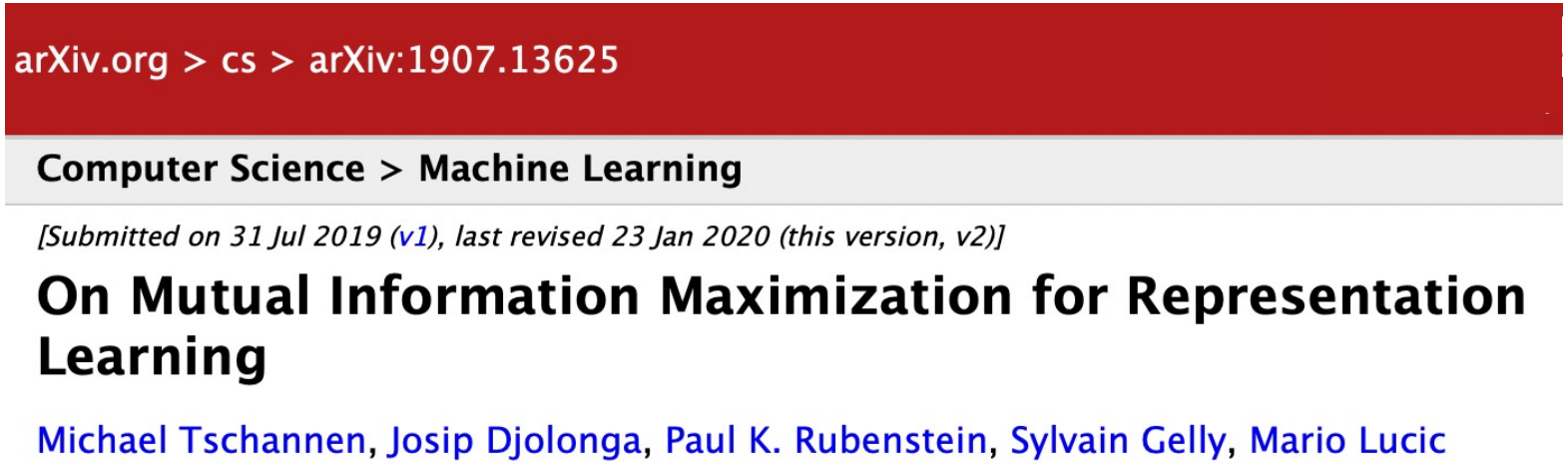
# Apprentissage de représentation par maximisation de l'information mutuelle

## une bonne idée, vraiment ?

Séminaire TALia du 11/02/2022

Source



Une bonne idée ?
Peut-on la justifier ?
Pourquoi ça marche, parfois ?

# Rappels de théorie de l'information

**Entropie** : nombre de bits minimal nécessaires en moyenne pour encoder un m

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = E_p \log \frac{1}{p(X)}$$

$$L(C) = \sum_{x \in \mathcal{X}} p(x) l(x)$$

**Entropie conditionnelle**, **divergence de KL** et **information mutuelle**

$$H(X|Y) = E_p \log \frac{1}{p(X|Y)}$$

$$D(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \geq 0$$

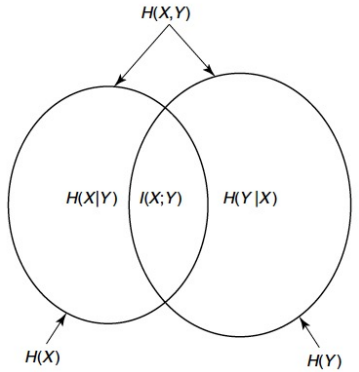
$$I($$

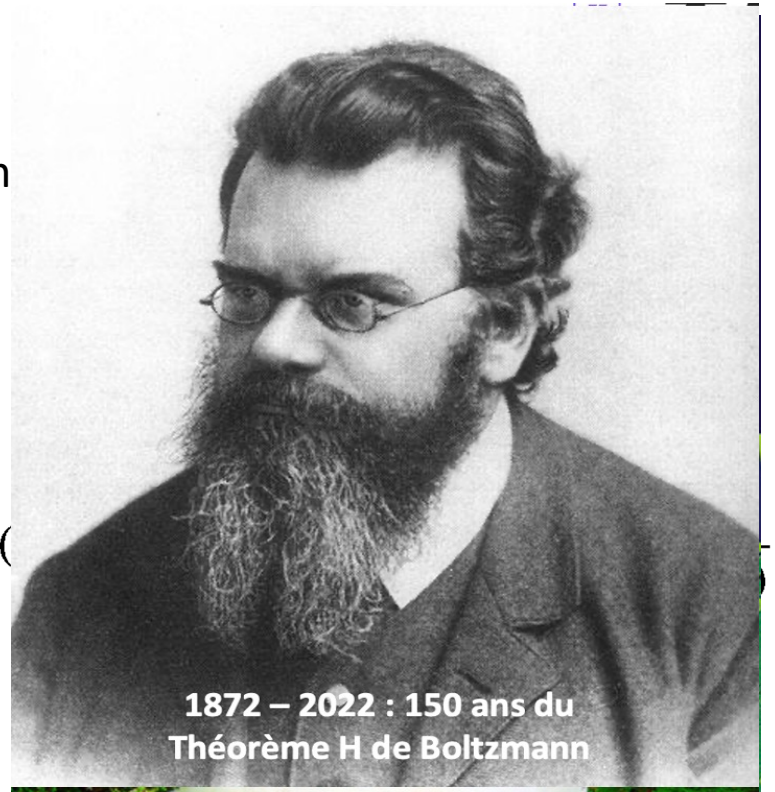
**Relations importantes**

$$H(X) = \log |\mathcal{X}| - D(p||u)$$

$$H(X|Y) \leq H(X)$$

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$



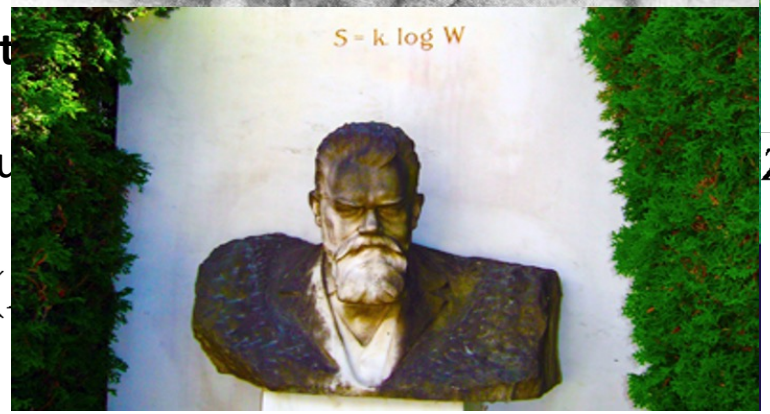




**Dat**

Pou

$Z$

$I($



# L'intuition derrière le principe InfoMax

1. Trouver une représentation $g(X)$ qui **maximise l'information mutuelle avec le signal** $X$.

$$\max_{g \in \mathcal{G}} I(X; g(X))$$

Si $g$ était une bijection on aurait : $I\big(X, g(X)\big) = H(X)$.

2. Si l'on a **deux vues** différentes $X^{(1)}$ et $X^{(2)}$ du signal $X$ on montre comme conséquence de la data processing inequality

$$I(g_1(X^{(1)}); g_2(X^{(2)})) \leq I(X; g_1(X^{(1)}), g_2(X^{(2)}))$$

Exemples pour $g_1$ et $g_2$ : les activations sur différentes couches d'un CNN.

on a donc intérêt à maximiser l'information mutuelle entre ces deux représentations.

Avantages de cette reformulation :

- La borne inférieure ne dépend que des représentations $g_1(X^{(1)})$ et $g_2(X^{(2)})$ qui sont généralement de **faible dimension**.
- On a une **grande flexibilité** dans le choix des encodeurs $g_1$ et $g_2$.

# Les difficultés commencent…

1. Estimer la valeur exacte de $I(X,Y)$ s'avère **difficile** (croissance exponentielle de la taille de l'échantillon avec la précision souhaitée).
2. **Invariance sous reparamétrisation** : $I(X,Y) = I(X',Y')$ si $X' = f_1(X)$ et $Y' = f_2(Y)$ avec des bijections $f_1$ et $f_2$.

Il existe différents **estimateurs empiriques** $I_{\text{EST}}(.,.)$ de $I(.,.)$ utiles trouver des représentations $g_1(X^{(1)})$ et $g_2(X^{(2)})$ qui maximisent approximativement l'information mutuelle entre les encodages des deux vues de $X$:

$$\max_{g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2} \quad I_{\text{EST}}\left(g_1(X^{(1)}); g_2(X^{(2)})\right)$$

Ces estimateurs reposent sur l'idée que si l'on peut trouver un classifieur capable de distinguer les couples $(x_i, y_i)$ tirés de la loi conjointe $P(X,Y)$ des couples tirés du produit des marginales $P(X)P(Y)$ alors l'information mutuelle est grande.

# Une approximation empirique de l'information mutuelle

L'approximation *InfoNCE*

$$I(X,Y) \geq \mathbb{E}_{\prod_k p(x_k,y_k)} \underbrace{\frac{1}{K}\sum_{i=1}^{K}\left[\log \frac{e^{f(x_i,y_i)}}{\frac{1}{K}\sum_{j=1}^{K} e^{f(x_j,y_i)}}\right]}_{I_{\mathrm{NCE}}} \quad \forall\, f.$$

Comme pour la borne ELBO (cf. p.ex. le séminaire n°30 sur GMNN ou la théorie des VAE) on a une borne variationnelle qui laisse la liberté de choisir judicieusement la fonction critique $f$.

- Pour un $x_i$ donné, la **fonction critique** $f$ peut s'interpréter comme l'association d'un score à chacun des $y_1, \ldots, y_K$ selon qu'il a été tirée ou non conjointement à $x_i$.
- La fonction critique $f$ est **arbitraire**. On a donc une **borne inférieure variationnelle** à l'IM (comme pour l'ELBO utilisée dans les VAE ou dans GMNN).
- La borne inférieure $\mathbb{E}\, I_{\mathrm{NCE}}(X,Y)$ est maximisée par $f(x,y) = \log p(y|x)$.
- Quelques **choix usuels** pour la fonction critique : $f(x,y) = x^{\mathrm{T}} W y$, $f(x,y) = \phi_1(x)^{\mathrm{T}}\phi_2(y)$ où $f(x,y) = \phi([x,y])$.
- Dans $I_{\mathrm{NCE}}\big(g_1(X^{(1)}), g_2(X^{(2)})\big)[f]$ on maximise simultanément sur $g_1, g_2$ et $f$.

# Maximiser $I$ n'induit pas des représentations optimales !

Le contenu de l'article = une liste d'observations expérimentales :

1. Si les encodeurs $g_1$ et $g_2$ sont choisis parmi un ensemble $\mathcal{G}$ de fonctions inversibles, **l'information mutuelle exacte** $I\left(g_1(X^{(1)}), g_2(X^{(2)})\right)$ **ne dépend pas des encodeurs**. Pourtant, on observe que **maximiser** $I_{NCE}$ (ou d'autres approximations $I_{\mathrm{EST}}$) **améliore l'efficacité** des représentations pour certaines tâches en aval !

2. Si l'ensemble $\mathcal{G}$ des encodeurs $g_1$ et $g_2$ contient des fonctions non inversibles, la maximisation d'une approximation $I_{\mathrm{EST}}$ favorise les **encodeurs difficiles à inverser** (dont le **conditionnement est élevé**).

3. Plus l'ensemble $\mathcal{F}$ de fonctions parmi lesquelles on choisit le critique $f$ est grand (« grand capacité »), meilleure sera l'estimation $I_{\mathrm{EST}}$ de l'information mutuelle. Pourtant **des critiques « simples »** associés à de mauvaises bornes $I_{\mathrm{EST}}$ **conduisent parfois à de meilleures représentations**.

4. Pour des bornes $I_{\mathrm{EST}}$ de qualités égales on constate que l'impact du **choix d'architecture** des encodeurs $g_1$ et $g_2$ est **plus important que la qualité de la borne**.

# Réécriture de l'approximation $I_{\mathrm{NCE}}$ de l'information mutuelle

$$
\begin{aligned}
\mathbb{E}\, I_{\mathrm{NCE}} &= \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\log\frac{e^{f(x_i,y_i)}}{\frac{1}{K}\sum_{j=1}^{K}e^{f(x_i,y_j)}}\right] \\
&= \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\log\frac{1}{\frac{1}{K}\sum_{j=1}^{K}e^{f(x_i,y_j)-f(x_i,y_i)}}\right] \\
&= \mathbb{E}\left[-\frac{1}{K}\sum_{i=1}^{K}\log\frac{1}{K}\sum_{j=1}^{K}e^{f(x_i,y_j)-f(x_i,y_i)}\right] \\
&= \log K - \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\log\left(1+\sum_{j\neq i}e^{f(x_i,y_j)-f(x_i,y_i)}\right)\right].
\end{aligned}
$$

# Relation de $I_{\mathrm{NCE}}$ avec l'apprentissage contrastif

Pour $f(x, y) = \phi^{\mathrm{T}}(x)\phi(y)$ et $g_1 = g_2$ **maximiser l'approximation $I_{\mathrm{NCE}}$ de l'information mutuelle** est équivalent à **minimiser** une fonction de **coût d'apprentissage contrastif**

$$\mathcal{L}_{\mathrm{NCE}}\left(\{(x_i, y_i)\}_{i=1}^{K}, \phi\right) = \frac{1}{K}\sum_{i=1}^{K}\log\left(1 + \sum_{j\neq i} e^{\phi(x_i)^{\top}\phi(y_j) - \phi(x_i)^{\top}\phi(y_i)}\right)$$

based metric learning !?!

$$I(X, Y) \geq \mathbb{E}_{\prod_k p(x_k, y_k)}\frac{1}{K}\sum_{i=1}^{K}\left[\log\frac{e^{f(x_i, y_i)}}{\frac{1}{K}\sum_{j=1}^{K} e^{f(x_j, y_i)}}\right]$$

Discussion incompréhensible dans l'appendice D à propos de l'importance pour les negative samples d'être tirés i.i.d !?!

# NCE dans deux séminaires précédents

## Séminaire n°27 : Spectral Contrastive Loss (SCL)

**Apprentissage self-supervisé** contrastif de features $f : \chi \to \mathbb{R}^k$

$$\mathcal{L}(f) := -2\,\mathbb{E}_{x,x^+}\left[f(x)^\mathsf{T} f(x^+)\right] + \mathbb{E}_{x,x^-}\left[\left(f(x)^\mathsf{T} f(x^-)\right)^2\right]$$

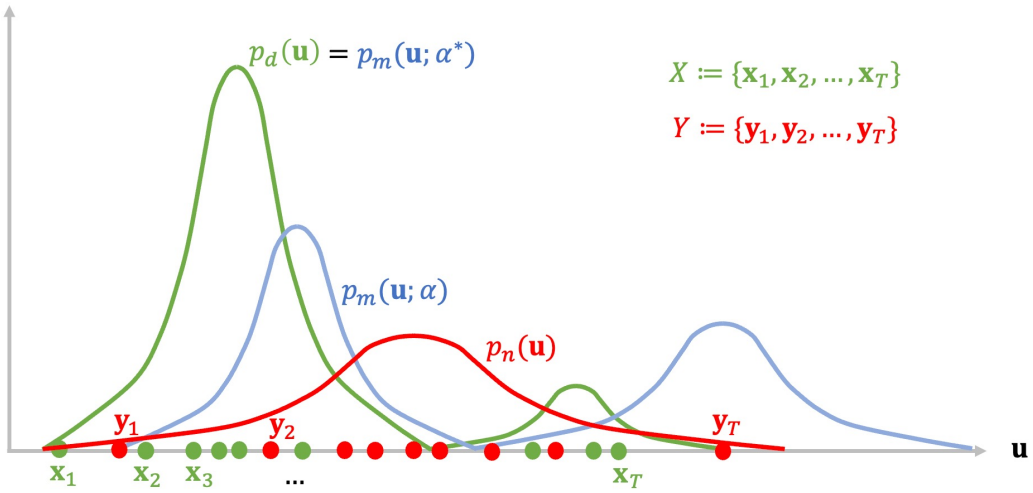
$(x, x^+)$ : même original

$(x, x^-)$ : originaux différents

## Séminaires n°15 et 17 : Noise Contrastive Estimation

$$P(C = 1|\mathbf{u};\theta) := h(\mathbf{u};\theta) = \frac{1}{1 + \exp[-\ln p_m(\mathbf{u};\theta) + p_n(\mathbf{u})]}$$

$$J_T(\theta) := \frac{1}{2T}\sum_{t=1}^{T} \ln[h(\mathbf{x}_t;\theta)] + \ln[1 - h(\mathbf{y}_t;\theta)]$$



# Résumé

- Des **résultats empiriques récents ont utilisé avec succès l'idée du principe InfoMax** pour générer des représentations utiles de données (pour de la classification d'images p.ex.).
- Utiliser **un concept classique de la théorie de l'information** est séduisant, à priori.
- L'information mutuelle exacte est cependant **difficile à estimer expérimentalement** et elle est **invariante sous les bijections**. Pour cette raison on cherche à **construire des approximations** qui sont des bornes inférieures.
- Mais à y regarder de plus près on constate qu'**il n'existe pas de raisons évidentes pour qu'InfoMax induise de bonnes représentations** pour des applications en aval qui les utilisent.
- Expérimentalement on constate que **des représentations « simples »** = de faible capacité, qui conduisent à de mauvaises bornes sur $I$, **fournissent parfois de meilleures représentations** pour les applications en aval.
- L'origine des performances pratiques d'InfoMax est plutôt à chercher dans les **biais inductifs cachés des représentations $g_1$ et $g_2$ utilisées**.
- On peut essayer de comprendre le succès pratique d'InfoMax en reliant l'approximations $I_{\mathrm{NCE}}$ de $I$ à une fonction de coût $\mathcal{L}_{\mathrm{NCE}}$ d'une méthode **d'apprentissage contrastif**.
- Pour une meilleure compréhension de ces phénomènes il faudrait élaborer des **notions d'informations plus fines** qui prennent en compte aussi bien la **quantité d'information** que la **géométrie des représentations** induites ou la **difficulté à extraire cette information** pour les tâches en aval.