Introduction to Information Theory and Machine Learning July 13th, 2020.
(Prof. Permuter Haim, Mr. Eli Shmuel, Mr. Dor Tsur and Mr. Ben Marinberg)

**Final Exam - Moed Alef**
Total time for the exam: 3 hours!

Please copy the following sentence and sign it: " I am respecting the rules of the exam: Signature:_________ "

1) **(7 points)** Assume $Y_1 - Y_2 - \cdots - Y_m$ forms a Markov chain. Simplify $I(Y_1; Y_2, Y_3, \ldots, Y_m)$ to its simplest form.
**Solution:** $I(Y_1; Y_2, Y_3, \ldots, Y_m) = H(Y_1) - H(Y_1|Y_2, \ldots, Y_m) = H(Y_1) - H(Y_1|Y_2)$ where the last equality follows from the Markovity. Hence, $I(Y_1; Y_2, Y_3, \ldots, Y_m) = I(Y_1; Y_2)$.
2) **(7 points)** Assume $X - Y - Z$ forms a Markov chain. Show that
$$I(X;Y) \geq I(X;Y|Z).$$
When does an equality hold?
Hint: Chain rule on $I(X;Y,Z)$.
**Solution:** From the information chain rule: on the one hand $I(X;Y,Z) = I(X;Y) + I(X;Z|Y)$, while on the other hand $I(X;Y,Z) = I(X;Z) + I(X;Y|Z)$. Hence $I(X;Y) - I(X;Y|Z) = I(X;Z) - I(X;Z|Y) = I(X;Z)$, where the last inequality follows from the given Markov chain. Hence $I(X;Y) \geq I(X;Y|Z)$, and an equality holds iff $I(X;Z) = 0$, i.e. $X \perp\!\!\!\perp Z$. Another solution is by using Question 1).
3) **(7 points)** Let $f(y)$ be an arbitrary function defined for $y \geq 1$. Let $X$ be a random variable taking values in $\mathcal{X} = \{x_1, x_2, ..., x_n\}$ with probability $p_i = \Pr(X = x_i), i = 1, 2, ..., n$. Define the $f$-entropy of $X$ by
$$H_f(X) \triangleq \sum_{i=1}^{n} p_i f\left(\frac{1}{p_i}\right).$$
If $f(\cdot)$ is concave, show that the following inequality is always satisfied:
$$H_f(X) \leq f(n).$$
**Solution:** Consider $x_i = \frac{1}{p_i}$, then
$$H_f(X) = \sum_{i=1}^{n} p_i f\left(\frac{1}{p_i}\right) = E[f(X)] \overset{(a)}{\leq} f(E[X]) = f(\sum_{i=1}^{n} p_i \frac{1}{p_i}) = f(n).$$
where (a) follows from Jensen's inequality because $f(\cdot)$ is concave.
4) **(17 points)** Assume $X$ is a random variable taking values in $\mathcal{X} = \{1, 2, 3, ...\}$ with $E[X] = M$.
   a) **(10 points)** Show: $H(X) \leq M$.
   b) **(7 points)** For $M = 2$, what distribution $P_X$ achieves an equality?

   **Solution:**
   a) Consider $Q(x) = \frac{1}{2^x}$ (make sure it is a legal probability measure). Then for any distribution $P(x)$:
   $$D(P||Q) = \sum_{x=1}^{\infty} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \geq 0.$$
   Simplifying $D(P||Q) = E[X] - H(X) \geq 0$ gives that $H(X) \leq M$.
   b) As we studied for divergence, $D(P||Q) = 0$ iff $P = Q$. For the choice $P(x) = \frac{1}{2^x}$ we then have $H(X) = M = 2$.
5) **(12 points)** Consider a ternary channel with input $X_i$ and output $Y_i$, i.e. $X_i, Y_i \in \{0, 1, 2\}$. Let $\oplus$ denote addition modulo-3. The channel law is given by
$$Y_i = X_i \oplus W_i$$
where noises $\{W_i\}$ are independent of $\{X_i\}$ and are distributed i.i.d. $\sim W$, $W_i \in \{0, 1, 2\}$.

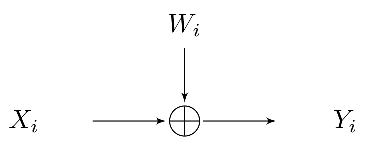



Fig. 1: An additive channel.

What is the capacity of this channel and what is the input distribution $P_X$ that achieves the capacity?

**Solution:** For any $P_X$ we have

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &\overset{(a)}{=} H(Y) - H(Y \oplus X|X) \\ &= H(Y) - H(W|X) \\ &\overset{(b)}{=} H(Y) - H(W) \\ &\overset{(c)}{\leq} \log 3 - H(W) \end{aligned} \tag{1}$$

where (a) is due to invariance of entropy to any one-to-one transformation of the random variable; (b) follows from $W \perp\!\!\!\perp X$; and (c) follows because $Y$ is ternary. (c) is achieved with equality if the distribution of $Y$ is uniform, and it can be induced when $P_X$ is distributed uniformly as well.

6) **Neural networks Highway gate (28 pt)** Fig. 2 visualizes a simple Highway gated network. The network has three linear layers, the first two is followed by ReLU activation function (marked by $\sigma$). The Highway gate $H$ and its complementary gate $\bar{H}$ are defined using a learnable parameter $h$ as follows:

$$H(x) = x \cdot h, \tag{2}$$
$$\bar{H}(x) = x \cdot (1-h). \tag{3}$$

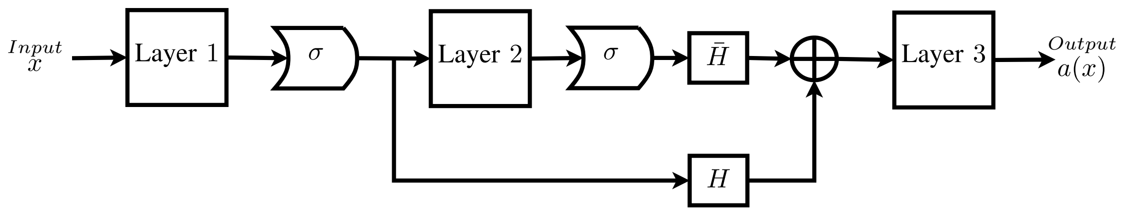



Fig. 2: A scheme of neural network with Highway gates

Initialize the network parameters as:

$$x = [0.1, 0.2, 0.6, 0.5]^T, y = 3, w^1 = \begin{bmatrix} 0.5 & 0.2 & 0.3 & -0.5 \\ 0.2 & -0.5 & 0.1 & 0.8 \\ -0.3 & 0.4 & 0.3 & -0.2 \end{bmatrix}, w^2 = \begin{bmatrix} 0.2 & 0.1 & 0.3 \\ 0.1 & -0.5 & 0.1 \\ 0 & 0.6 & -0.7 \end{bmatrix}, w^3 = [1.5, 1, 0.5], h = 0.4.$$

a) **(10 points)** Calculate the derivatives $\frac{\partial C}{\partial w^2_{3,1}}, \frac{\partial C}{\partial h}$. Consider MSE cost function.
b) **(3 points)** Explain for what purpose one need to calculate the derivative in a).
c) **(8 points)** Calculate the derivative $\frac{\partial C}{\partial w^1_{2,2}}$ for $h = 0, h = 0.5$ and $h = 1$. In which case the parameter update is largest?
d) **(7 points)** In feed-forward neural networks with many layers, Highway gates are very common. Explain the motivation of using Highway gates in deep networks?

**Solution** $MSE \triangleq \frac{1}{2}(y-a)^2$ **NO POINTS WERE TAKEN IF YOU HAVE NOT USED $\frac{1}{2}$ FACTOR**
First, we feed forward

$$\begin{bmatrix} 0.1 \\ 0.2 \\ 0.6 \\ 0.5 \end{bmatrix} \rightarrow \begin{bmatrix} 0.02 \\ 0.38 \\ 0.13 \end{bmatrix} \rightarrow \begin{bmatrix} 0.082h \ + 0.081(1-h) \\ 0.38h + 0 \cdot (1-h) \\ 0.13h + 0.137(1-h) \end{bmatrix} = \begin{bmatrix} 0.0566 \\ 0.152 \\ 0.1342 \end{bmatrix} \rightarrow 0.304 \tag{4}$$

a)

$$\frac{\partial C}{\partial w^2_{3,1}} = \frac{\partial C}{\partial a^3}\frac{\partial a^3}{\partial a^2_3}\frac{\partial a^2_3}{\partial \bar{H}_3}\frac{\partial \bar{H}_3}{\partial \tilde{a}^2_3}\frac{\partial \tilde{a}^2_3}{\partial \tilde{z}^2_3}\frac{\partial \tilde{z}^2_3}{\partial w^2_{3,1}} = (a-y) \cdot w^3_{1,3} \cdot (1-h) \cdot a^1_1 = (0.304-3) \cdot 0.5 \cdot 0.6 \cdot 0.02 \simeq -0.0161 \tag{5}$$

$$\frac{\partial C}{\partial h} = \frac{\partial C}{\partial a^3}\frac{\partial a^3}{\partial a^2}\left(\frac{\partial a^2}{\partial \bar{H}}\frac{\partial \bar{H}}{\partial h} + \frac{\partial a^2}{\partial H}\frac{\partial H}{\partial h}\right) = (a-y) \cdot w^3(a^1 - ReLU(w^2 a^1)) \tag{6}$$

$$= -2.696 \cdot [1.5, 1, 0.5] \begin{bmatrix} 0.02 \ - 0.081 \\ 0.38 - 0 \\ 0.13 - 0.137 \end{bmatrix} \simeq -0.7684 \tag{7}$$

b) We need the derivatives to update the learnable parameters.

c)

$$\frac{\partial C}{\partial w^1_{2,2}} = \frac{\partial C}{\partial a^3}\frac{\partial a^3}{\partial a^2}\left(\frac{\partial a^2}{\partial \bar{H}}\frac{\partial \bar{H}}{\partial \tilde{a}^2}\frac{\partial \tilde{a}^2}{\partial \tilde{z}^2}\frac{\partial \tilde{z}^2}{\partial a^1_2} + \frac{\partial a^2}{\partial H}\frac{\partial H}{\partial a^1_2}\right)\frac{\partial a^1_2}{\partial z^1_2}\frac{\partial z^1_2}{\partial w^1_{2,2}} = (a-y)w^3\left(\begin{bmatrix}(1-h)\cdot 1\cdot w^2_{1,2}\\(1-h)\cdot 0\cdot w^2_{2,2}\\(1-h)\cdot 1\cdot w^2_{3,2}\end{bmatrix} + \begin{bmatrix}0\\h\\0\end{bmatrix}\right)x_2 \tag{8}$$

$$= -2.696\cdot[1.5,1,0.5]\begin{bmatrix}(1-h)0.2\\h\\(1-h)0.4\end{bmatrix}\cdot 0.2 = -0.2696(1+h) \tag{9}$$

$$\frac{\partial C}{\partial w^1_{2,2}}|_{h=0} = -0.2696,\quad \frac{\partial C}{\partial w^1_{2,2}}|_{h=0.5} = -0.4044,\quad \frac{\partial C}{\partial w^1_{2,2}}|_{h=1} = -0.5392 \tag{10}$$

d) Highway gates improve gradient flow. We derive using the chain rule , therefore the more layers we have the more gradients multiplications we have on our derivation chain. This phenomena is known as vanishing gradients and Highway gates overcomes this by providing a better route for the gradients to flow in.

7) **Variant of MINE (32 pt)**

In this question we investigate an algorithm based on the mutual information neural estimator, using the following representation of mutual information:

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \tag{11}$$

Let $X \sim P_X$, $Y \sim P_Y$ and denote the joint PMF of $(X,Y)$ by $P_{XY}$. Let $U_X$ be the PMF of the uniform discrete probability measure over $\mathcal{X}$, the alphabet of $X$ (namely, $U_X(x) = \frac{1}{|\mathcal{X}|} \quad \forall x \in \mathcal{X}$).

a) **(5 points)** Prove the following equality:

$$H(X) = H(P_X, U_X) - D_{KL}(P_X \| U_X), \tag{12}$$

where $H(P_X, U_X)$ is the cross-entropy between $P_X$ and $U_X$.

b) **(5 points)** If we replace the uniform PMF $U_X$ by an arbitrary PMF $V_X$, does Eq. (12) still hold? Prove or disprove it.

c) **(5 points)** Based on the result of (a), prove the following equation:

$$I(X;Y) = D_{KL}(P_{XY}\|U_{XY}) - D_{KL}(P_X\|U_X) - D_{KL}(P_Y\|U_Y), \tag{13}$$

where $U_Y$ and $U_{XY}$ are defined in the same sense as $U_X$, on $\mathcal{Y}$ and $\mathcal{X}\times\mathcal{Y}$ respectively (assume that $|\mathcal{X}\times\mathcal{Y}| = |\mathcal{X}||\mathcal{Y}|$).

d) **(10 points)** Based on the KL divergence estimation method taught in class, propose an algorithm for the estimation of $I(X;Y)$ from a sample set $\{(x_i,y_i)\}_{i=1}^n \sim P_{XY}$, based on the equality proved in (b). Denote by $\widehat{I}_n^{(H)}(X,Y)$:

i) Write the optimization objective

ii) Give a block diagram of the proposed algorithm for estimating $\widehat{I}_n^{(H)}(X,Y)$. Assume the neural network consists of a single hidden layer with M units.

e) **(7 points)** We now wish to calculate the optimization objective $\widehat{I}_n^{(H)}(X,Y)$. For sufficiently large $n$, does the following hold? explain.

$$\widehat{I}_n^{(H)}(X,Y) \le I(X;Y) \tag{14}$$

**Solution**

a) Proof:

$$\begin{aligned} H(X) &= \mathbb{E}_{P_X}\left[\log\frac{1}{P_X}\right] \\ &= \mathbb{E}_{P_X}\left[\log\frac{U_X}{P_X U_x}\right] \\ &= \mathbb{E}_{P_X}\left[\log\frac{1}{U_X}\right] - \mathbb{E}_{P_X}\left[\log\frac{P_x}{U_Y}\right] \\ &= H(P_X, U_x) - D_{KL}(P_X\|U_X) \end{aligned}$$

b) We did not use the fact that $U_X$ is a uniform PMF, therefore, the above equality is true for every PMF $V_X$ such that the KL-divergence is well defined.

c) Proof:

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(P_X,U_X) - D_{KL}(P_X\|U_X) + H(P_Y,U_Y) - D_{KL}(P_Y\|U_Y) \\ &- (H(P_{XY},U_{XY}) - D_{KL}(P_{XY}\|U_{XY})) \\ &= D_{KL}(P_{XY}\|U_{XY}) - D_{KL}(P_X\|U_X) - D_{KL}(P_Y\|U_Y) \\ &+ H(P_X,U_X) + H(P_Y,U_Y) - H(P_{XY},U_{XY}). \end{aligned}$$

Let us show that the cross entropies cancel out (denote by $H_X, H_Y, H_{XY}$):

$$\begin{aligned}
H_X + H_Y - H_{XY} &= \mathbb{E}_{P_X}\left[\log\frac{1}{U_X}\right] + \mathbb{E}_{P_Y}\left[\log\frac{1}{U_Y}\right] - \mathbb{E}_{P_{XY}}\left[\log\frac{1}{U_{XY}}\right] \\
&= \sum_{x\in\mathcal{X}} P_X(x)\log\frac{1}{U_X(x)} + \sum_{Y\in\mathcal{Y}} P_Y(y)\log\frac{1}{U_Y(y)} - \sum_{x,y\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\log\frac{1}{U_{X,Y}(x,y)} \\
&= \sum_{x\in\mathcal{X}} P_X(x)\log\frac{1}{|\mathcal{X}|} + \sum_{Y\in\mathcal{Y}} P_Y(y)\log\frac{1}{|\mathcal{Y}|} - \sum_{x,y\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\log\frac{1}{|\mathcal{X}||\mathcal{Y}|} \\
&= \log\frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} P_X(x) + \log\frac{1}{|\mathcal{Y}|}\sum_{Y\in\mathcal{Y}} P_Y(y) - \log\frac{1}{|\mathcal{X}||\mathcal{Y}|}\sum_{x,y\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y) \\
&= \log\frac{1}{|\mathcal{X}|} + \log\frac{1}{|\mathcal{Y}|} - \log\frac{1}{|\mathcal{X}||\mathcal{Y}|} \\
&= \log\frac{|\mathcal{X}||\mathcal{Y}|}{|\mathcal{X}||\mathcal{Y}|} \\
&= 0.
\end{aligned}$$

Therefore,

$$I(X:Y) = D_{KL}(P_{XY}\|U_{XY}) - D_{KL}(P_X\|U_X) - D_{KL}(P_Y\|U_Y) \tag{15}$$

d) Solution:

i) We follow the steps taken in class - we use the Donsker-Varadhan representation and replace the expectations with empirical means. The objective is of the form:

$$\begin{aligned}
I_n^{(H)}(X;Y) = &\sup_{\theta_{XY}\in\Theta_{XY}} \frac{1}{n}\sum_{i=1}^{n} T_{\theta_{XY}}(x_i, y_i) - \log\left(\frac{1}{n}\sum_{i=1}^{n} e^{T_{\theta_{XY}}(\widetilde{x}_i,\widetilde{y}_i)}\right) \\
&- \sup_{\theta_X\in\Theta_X} \frac{1}{n}\sum_{i=1}^{n} T_{\theta_X}(x_i) - \log\left(\frac{1}{n}\sum_{i=1}^{n} e^{T_{\theta_X}(\widetilde{x}_i)}\right) \\
&- \sup_{\theta_Y\in\Theta_Y} \frac{1}{n}\sum_{i=1}^{n} T_{\theta_Y}(y_i) - \log\left(\frac{1}{n}\sum_{i=1}^{n} e^{T_{\theta_Y}(\widetilde{y}_i)}\right)
\end{aligned}$$

ii) We denote the objectives of the supremization problems by $L_{XY}$, $L_X$, and $L_Y$ respectively. A block diagram is provided in the attached figure.

e) Our optimization objective consists of a difference of supremums. Therefore, we cannot claim it is either a lower or upper bound on the true value of the mutual information. Consequently, we cannot state that the inequality hold.

Good Luck!

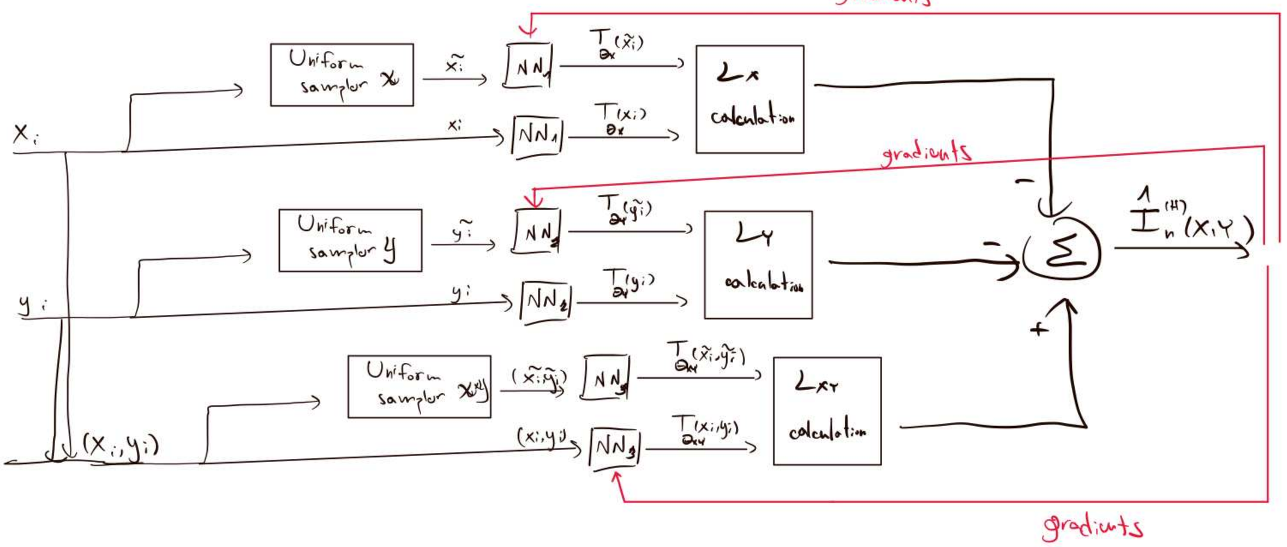
Our building blocks:

1) Uniform samplers, one for each of the alphabets ($\mathcal{X}, \mathcal{Y}, \mathcal{X} \times \mathcal{Y}$)

2) Three neural nets, each of the same structure (1 hidden layer, M units)
Denote by NN1, NN2, NN3

3) Calculator of the Donsker-Varadhan loss for each estimated KL divergence.

The block diagram




* In practice we don't know $\mathcal{X}, \mathcal{Y}$, therefore we estimate it from the symbol values we observe from $\{X_i, Y_i\}_{i=1}^{n}$

Fig. 3: Proposed block diagram for 7.d.