Theoretical assignment 4; 15 points total + 3 points extra

Theoretical Deep Learning course, MIPT

Here by capital letters we denote random variables, and by small letters — their values. That is, $x \in \operatorname{supp} X$ is a value of random variable X with support $\operatorname{supp} X$.

Problem 1

5 points total.

1. **1 point.** Prove that Kullback-Leibler divergence is non-negative.
2. **0.5 points.** Prove that mutual information is non-negative.
3. **0.5 points.** Prove that conditioning reduces entropy: $H(X|Y) \le H(X)$ (this holds for both entropy and differential entropy).
4. **2 points.** Is it true, that conditioning reduces mutual information: $I(X;Y|Z) \le I(X;Y)$? Or maybe conditioning increases it $I(X;Y|Z) \ge I(X;Y)$?
5. **1 point.** Prove that $I(X; f(Y)) \le I(X;Y)$ for any deterministic function f.

Problem 2

1 point.
Consider a Markov chain $X_1 \to X_2 \to \ldots \to X_n$. Prove that a sequence of random variables $X_n, \ldots, X_1$ is also a Markov chain.

Problem 3

1 point.
Let $\mathcal{Q}$ be a set of all factorized density functions, i.e.

$$\mathcal{Q} = \big\{ q(\boldsymbol{x}) \mid q(\boldsymbol{x}) = \prod_{i=1}^{n} q(x_i) \big\}$$

Consider some density $p(\boldsymbol{x})$ (not necessarily from $\mathcal{Q}$). Prove that

$$q_*(\boldsymbol{x}) = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}[p(\boldsymbol{x}) \| q(\boldsymbol{x})] \Longleftrightarrow q_*(\boldsymbol{x}) = \prod_{i=1}^{n} p(x_i)$$

Hence if we want to approximate some distribution $p(\boldsymbol{x})$ with a factorized one, we should better take the product of its marginals.

Problem 4

2 points.
Consider a Markov chain $X_1 \to X_2 \to \ldots \to X_n$. Prove that

$$I(X_1; X_n) \le \min_{k<n} I(X_k, X_{k+1})$$

This means that amount of information passed along the chain can't be larger than the "bandwidth" of its tightest link.

Problem 5

6 points total + 3 points extra.

Consider a Markov chain $Y \to X \to Z$, where $Z = f(X)$ for a *deterministic* neural network f. We are going to prove that in this case $I(X;Z)$ is either infinite (if X is continuous) or constant (if X is discrete) regardless of training process. Proofs will be completely different for these two cases.

1. **Continuous case.**

   Let X be a continuous random variable.

   (a) **1 point.** Prove that if f is injective, we have:

   $$I(X; f(X)) = \infty.$$

   (b) **2 points.** Prove that the same holds for f such that a set $f^{-1}(x)$ is finite for every $x \in \operatorname{supp} X$.

   (c) **3 points extra.** Prove that the same holds if $f(X)$ has support of positive Lebesgue measure. The proof of point 1 of Theorem 2.4 in `http://people.lids.mit.edu/yp/homepage/data/itlectures_v5.pdf` might help you.

2. **Discrete case.**

   (a) **2 points.** Let X be some discrete random variable with a finite or countable support $S_X = \operatorname{supp} X$ and $f_\theta(x)$ be a neural network with any injective non-linearity σ (sigmoid, tanh, LeakyReLU, etc):

   $$f_\theta(x) = \sigma(W_k(...(W_2(W_1(x) + b_1) + b_2)...) + b_k) \qquad \theta = \{W_1, ..., W_k, b_1, ..., b_k\}$$

   Prove that the set of weights $\tilde{\Theta}$ for which $f_\theta(x)$ is not injective on S_X:

   $$\tilde{\Theta} = \{\theta \mid \exists\, x_i, x_j \in S_X, x_i \neq x_j, f_\theta(x_i) = f_\theta(x_j)\}$$

   is a measure zero set. This means that if we have a discrete distribution then our output (and all intermediate activations) are different for different inputs almost surely.

   (b) **1 point.** Let X be some discrete random variable. Prove that there is some $c \in \mathbb{R}$ such that for any function f which is injective for X (i.e. it's injective on $\operatorname{supp} X$) we have $I(X; f(X)) = c$.