# Theoretical assignment 4

Theoretical Deep Learning course, MIPT

## Problem 1 (2 points)

(0.25 pts) Prove that conditioning reduces entropy: $H(x|y) \leq H(x)$ (this holds for both entropy and differential entropy).

(0.75 pts) Is it true, that conditioning reduces mutual information: $I(x; y|z) \leq I(x; y)$?

(1 pts) Prove that $I(x; f(y)) \leq I(x; y)$ for any deterministic function $f$.

## Problem 2 (1 point)

Consider a Markov chain $x_1 \to x_2 \to ... \to x_n$. Prove that $x_n \to x_{n-1} \to ... \to x_1$ is also a Markov chain.

## Problem 3 (1 point)

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

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

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

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

This means, that 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 (1 point)

Consider a Markov chain $x_1 \to x_2 \to ... \to x_n$. Prove that

$$I(x_1; x_n) \leq \min_{k<n} I(x_k, x_{k+1})$$

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

## Problem 5

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.

### Continuous case (1 point)

Prove that for any continuous random variable $x$ and any continuous function $f$ we have:

$$I(x; f(x)) = \infty$$

**Discrete case (4 points)**

(3 points) Let $X$ be some discrete random variable with a finite or countable support $S_X = \text{supp}(X)$ and $f_\theta(x)$ be a neural network with any injective non-linearity $\sigma$ (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.

(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 $\text{supp}(X)$) we have $I(x; f(x)) = c$.