

## Lecture 7: January 28, 2025

Lecturer: Madhur Tulsiani

## 1 Some more consequences of Gaussian KL-divergence

In the previous lecture, we defined and computed differential entropy for continuous random variables. The differential analogue of KL-divergence, which can be used to rigorously extend much of the theory from finitely supported random variables, leads to some useful inequalities.

### 1.1 Maximum Entropy

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

**Theorem 1.1.** *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  $X \sim N(0, \Sigma)$ .

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

$$\begin{aligned} 0 \leq D(p||q) &= \int p(x) \log \left( \frac{p(x)}{q(x)} \right) dx \\ &= \int p(x) \log p(x) dx - \int p(x) \log q(x) dx \\ &= -h(p) - \int p(x) \log q(x) dx \\ &= -h(p) - \int q(x) \log q(x) dx \\ &= -h(p) + h(q), \end{aligned}$$

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. ■

Note that while in general it does not make sense to compare differential entropy, in the above theorem, we fixed the "scale" of the random variable by fixing the covariance matrix. The above theorem also has a useful consequence, which says that for *any* continuous real-valued random variable, the differential entropy is bounded by a function of the variance.

**Corollary 1.2.** *Let  $X$  be any continuous random variable taking values in  $\mathbb{R}$ . Then,*

$$h(X) \leq \frac{1}{2} \cdot \log(2\pi e \cdot \text{Var}[X]).$$

**Proof:** Let  $Y$  be a Gaussian random variable with  $\text{Var}[Y] = \text{Var}[X]$ . Then, we have  $h(X) \leq h(Y) = \frac{1}{2} \cdot \log(2\pi e \cdot \text{Var}[Y]) = \frac{1}{2} \cdot \log(2\pi e \cdot \text{Var}[X])$ . ■

## 1.2 A continuous analogue of Fano's inequality

The above relation between the entropy and variance of a continuous random variable, can also be used to prove an analogue of Fano's inequality for continuous random variables, which says that the error of an estimator can be lower bounded in terms of the conditional entropy. Formally, for a Markov chain  $X \rightarrow Y \rightarrow \hat{X}$ , we are interested in understanding the accuracy of the estimator  $\hat{X}$  for predicting  $X$ . However, since the variables now take values in an infinite universe (say  $\mathbb{R}$ ), we will measure the error not by how often  $X$  and  $\hat{X}$  are on average, but rather how close they are on average.

**Claim 1.3.** *Let  $X \rightarrow Y \rightarrow \hat{X}$  real-valued continuous random variables. Then,*

$$\mathbb{E}[(X - \hat{X})^2] \geq \frac{1}{2\pi e} \cdot 2^{h(X|Y)}.$$

**Proof:** Note that for any fixed  $a \in \mathbb{R}$  and any random variable  $Z$ , we can say that  $\mathbb{E}[Z - a] \geq \text{Var}[Z]$ . This implies that for any fixed value  $y$  of  $Y$ , we get that

$$\mathbb{E}[(X - \hat{X})^2 | Y = y] \geq \text{Var}[X | Y = y] \geq \frac{1}{2\pi e} \cdot 2^{h(X|Y=y)},$$

where the second inequality used [Corollary 1.2](#). Taking an expectation over the values  $y$  (according to the distribution of the random variable  $Y$ ) gives

$$\mathbb{E}_y \mathbb{E}[(X - \hat{X})^2 | Y = y] \geq \mathbb{E}_y \frac{1}{2\pi e} \cdot 2^{h(X|Y=y)} \geq \frac{1}{2\pi e} \cdot 2^{\mathbb{E}_y h(X|Y=y)} = \frac{1}{2\pi e} \cdot 2^{h(X|Y)},$$

where we used Jensen's inequality to deduce the second inequality. ■

## 2 The Method of Types

We now return to finitely supported random variables. We will take  $\mathcal{X}$  to be a finite universe  $|\mathcal{X}| = r$ , and use  $\bar{x} = (x_1, x_2, \dots, x_n)$  to denote a sequence of  $n$  elements from  $U$ .

**Definition 2.1.** The type  $P_{\bar{x}}$  of  $\bar{x}$ , also called the empirical distribution of  $\bar{x}$ , is a distribution  $\hat{P}$  on  $\mathcal{X}$ , defined as

$$\hat{P}(a) := \frac{|\{i : x_i = a\}|}{n} \quad \forall a \in \mathcal{X}.$$

We use  $\mathcal{T}_n$  to denote the set of all types coming from sequences of length  $n$ . We also use  $\mathcal{C}_P$  to denote the set of all sequences with the type  $P$ .  $\mathcal{C}_P$  is called the type class of  $P$ .

$$\mathcal{C}_P := \{\bar{x} \in \mathcal{X}^n \mid P_{\bar{x}} = P\}.$$

**Exercise 2.2.** Check that  $|\mathcal{T}_n| = \binom{n+r-1}{r-1} \leq (n+1)^r$ .

Next, we bound the size of a given type class in terms of the entropy of that type.

**Proposition 2.3.** For any type  $P \in \mathcal{T}_n$ , we have

$$\frac{2^{n \cdot H(P)}}{(n+1)^r} \leq |\mathcal{C}_P| \leq 2^{n \cdot H(P)}.$$

**Proof:** For each  $a_i \in U$ , let  $P(a_i) = k_i/n$ . Then  $|\mathcal{C}_P| = n!/(k_1!k_2! \dots k_r!)$ . We prove the lower bound by considering

$$\begin{aligned} n^n &= (k_1 + k_2 + \dots + k_r)^n = \sum_{j_1 + \dots + j_r = n} \frac{n!}{j_1! \dots j_r!} (k_1^{j_1} \dots k_r^{j_r}) \\ &\leq \binom{n+r-1}{r-1} \cdot \max_{j_1 + \dots + j_r = n} \frac{n!}{j_1! \dots j_r!} \cdot (k_1^{j_1} \dots k_r^{j_r}), \end{aligned}$$

where each tuple  $(j_1, \dots, j_r)$  corresponds to a distinct type. We leave it as an exercise to check that the maximum term in the expression above is when  $(j_1, \dots, j_r) = (k_1, \dots, k_r)$ .

**Exercise 2.4.** Show that

$$\frac{n!}{j_1! \dots j_r!} \cdot (k_1^{j_1} \dots k_r^{j_r}) \leq \frac{n!}{k_1! \dots k_r!} \cdot (k_1^{k_1} \dots k_r^{k_r})$$

for all  $(j_1, \dots, j_r)$  such that  $j_1 + \dots + j_r = n$ . (Hint: if  $j_s > k_s$  for some  $s$ , then  $j_t < k_t$  for some  $t$ .)

Using the above, we can now prove the lower bound.

$$n^n \leq \binom{n+r-1}{r-1} \cdot \frac{n!}{k_1! \dots k_r!} \cdot (k_1^{k_1} \dots k_r^{k_r}) \leq (n+1)^r \cdot |\mathcal{C}_P| \cdot (k_1^{k_1} \dots k_r^{k_r}).$$

We get

$$\begin{aligned}
|\mathcal{C}_P| &\geq \frac{1}{(n+1)^r} \cdot \frac{n^{k_1+k_2+\dots+k_r}}{k_1^{k_1} \dots k_r^{k_r}} \\
&= \frac{1}{(n+1)^r} \cdot \prod_{i=1}^r \left(\frac{n}{k_i}\right)^{k_i} \\
&= \frac{1}{(n+1)^r} \cdot \prod_{i=1}^r 2^{k_i \cdot \log(n/k_i)} = \frac{1}{(n+1)^r} \cdot 2^{n \cdot H(P)}.
\end{aligned}$$

The proof of the upper bound is similar and left as an exercise. ■

Next, we need the observation that the probability of a sequence according to a product distribution only depends on its type.

**Proposition 2.5.** *Let  $Q$  be any distribution on  $U$  and let  $Q^n$  the product distribution on  $\mathcal{X}^n$ . Let  $\bar{x}, \bar{y} \in \mathcal{X}^n$  be such that  $P_{\bar{x}} = P_{\bar{y}}$ . Then,  $Q^n(\bar{x}) = Q^n(\bar{y})$ .*

**Proof:** Let  $P = P_{\bar{x}} = P_{\bar{y}}$ . Then we have:

$$Q^n(\bar{x}) = \prod_{a \in \mathcal{X}} (Q(a))^{\#\{i: x_i=a\}} = \prod_{a \in \mathcal{X}} (Q(a))^{n \cdot P(a)} = Q^n(\bar{y}).$$

■

Now we give bounds on the probability of a certain type occurring, in terms of the KL divergence between the true distribution and the empirical distribution.

**Theorem 2.6.** *For any product distribution  $Q^n$  and type  $P$  on  $\mathcal{X}^n$ , we have*

$$\frac{2^{-n \cdot D(P||Q)}}{(n+1)^r} \leq \mathbb{P}_{\bar{x} \sim Q^n} [P_{\bar{x}} = P] \leq 2^{-n \cdot D(P||Q)}.$$

**Proof:** Let  $\bar{x}$  be of type  $P_{\bar{x}} = P$ . For the lower bound, we note that

$$\frac{Q^n(\bar{x})}{P^n(\bar{x})} = \frac{\prod_{a \in \mathcal{X}} (Q(a))^{n P(a)}}{\prod_{a \in \mathcal{X}} (P(a))^{n P(a)}} = \prod_{a \in \mathcal{X}} \left(\frac{Q(a)}{P(a)}\right)^{n P(a)} = 2^{n \sum_{a \in \mathcal{X}} P(a) \log\left(\frac{Q(a)}{P(a)}\right)} = 2^{-n \cdot D(P||Q)}$$

We also know from the previous proposition that for any  $\bar{x} \in \mathcal{C}_P$ , we have

$$P^n(\bar{x}) = \prod_{a \in U} (P(a))^{n \cdot P(a)} = 2^{-n \cdot H(P)}.$$

Finally, using Proposition 2.3, we get

$$\begin{aligned}
\mathbb{P}_{\bar{x} \sim Q^n} [P_{\bar{x}} = P] &= \sum_{\bar{x} \in \mathcal{C}_P} Q^n(\bar{x}) = \sum_{\bar{x} \in \mathcal{C}_P} 2^{-n \cdot H(P)} \cdot 2^{-n \cdot D(P \| Q)} \\
&= |\mathcal{C}_P| \cdot 2^{-n \cdot H(P)} \cdot 2^{-n \cdot D(P \| Q)} \\
&\geq \frac{2^{n \cdot H(P)}}{(n+1)^r} \cdot 2^{-n \cdot H(P)} \cdot 2^{-n \cdot D(P \| Q)} \\
&= \frac{2^{-n \cdot D(P \| Q)}}{(n+1)^r}
\end{aligned}$$

The proof of the upper bound is left as an exercise. Note that It may be that  $\text{Supp}(Q) \subsetneq \text{Supp}(P)$  i.e.,  $\exists a \in \mathcal{X} : Q(a) = 0, P(a) \neq 0$ . Then the  $\log(1/Q(a))$  term makes  $D(P \| Q)$  undefined, so thinking of  $D(P \| Q)$  as  $+\infty$ , we get  $2^{-nD(P \| Q)} = \text{Prob}_{Q^n}(T_P^n) = 0$ . ■

### 3 Chernoff bounds

The above counting can be used to prove the Chernoff bound. Let  $\mathcal{X} = \{0, 1\}$ , and let  $\bar{x} = (x_1, \dots, x_n)$  be a sequence drawn from  $\mathcal{X}^n$  according to  $Q^n$ , where

$$Q = \begin{cases} 0 & : \text{ with probability } 1/2 \\ 1 & : \text{ with probability } 1/2. \end{cases}$$

We expect there to be around  $n/2$  occurrences of 1 in  $\bar{X}$ ; that is,  $\mathbb{E}[\sum_{i=1}^n x_i] = n/2$ . It is natural to ask how much the empirical distribution is likely to deviate from  $n/2$ . If we set

$$P = \begin{cases} 0 & : \text{ with probability } 1/2 - \varepsilon \\ 1 & : \text{ with probability } 1/2 + \varepsilon, \end{cases}$$

then we have

$$\mathbb{P}_{Q^n} \left[ X_1 + \dots + X_n = \frac{n}{2} + \varepsilon n \right] = \mathbb{P}_{\bar{x} \sim Q^n} [P_{\bar{x}} = P] \leq 2^{-n \cdot D(P \| Q)} = 2^{-c \cdot n \cdot \varepsilon^2},$$

by Theorem 2.6, for a constant  $c$ . This is sort of like Chernoff bounds, but we may want to know how likely we are to see *any* sufficiently large deviation, and not just the deviation exactly equal to  $\varepsilon n$ .

**Theorem 3.1** (Chernoff bound). *For  $\bar{X} = (X_1, \dots, X_n) \sim_{Q^n} U^n$  with  $Q$  the uniform distribution on  $\mathcal{X} = \{0, 1\}$ , we have*

$$\mathbb{P}_{Q^n} \left[ \sum_{i=1}^n X_i \geq \frac{n}{2} + \varepsilon n \right] \leq (n+1) \cdot 2^{-c \cdot n \cdot \varepsilon^2}.$$

**Proof:** Let  $\mathcal{X} = \{0, 1\}$  and note that each type class corresponds to a unique value of  $x_1 + \cdots + x_n$ . From the above bound, we have that for any  $\eta > 0$ ,

$$\mathbb{P}_{Q^n} \left[ X_1 + \cdots + X_n = \frac{n}{2} + \eta n \right] \leq 2^{-c \cdot n \cdot \eta^2}.$$

Going over all types for all  $\eta \geq \varepsilon$ , and noting that the number of types is at most  $n + 1$ , we get

$$\mathbb{P}_{Q^n} \left[ \sum_{i=1}^n X_i \geq \frac{n}{2} + \varepsilon n \right] \leq (n + 1) \cdot 2^{-c \cdot n \cdot \varepsilon^2},$$

as claimed. ■

The above idea can be generalized for product distributions over arbitrary (finite) universes to prove a general large deviation result known as Sanov's theorem.