Multi-armed Bandits

Kjell Raaijmakers (1244095), Jeroen van Riel (1236068)

June 1, 2022

Question 1

Our intuitive description of Kullback-Leibler divergence is largely based on Chapter 2, 4 and 6 in [1], which provides a very foundational treatment of probability and entropy while showing their importance in physics. The main point that we are trying to convey is that an intuitive interpretation of Kullback-Leibler divergence should be based on the axioms that give rise to its definition.

Probabilities as beliefs. Probability theory can be regarded an extension of classical logic (which deals with statements that are either true or false) to support reasoning with degrees of belief [5]. Richard Cox [2] developed a calculus for rational reasoning to represent and manipulate these degrees of belief as numbers. Based on assumptions of transitivity and consistency, he derives the foundational sum and product rules in probability theory, see [7] or pages 7-26 in [1].

Missing information. Suppose we need to determine the state i of some system. Based on incomplete information I, we have assigned probabilities $p(i|I) = p_i$ to each of the possible states. In order to exactly determine the state of the system, we need more information. Claude Shannon [6] came up with a definition of this amount of *missing information* S, based on the following axioms (pages 392-393). (i) S is a function of the probabilities p_i. (ii) If all states are equally likely, so $p_i = 1/n$ for n possible states, then $S = F(n)$ is an increasing function of n. The idea is that we need more information when the number of states we can choose from is large. (iii) The third axiom is less intuitive, but essentially requires that the particular sequence in which we obtain partial missing information does not make a difference: it should not matter if we first obtain information "$i < n/2$" and then learn the actual state i or that we learn the actual state in one step. The *entropy* of the probability distribution $p_1, \ldots, p_n$ is defined as

$$S[p] = -\sum_{i=1}^{n} p_i \log p_i$$

and is the unique S satisfying these axioms, up to the base of the logarithm. Shannon was interested in the fundamental limits of data compression for communication. Informally, Shannon's *source coding theorem* states that a stream of symbols (e.g. numbers or characters), considered as discrete random variables with distribution p, need to be transmitted using at least $S[p]$ bits on average in order to avoid loss of information.

Updating beliefs. Inspired by Shannon's information theory, Edwin Jaynes proposed the method of maximum-entropy inference: when faced with new data, we should prefer posterior distributions that make as few additional assumptions as possible, i.e., have maximum entropy [3, 4]. In the particular case of discrete random variables, the *relative entropy* defined as

$$K[p, q] = \sum_i p_i \log \frac{p_i}{q_i},$$

can be used to measure the change in entropy due to learning that the actual distribution is p, while we believed it to be q. In the continuous case, its equivalent is known as Kullback-Leibler divergence, because it was later found that it may be defined as a divergence between distributions considered as vectors in some vector space. Informally, both quantities may be regarded as a measure of how far apart two distributions are.

Returning to the example of transmitting symbols, relative entropy can be interpreted as the average additional number of bits that we require to encode symbols that are actually distributed according to q, while using a code that is optimal for symbols distributed according to p.

The entropy maximization principle also provides a very natural interpretation of some common probability distributions. When regarding the uniform distribution as the distribution that encodes the least prior bias, requiring a fixed mean μ and variance σ^2, we find that the normal distribution is the unique distribution that minimizes the relative entropy with respect to the uniform distribution. In other words, the normal distribution introduces no more information beyond the uniform distribution than specifying the mean and variance.

Question 2

Let $\Delta_\mu(i) := \max_{k\in[K]} \mu_k - \mu_i$ denote the suboptimality parameter of arm i under reward distribution μ. Expressing the pseudo-regret as in in Exercise 8 and using Markov's inequality, we obtain

$$\begin{aligned}
\overline{R}_{B_\mu,\pi}(t) &= \sum_{i=1}^{K} \Delta_\mu(i)\ \mathbb{E}_{\mathbb{P}_\pi}[T_i(t)] \\
&= \sum_{i=2}^{K} m\ \mathbb{E}_{\mathbb{P}_\pi}[T_i(t)] \\
&= m\ \mathbb{E}_{\mathbb{P}_\pi}\left[\sum_{i=2}^{K} T_i(t)\right] \\
&= m\ \mathbb{E}_{\mathbb{P}_\pi}\left[t - T_1(t)\right] \\
&\geq \frac{mt}{2}\ \mathbb{P}_\pi(t - T_1(t) \geq t/2) \\
&= \frac{mt}{2}\ \mathbb{P}_\pi(T_1(t) \leq t/2).
\end{aligned}$$

Similarly for $\Delta_\nu(i) := \max_{k\in[K]} \nu_k - \nu_i$, we derive

$$\begin{aligned}
\overline{R}_{B_\nu,\pi}(t) &= \sum_{i=1}^{K} \Delta_\nu(i)\ \mathbb{E}_{\mathbb{Q}_\pi}[T_i(t)] \\
&= m\ \mathbb{E}_{\mathbb{Q}_\pi}[T_1(t)] + 2m \sum_{i\notin\{1,l^*\}} \mathbb{E}_{\mathbb{Q}_\pi}[T_i(t)] \\
&\geq m\ \mathbb{E}_{\mathbb{Q}_\pi}[T_1(t)] \\
&\geq \frac{mt}{2}\ \mathbb{Q}_\pi(T_1(t) > t/2).
\end{aligned}$$

Combining both inequalities using Bretagnolle-Huber's inequality yields

$$\overline{R}_{B_\mu,\pi}(t) + \overline{R}_{B_\nu,\pi}(t) \geq \frac{mt}{2}\Big(\mathbb{P}_\pi(T_1(t) \leq t/2) + \mathbb{Q}_\pi(T_1(t) > t/2)\Big) \geq \frac{mt}{4} e^{-\mathrm{KL}(\mathbb{P}_\pi,\mathbb{Q}_\pi)}.$$

Question 3

Now we want to show that for any policy π for which $\arg\min_{l>1} \mathbb{E}_{\mathbb{P}_\pi}[T_l(t)] = l^*$ is constant, there exists a vector ξ such that we can find a lower bound for $\overline{R}_{B_\xi,\pi}(t)$. Because we may choose any vector

ξ, we may always pick the larger one of both candidates

$$\begin{aligned}\overline{R}_{B_\xi,\pi}(t) &\geq \max\left(\overline{R}_{B_\mu,\pi}(t), \overline{R}_{B_\nu,\pi}(t)\right)\\ &\geq \frac{1}{2}\left(\overline{R}_{B_\mu,\pi}(t) + \overline{R}_{B_\nu,\pi}(t)\right)\\ &\geq \frac{mt}{8}e^{-\mathrm{KL}(\mathbb{P}_\pi,\mathbb{Q}_\pi)}. \end{aligned} \tag{1}$$

Now to get an further lower bound for this, we need to find what $\mathrm{KL}(\mathbb{P}_\pi, \mathbb{Q}_\pi)$ looks like. Let us first compute

$$\begin{aligned}\mathrm{KL}(P_k, Q_k) &= \int_{-\infty}^{\infty} p_k(x)\ln\left(\frac{p_k(x)}{q_k(x)}\right)dx\\ &= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\mu_k)^2}\ln\left(\frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\mu_k)^2}}{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\nu_k)^2}}\right)dx\\ &= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\mu_k)^2}\frac{1}{2}((x-\nu_k)^2-(x-\mu_k)^2)dx\\ &= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\mu_k)^2}(x(\mu_k-\nu_k)+\frac{1}{2}(\nu_k^2-\mu_k^2))dx\\ &= (\mu_k-\nu_k)\int_{-\infty}^{\infty}\frac{x}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\mu_k)^2}dx + \frac{1}{2}(\nu_k^2-\mu_k^2)\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\mu_k)^2}dx\\ &= \mu_k(\mu_k-\nu_k)+\frac{1}{2}(\nu_k^2-\mu_k^2)\\ &= \frac{1}{2}(\nu_k-\mu_k)^2. \end{aligned}$$

Now for arm l^*, we must have $\mathbb{E}_{\mathbb{P}_\pi}[T_{l^*}(t)] \leq \frac{t}{K-1}$, because otherwise the other $K-1$ arms satisfy $\mathbb{E}_{\mathbb{P}_\pi}[T_i(t)] > \frac{t}{K-1}$, so that $\sum_{i\neq l^*}\mathbb{E}_{\mathbb{P}_\pi}[T_i(t)] > t$, which is not possible. Note that the only difference between μ_k and ν_k occurs at $k = l^*$, so we can bound the Kullback-Leibler divergence like

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_\pi, \mathbb{Q}_\pi) &= \sum_{k=1}^{K}\mathbb{E}_{\mathbb{P}_\pi}[T_k(t)]\ \mathrm{KL}(P_k, Q_k)\\ &= \sum_{k=1}^{K}\mathbb{E}_{\mathbb{P}_\pi}[T_k(t)]\ \frac{1}{2}(\nu_k-\mu_k)^2\\ &= \mathbb{E}_{\mathbb{P}_\pi}[T_{l^*}(t)]\ 2m^2\\ &\leq \frac{2m^2t}{K-1}. \end{aligned}$$

Since $e^{-x} \geq e^{-y}$ for $x \leq y$, this yield the following lower bound

$$(1) \geq \frac{mt}{8}\exp\left(-\frac{2m^2t}{K-1}\right). \tag{2}$$

This lower bound holds for all $m > 0$, so we may optimize over m:

$$\begin{aligned}
\frac{d}{dm}\frac{mt}{8}\exp\left(-\frac{2m^2t}{K-1}\right) &= 0 \\
\frac{t}{8}\exp\left(-\frac{2m^2t}{K-1}\right) - \frac{m^2t^2}{2(K-1)}\exp\left(-\frac{2m^2t}{K-1}\right) &= 0 \\
\frac{t}{8} &= \frac{m^2t^2}{2(K-1)} \\
\frac{K-1}{4t} &= m^2 \\
m &= \frac{1}{2}\sqrt{\frac{(K-1)}{t}}
\end{aligned}$$

Note that we only have to take the positive value for m. Substituting $m = \frac{1}{2}\sqrt{\frac{(K-1)}{t}}$ into equation (2) gives us

$$\begin{aligned}
\frac{mt}{8}e^{-\frac{2m^2t}{K-1}} &\geq \frac{1}{2}\sqrt{\frac{(K-1)}{t}}\frac{t}{8}e^{-2\frac{K-1}{4t}\frac{t}{K-1}} \\
&= \frac{1}{16}\sqrt{(K-1)t}e^{-\frac{1}{2}} \\
&= \frac{1}{16\sqrt{e}}\sqrt{(K-1)t}\,.
\end{aligned}$$

References

[1] Ariel Caticha et al. "Entropic inference and the foundations of physics". In: *Brazilian Chapter of the International Society for Bayesian Analysis-ISBrA, Sao Paulo, Brazil* (2012).

[2] R. T. Cox. "Probability, Frequency and Reasonable Expectation". In: *American Journal of Physics* 14.2 (1946), pp. 1–13. DOI: http://doi.org/10.1119/1.1990764.

[3] E. T. Jaynes. "Information Theory and Statistical Mechanics". In: *Phys. Rev.* 106 (4 May 1957), pp. 620–630. DOI: 10.1103/PhysRev.106.620. URL: https://link.aps.org/doi/10.1103/PhysRev.106.620.

[4] E. T. Jaynes. "Information Theory and Statistical Mechanics. II". In: *Phys. Rev.* 108 (2 Oct. 1957), pp. 171–190. DOI: 10.1103/PhysRev.108.171. URL: https://link.aps.org/doi/10.1103/PhysRev.108.171.

[5] E. T. Jaynes. *Probability Theory: The Logic of Science*. Ed. by G. Larry Bretthorst. Cambridge University Press, 2003. DOI: 10.1017/CBO9780511790423.

[6] C. E. Shannon. "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[7] Kevin S. Van Horn. "Constructing a logic of plausible inference: a guide to Cox's theorem". In: *International Journal of Approximate Reasoning* 34.1 (2003), pp. 3–24. ISSN: 0888-613X. DOI: https://doi.org/10.1016/S0888-613X(03)00051-3. URL: https://www.sciencedirect.com/science/article/pii/S0888613X03000513.

Distribution of work

- Question 1 is due to Jeroen and Kjell provided some useful feedback.
- We both independently managed to solve Question 2 and the answer above is a result of combining our findings.
- Kjell derived the argument for Question 3 and Jeroen fixed some details.