

Lecture 6: January 23, 2025

Lecturer: Madhur Tulsiani

1 KL-divergence for continuous random variables

We define KL-divergence for two distributions analogously, when both distributions have associated density functions.

Definition 1.1. *If P and Q are two distributions with densities p and q , then their KL-divergence is defined by the integral*

$$D(P\|Q) := \int p(x) \cdot \log \left(\frac{p(x)}{q(x)} \right) dx.$$

Unlike the concept of differential entropy, that of KL-divergence is a direct generalization of KL-divergence for distributions on finite universes. A measure-theoretic definition of KL-divergence was developed in the works of Kolmogorov and Pinsker. A detailed treatment can be found in Chapter 7 of the book by Gray [Gra11] (Chapter 5 of the older edition linked from the author's webpage).

In general, consider any two probability measures P, Q on a space Ω with underlying σ -algebra $\mathcal{F} \subseteq 2^\Omega$ (defining the notion of "valid events" which one can talk about). A random variable X taking values in a finite set $[n]$ is defined to be a *measurable function* $X : \Omega \rightarrow [n]$ i.e., we require $X^{-1}(S)$ to be a valid event in \mathcal{F} , for all subsets $S \subseteq [n]$. Then, the KL-divergence of P and Q is defined to be

$$D(P\|Q) = \sup_{X,n} D(P(X)\|Q(X)),$$

for X and n as above. When P and Q have densities p and q , this definition can be shown to converge to the one defined above.

Note that the measure-theoretic definition reduces the infinite case to the (supremum over) finite cases.

Since mutual information of two random variables X, Y can be defined in terms of the KL-divergence as (see Homework 1)

$$I(X; Y) = D(P(X, Y) \| P(X)P(Y)),$$

this also gives a measure-theoretic definition for mutual information.

Also, since $D(P(X)||Q(X)) \geq 0$ for each of the finite cases, we still have $D(P||Q)$ for any two distributions over \mathbb{R}^n . Thus, any inequalities between entropies which were derived using the non-negativity of KL-divergence are still valid. These include the non-negativity of mutual information or (equivalently) the fact that conditioning reduces entropy, the sub-additivity of entropy and also Shearer's lemma. In addition, Pinsker's inequality also holds for the infinite setting, since the total variation distance can also be defined by a similar expression in terms of finite distributions.

2 Gaussian computations

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

2.1 Differential entropy

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

$$\begin{aligned} h(X) &= \int p(x) \cdot \frac{1}{\ln 2} \cdot \left(\frac{(x - \mu)^2}{2\sigma^2} + \frac{1}{2} \ln(2\pi\sigma^2) \right) dx \\ &= \frac{1}{\ln 2} \cdot \left(\frac{1}{2} + \frac{1}{2} \ln(2\pi\sigma^2) \right) \\ &= \frac{1}{2} \cdot \log(2\pi \cdot e \cdot \sigma^2). \end{aligned}$$

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 $X = (X_1, \dots, X_n)$ 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

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

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

Exercise 2.1 (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 2.2. 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

$$\begin{aligned}
 h(Y) &= \int_{\mathbb{R}^n} p_Y(y) \cdot \log \left(\frac{1}{p_Y(y)} \right) dy \\
 &= \int_{\mathbb{R}^n} p_X(y - c) \cdot \log \left(\frac{1}{p_X(y - c)} \right) dy \\
 &= \int_{\mathbb{R}^n} p_X(x) \cdot \log \left(\frac{1}{p_X(x)} \right) dx && \text{(substituting } x = y - c) \\
 &= h(X)
 \end{aligned}$$

Similarly, for $Y = AX$, we have

$$\begin{aligned}
 h(Y) &= \int_{\mathbb{R}^n} p_Y(y) \cdot \log \left(\frac{1}{p_Y(y)} \right) dy \\
 &= \int_{\mathbb{R}^n} \frac{p_X(A^{-1}y)}{|A|} \cdot \log \left(\frac{|A|}{p_X(A^{-1}y)} \right) dy \\
 &= \int_{\mathbb{R}^n} \frac{p_X(x)}{|A|} \cdot \log \left(\frac{|A|}{p_X(x)} \right) |A| dx && \text{(substituting } x = A^{-1}y) \\
 &= h(X) + \log(|A|).
 \end{aligned}$$

■

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

$$h(Y) = h(X) + \log(|\Sigma^{1/2}|) = \frac{n}{2} \cdot \log(2\pi \cdot e) + \frac{1}{2} \cdot \log |\Sigma|.$$

2.2 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

$$\begin{aligned}
 D(P \parallel Q) &= \int_{\mathbb{R}} p(x) \cdot \log \left(\frac{p(x)}{q(x)} \right) dx \\
 &= \mathbb{E}_{x \sim P} \left[\log \left(\frac{p(x)}{q(x)} \right) \right] \\
 &= \mathbb{E}_{x \sim P} \left[\frac{1}{\ln 2} \cdot \ln \left(\frac{\exp(-(x - \mu_1)^2 / 2\sigma_1^2)}{\sqrt{2\pi}\sigma_1} \cdot \frac{\sqrt{2\pi}\sigma_2}{\exp(-(x - \mu_2)^2 / 2\sigma_2^2)} \right) \right] \\
 &= \frac{1}{\ln 2} \cdot \mathbb{E}_{x \sim P} \left[\frac{(x - \mu_2)^2}{2\sigma_2^2} - \frac{(x - \mu_1)^2}{2\sigma_1^2} + \ln \left(\frac{\sigma_2}{\sigma_1} \right) \right] \\
 &= \frac{1}{\ln 2} \cdot \left(\frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} + \ln \left(\frac{\sigma_2}{\sigma_1} \right) \right) \\
 &= \frac{1}{\ln 2} \cdot \left(\frac{\sigma_1^2 - \sigma_2^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} + \ln \left(\frac{\sigma_2}{\sigma_1} \right) \right).
 \end{aligned}$$

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 2.3. Let P and Q be Gaussian distributions with means μ_1 and μ_2 respectively, and variance σ^2 in both cases. Use Pinsker's inequality to show that

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

Exercise 2.4. Compute $D(P \parallel Q)$ for the n -dimension Gaussian distributions $P = N(\mu_1, \Sigma_1)$ and $Q = N(\mu_2, \Sigma_2)$.

3 Gaussian computations

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

3.1 Differential entropy

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

$$\begin{aligned} h(X) &= \int p(x) \cdot \frac{1}{\ln 2} \cdot \left(\frac{(x - \mu)^2}{2\sigma^2} + \frac{1}{2} \ln(2\pi\sigma^2) \right) dx \\ &= \frac{1}{\ln 2} \cdot \left(\frac{1}{2} + \frac{1}{2} \ln(2\pi\sigma^2) \right) \\ &= \frac{1}{2} \cdot \log(2\pi \cdot e \cdot \sigma^2). \end{aligned}$$

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 $X = (X_1, \dots, X_n)$ 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

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

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

Exercise 3.1 (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 3.2. *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

$$\begin{aligned}
 h(Y) &= \int_{\mathbb{R}^n} p_Y(y) \cdot \log\left(\frac{1}{p_Y(y)}\right) dy \\
 &= \int_{\mathbb{R}^n} p_X(y - c) \cdot \log\left(\frac{1}{p_X(y - c)}\right) dy \\
 &= \int_{\mathbb{R}^n} p_X(x) \cdot \log\left(\frac{1}{p_X(x)}\right) dx && \text{(substituting } x = y - c\text{)} \\
 &= h(X)
 \end{aligned}$$

Similarly, for $Y = AX$, we have

$$\begin{aligned}
 h(Y) &= \int_{\mathbb{R}^n} p_Y(y) \cdot \log\left(\frac{1}{p_Y(y)}\right) dy \\
 &= \int_{\mathbb{R}^n} \frac{p_X(A^{-1}y)}{|A|} \cdot \log\left(\frac{|A|}{p_X(A^{-1}y)}\right) dy \\
 &= \int_{\mathbb{R}^n} \frac{p_X(x)}{|A|} \cdot \log\left(\frac{|A|}{p_X(x)}\right) |A| dx && \text{(substituting } x = A^{-1}y\text{)} \\
 &= h(X) + \log(|A|).
 \end{aligned}$$

■

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

$$h(Y) = h(X) + \log(|\Sigma^{1/2}|) = \frac{n}{2} \cdot \log(2\pi \cdot e) + \frac{1}{2} \cdot \log |\Sigma|.$$

3.2 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

$$\begin{aligned}
 D(P \parallel Q) &= \int_{\mathbb{R}} p(x) \cdot \log \left(\frac{p(x)}{q(x)} \right) dx \\
 &= \mathbb{E}_{x \sim P} \left[\log \left(\frac{p(x)}{q(x)} \right) \right] \\
 &= \mathbb{E}_{x \sim P} \left[\frac{1}{\ln 2} \cdot \ln \left(\frac{\exp(-(x - \mu_1)^2 / 2\sigma_1^2)}{\sqrt{2\pi}\sigma_1} \cdot \frac{\sqrt{2\pi}\sigma_2}{\exp(-(x - \mu_2)^2 / 2\sigma_2^2)} \right) \right] \\
 &= \frac{1}{\ln 2} \cdot \mathbb{E}_{x \sim P} \left[\frac{(x - \mu_2)^2}{2\sigma_2^2} - \frac{(x - \mu_1)^2}{2\sigma_1^2} + \ln \left(\frac{\sigma_2}{\sigma_1} \right) \right] \\
 &= \frac{1}{\ln 2} \cdot \left(\frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} + \ln \left(\frac{\sigma_2}{\sigma_1} \right) \right) \\
 &= \frac{1}{\ln 2} \cdot \left(\frac{\sigma_1^2 - \sigma_2^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} + \ln \left(\frac{\sigma_2}{\sigma_1} \right) \right).
 \end{aligned}$$

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 3.3. Let P and Q be Gaussian distributions with means μ_1 and μ_2 respectively, and variance σ^2 in both cases. Use Pinsker's inequality to show that

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

Exercise 3.4. Compute $D(P \parallel Q)$ for the n -dimension Gaussian distributions $P = N(\mu_1, \Sigma_1)$ and $Q = N(\mu_2, \Sigma_2)$.

References

[Gra11] Robert M Gray, *Entropy and information theory*, Springer Science & Business Media, 2011. 1