}^n} p_X(y - c) \cdot \log \left( \frac{1}{p_X(y - c)} \right) \, dy
\]

\[
= \int_{\mathbb{R}^n} p_X(x) \cdot \log \left( \frac{1}{p_X(x)} \right) \, dx \quad \text{(substituting $x = y - c$)}
\]

\[
= h(X)
\]

Similarly, for $Y = AX$, we have

\[
h(Y) = \int_{\mathbb{R}^n} p_Y(y) \cdot \log \left( \frac{1}{p_Y(y)} \right) \, dy
\]

\[
= \int_{\mathbb{R}^n} \frac{p_X(A^{-1}y)}{|A|} \cdot \log \left( \frac{|A|}{p_X(A^{-1}y)} \right) \, dy
\]

\[
= \int_{\mathbb{R}^n} \frac{p_X(x)}{|A|} \cdot \log \left( \frac{|A|}{p_X(x)} \right) \cdot |A| \, dx \quad \text{(substituting $x = A^{-1}y$)}
\]

\[
= h(X) + \log(|A|).
\]

Using the fact that $Y \sim N(\mu, \Sigma)$ can be written as $Y = \Sigma^{1/2}X + \mu$, where $X = N(0, I_n)$ (check this!) we get that

\[
h(Y) = h(X) + \log \left( |\Sigma^{1/2}| \right) = \frac{n}{2} \cdot \log(2\pi \cdot e) + \frac{1}{2} \cdot \log |\Sigma|.
\]
1.2 KL-divergence

We can compute the KL-divergence of two Gaussian distributions $P = N(\mu_1, \sigma_1^2)$ and $Q = N(\mu_2, \sigma_2^2)$ as

$$D(P \parallel Q) = \int_{\mathbb{R}} p(x) \cdot \log \left( \frac{p(x)}{q(x)} \right) dx = \mathbb{E}_{x \sim p} \left[ \log \left( \frac{p(x)}{q(x)} \right) \right]$$

$$= \mathbb{E}_{x \sim p} \left[ \frac{1}{\ln 2} \cdot \ln \left( \frac{\exp \left( -\frac{(x - \mu_1)^2}{2\sigma_1^2} \right)}{\sqrt{2\pi\sigma_1}} \cdot \frac{\sqrt{2\pi\sigma_2}}{\exp \left( -\frac{(x - \mu_2)^2}{2\sigma_2^2} \right)} \right) \right]$$

$$= \frac{1}{\ln 2} \cdot \mathbb{E}_{x \sim p} \left[ \frac{(x - \mu_2)^2}{2\sigma_2^2} - \frac{(x - \mu_1)^2}{2\sigma_1^2} + \ln \left( \frac{\sigma_2^2}{\sigma_1^2} \right) \right]$$

$$= \frac{1}{\ln 2} \cdot \left( \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} + \ln \left( \frac{\sigma_2^2}{\sigma_1^2} \right) \right)$$

The above is a common way of showing that changing the parameters of a Gaussian distribution by a small amount does not alter the behavior of an algorithm using the corresponding random variable as input, by too much.

**Exercise 1.3.** Let $P$ and $Q$ be Gaussian distributions with means $\mu_1$ and $\mu_2$ respectively, and variance $\sigma^2$ in both cases. Use Pinsker's inequality to show that

$$\|P - Q\|_1 \leq \frac{|\mu_1 - \mu_2|}{\sigma}.$$ 

**Exercise 1.4.** Compute $D(P \parallel Q)$ for the $n$-dimension Gaussian distributions $P = N(\mu_1, \Sigma_1)$ and $Q = N(\mu_2, \Sigma_2)$.

1.3 Maximum Entropy

We will now see that the multivariate Gaussian distribution maximizes differential entropy across all distributions with the same covariance.

**Theorem 1.5.** Let $X$ be a continuous random variable taking values in $\mathbb{R}^n$ with mean $\mathbb{E}[X] = 0$ and covariance matrix $\mathbb{E}[XX^T] = \Sigma$. Then,

$$h(X) \leq \frac{n}{2} \log(2\pi e) + \log(|\det(\Sigma)|),$$

with equality iff $X \sim N(0, \Sigma)$. 

3
Proof: Let $p$ be the density of $X$, and $q$ be the density of a gaussian random variable $N(0, \Sigma)$. Then,

$$
0 \leq D(p||q) = \int p(x) \log \left( \frac{p(x)}{q(x)} \right) dx
$$

$$
= \int p(x) \log p(x) dx - \int p(x) \log q(x) dx
$$

$$
= -h(p) - \int p(x) \log q(x) dx
$$

$$
= -h(p) - \int q(x) \log q(x) dx
$$

$$
= -h(p) + h(q),
$$

where the substitution $\int p(x) \log q(x) dx = \int q(x) \log q(x) dx$ follows from the definition of the density function $q$ (for a Gaussian random variable) and the fact the both $p$ and $q$ are densities for different random variables admitting the same first and second moments (Use these observations to verify that $\int p(x) \log q(x) dx = \int q(x) \log q(x) dx$). By rearranging terms, we arrive at the stated inequality. □