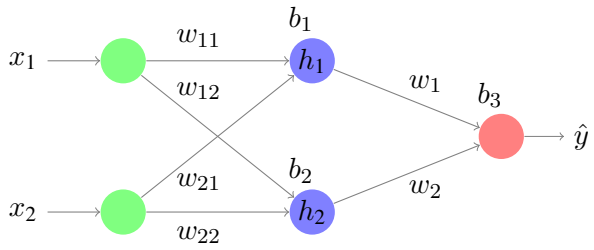
CORRECTIONS DE L'EXAMEN D'APPRENTISSAGE DU LUNDI 12 MAI 2025

Préambule

Cette première question sera utile pour répondre à la question 2 de l'exercice suivant

Considérons le réseau à 2 couches suivant :



$$\begin{cases} \hat{y} = \sigma(z) \text{ avec } z = w_1 h_1 + w_2 h_2 + b_3 \\ h_1 = f(z_1) \text{ avec } z_1 = w_{11} x_1 + w_{21} x_2 + b_1 \\ h_2 = f(z_2) \text{ avec } z_2 = w_{12} x_1 + w_{22} x_2 + b_2 \end{cases}$$

1. (2 pts) Exprimez le gradient $\frac{\partial J}{\partial w_{11}}$ de la fonction objectif J par rapport au paramètre w_{11} comme un produit de dérivées partielles, et explicitez les différents termes de ce produit.

   Réponse : La *chain-rule* nous permet de décomposer ce gradient en le produit suivant :

   $$\frac{\partial J}{\partial w_{11}} = \frac{\partial J}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial h_1} \frac{\partial h_1}{\partial z_1} \frac{\partial z_1}{\partial w_{11}}$$

   où

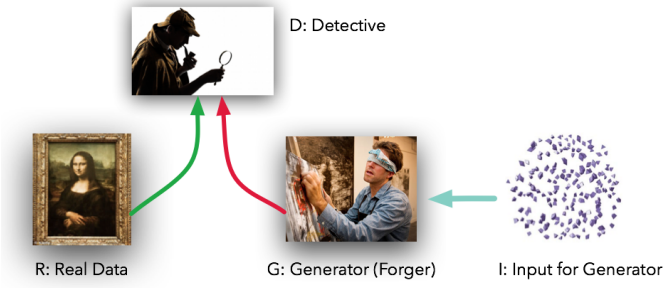
   - $\frac{\partial J}{\partial \hat{y}}$ dépend de la dérivée de la fonction de coût utilisée pour évaluer la qualité des prédictions du modèle.
   - $\frac{\partial \hat{y}}{\partial z}$ est la dérivée de la fonction sigmoide évaluée en z.
   - $\frac{\partial z}{\partial h_1}$ est la valeur du poids synaptique w_1
   - $\frac{\partial h_1}{\partial z_1}$ est la dérivée de la fonction d'activation des couches cachées évaluée en z_1.
   - $\frac{\partial z_1}{\partial w_{11}}$ est la valeur de l'entrée x_1

   .

Exercice 1 : Réseaux Génératifs Antagonistes (GAN)

Les GAN (*Generative Adversarial Networks*, en français réseaux génératifs antagonistes) sont des réseaux de neurones appartenant à la famille des méthodes génératives. Ils ont été très populaires entre 2015 (année de leur première proposition par Goodfellow et al.) et 2020 et sont à l'origine, par exemple, des premiers *deep fakes*. Dans ce problème, nous allons présenter quelques idées derrière leur fonctionnement et leur implémentation.

Il n'est, bien sûr, pas nécessaire de connaître ces réseaux pour pouvoir répondre aux questions.



La figure ci-dessus donne l'idée derrière le principe des GAN. L'idée d'un GAN est de mettre deux réseaux de neurones en compétition :

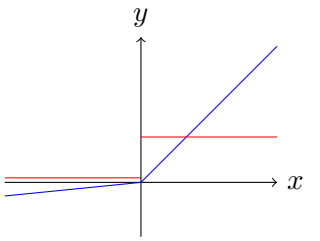
- Le **Générateur** est le réseau principal, sa tâche est d'apprendre à générer des données le plus réalistes possibles. Sur la figure, il apparait comme *Forger*, car tel un faussaire, ce réseau doit générer des données qui ont l'air réelles.
- Le **Discriminateur** est un second réseau créé dans le but d'entraîner le Générateur. Sa tâche est de déterminer, pour une donnée fournie en entrée, si la donnée est réelle ou si elle a été générée par le Générateur. Sur la figure, ce réseau apparait sous la forme d'une *Detective*, il doit être capable de discerner des données réelles des "faux" créés par le Générateur.

Les deux réseaux ont des objectifs antagonistes : l'objectif du Générateur est de produire des données qui vont tromper le Discriminateur, alors que l'objectif du Discriminateur est de réussir à discerner les "fausses" données produites par le Générateur. Les deux réseaux sont entraînés en même temps et s'améliorent conjointement, aboutissant si tout se passe bien à un Générateur capable de produire des données parfaitement réalistes, et un Discriminateur incapable de distinguer une donnée réelle d'une donnée générée.

1. (2 pts) Le Discriminateur est un réseau de neurones qui prend en entrée une donnée (par exemple, une image de visage dans le cas des *deep fakes*) et qui doit déterminer si elle est réelle ou générée. Coment s'appelle le type de problème auquel répond ce réseau ? Comment, en conséquence, doit-on construire la couche de sortie du Discriminateur ? (précisez le nombre de neurones, et la fonction d'activation qui doit être utilisée).
   Réponse : Le Discriminateur doit résoudre un problème de classification binaire; sa couche de sortie doit donc posséder un seul neurone, et porter une fonction d'activation sigmoide.

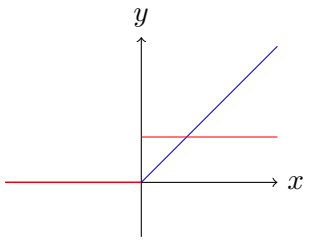
   L'article DCGAN (Radford et al.) propose une étude approfondie des différentes architectures possibles pour la construction des Générateur et Discriminateur dans le cadre de la génération d'image. L'une des propositions de l'article est d'utiliser une fonction d'activation des couches cachées différente pour le Générateur et le Discriminateur. En effet, pour que l'apprentissage se passe bien, il est nécessaire que le Discriminateur soit un peu "meilleur" que le Générateur. Radford et al. ont donc choisi une fonction d'activation pour le Discriminateur ayant pour objectif que celui-ci optimise sa fonction de perte plus rapidement que le Générateur.

   Les courbes suivantes décrivent les deux fonctions d'activation (en bleu) utilisées sur les couches cachées du Générateur et du Discriminateur avec leurs dérivées respectives (en rouge) :



$$\text{leakyReLU}(x) = \begin{cases} \alpha x & \text{si } x < 0 \\ x & \text{sinon} \end{cases}$$

Leaky Rectified Linear Unit



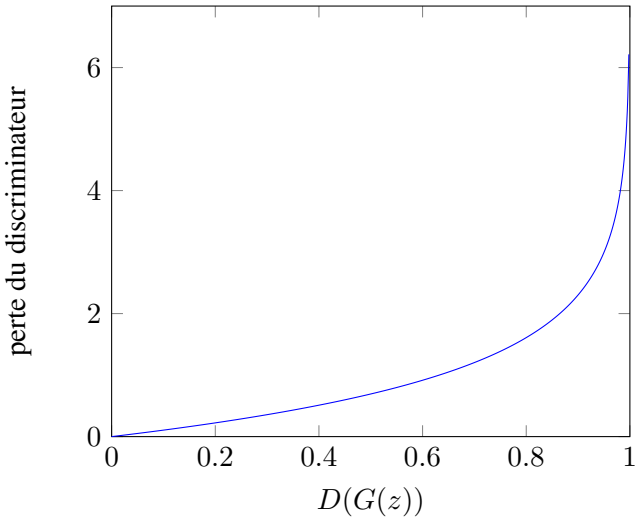
$$\text{reLU}(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{sinon} \end{cases}$$

Rectified Linear Unit

2. (1 pt) En regardant les valeurs que peut prendre la dérivée de ces deux fonctions, expliquez pourquoi une des deux fonctions est susceptible de favoriser l'optimisation d'un réseau de neurones par rapport à l'autre.
   Réponse : La fonction `reLU` a une dérivée nulle sur $\mathbb{R}^-$, ce qui signifie que lorsque l'activation d'un neurone est négative, les gradients des paramètres liés à ce neurones seront nuls : ils n'évolueront plus. Si l'on se réfère au préambule, cela signifie par exemple que si l'activation h_1 est négative, le paramètre $w_1 1$ dont la mise à jour dépend de $\frac{\partial h_1}{\partial z_1}$ (qui vaut donc 0 dans ce cas) n'évolue pas. Ce problème n'arrive pas avec la fonction `leakyReLU` dont la dérivée est non nulle sur $\mathbb{R}^-$ pour éviter ce cas de figure.

3. (1 pt) Compte tenu de votre réponse précédente, quelle fonction d'activation est utilisée pour quel réseau ? Justifiez votre réponse.
   Réponse : L'apprentissage étant moins pénalisé avec la fonction `leakyReLU` qu'avec la fonction `reLU`, il faut utiliser la fonction `leakyReLU` pour les couches cachées du Discriminateur et la fonction `reLU` pour les couches cachées du Générateur.

   La figure suivante présente l'allure de la fonction de perte du Discriminateur en fonction de la valeur $D(G(z))$, c'est-à-dire de la prédiction du Discriminateur appliqué à une donnée générée. Si $D(G(z))$ est proche de 0, le Discriminateur reconnaît à juste titre que son entrée est une donnée générée. Si $D(G(z))$ est proche de 1, alors le Discriminateur est en train d'être trompé par le Générateur.



4. (1,5 pts) Au début de l'apprentissage, le Générateur est très mauvais et il est assez facile pour le Discriminateur de détecter qu'une donnée est réelle ou générée. En observant l'allure de

la fonction (et en réfléchissant aux valeurs de la dérivée de cette fonction), expliquez pourquoi l'entraînement des GAN est souvent très lent.
Réponse : Si le Discriminateur arrive facilement à dire qu'une donnée est générée, cela signifie que $D(G(z))$ est proche de 0. La fonction de perte a des gradients très faibles lorsque $D(G(z))$ tend vers 0, ce qui signifie que les mises à jour des paramètres seront également très faibles et donc que le réseau évolue lentement.

L'une des raisons fondamentales pour lesquelles les GAN sont aujourd'hui supplantés par d'autres types de méthodes génératives réside dans l'instabilité de leur entraînement. Lorsque le Générateur commence à devenir performant, il arrive fréquemment que l'algorithme d'optimisation se mette à diverger soudainement.

5. (1,5 pts) D'après vous, et toujours en observant l'allure de la courbe, d'où vient l'instabilité de l'entraînement des GAN ?
Réponse : Lorsque le Générateur devient trop performant, le Discriminateur a de plus en plus de mal à discerner une donnée réelle d'une donnée générée, et donc $D(G(z))$ s'approche de 1. La fonction de perte a des gradients très forts lorsque $D(G(z))$ tend vers 1, ce qui signifie que les mises à jour des paramètres seront de grande amplitude et sont donc susceptibles de modifier fortement les paramètres du réseau, jusqu'à parfois le faire diverger.

Le tableau ci-dessous résume les couches que l'on peut trouver dans un Discriminateur construit pour déterminer si des images de chiffres manuscrits sont réelles ou générées.

Type de couche	Caractéristiques	Dim. tenseur d'entrée	Dim. tenseur de sortie
Convolution	32 filtres de taille 3×3 **padding = 1** et **stride = 2**	$32 \times 32 \times 1$	$16 \times 16 \times 32$
Convolution	64 filtres de taille 3×3 **padding = 1** et **stride = 2**	$16 \times 16 \times 32$	$8 \times 8 \times 64$
Vectorisation (Flatten)		$8 \times 8 \times 64$	4096
Dense	32 neurones	4096	32
Dense (sortie)	1 neurone	32	1

6. (2 pts) Complétez le tableau en indiquant la dimension des tenseurs.

7. (1 pt) Combien de paramètres compte la première couche de convolution (première ligne du tableau) ?
Réponse : Le stride et le padding n'ont aucune influence sur le nombre de paramètres de la couche. Il y a 32 filtres de taille $3 \times 3 \times 1$ (car l'image d'entrée n'a qu'un seul canal) ce qui fait donc $32 \times 3 \times 3 = 288$ coefficients de filtre auxquels s'ajoutent 1 biais pour chaque filtre, soit un total de $288 + 32 = 320$ paramètres.

Questions sur l'article (10 **points)**

1. (1 pt) Expliquer ce qu'on entend par "segmentation d'images" qui apparaît dans le titre de cet article.
   Réponse : La segmentation d'images consiste à découper une image en régions homogènes. Il s'agit de classifier les pixels de l'image en différents classes associées à ces régions homogènes.

2. (1 pt) Dans leur introduction, les auteurs parlent d'un modèle de mélange de gaussiennes (Gaussian mixture model). Pouvez-vous expliquer de quoi il s'agit et donner la densité de probabilité d'un vecteur $\boldsymbol{x}$ issu de ce modèle ?
   Réponse : On parle de mélange de gaussiennes lorsque le vecteur $\boldsymbol{x}$ peut provenir de K classes différentes avec des probabilités potentiellement différentes. Si on note ω_K ces classes et π_k les probabilités d'appartenance à ces classes, la densité de $\boldsymbol{x}$ s'écrit

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|\omega_k),$$

   où $p(\boldsymbol{x}|\omega_k)$ est la loi de $\boldsymbol{x}$ conditionnellement à la classe ω_k. Pour un mélange de gaussiennes, $p(\boldsymbol{x}|\omega_k)$ est la densité d'un vecteur gaussien définie par son vecteur moyenne $\boldsymbol{m}_k$ et sa matrice de covariance $\boldsymbol{\Sigma}_k$.

3. (1 pt) Dans la section II, les auteurs proposent de normaliser les données à l'aide de l'équation (1). Si $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ désignent les n pixels de l'image avec $\boldsymbol{x}_j = [x_j(1), ..., x_j(n)]^T$, rappeler les expressions de $\mu_i \in \mathbb{R}$ et de $\sigma_i \in \mathbb{R}$ intervenant dans l'équation (1).
   Réponse : Les valeurs de μ_i et de σ_i intervenant dans l'équation (1) sont des estimateurs de la moyenne et de l'écart-type des valeurs de la $i$ème bande spectrale. On pourra donc utiliser les expressions suivantes :

$$\mu_i = \frac{1}{n}\sum_{j=1}^{n} x_j(i), \ \sigma_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}[x_j(i) - \mu_i]^2.$$

4. (1 pt) Après l'équation (3), les auteurs parlent du vecteur $\boldsymbol{x}_i^{\max}$. Expliquer comment déterminer ce vecteur à l'aide des n pixels de l'image $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ et ce que ce vecteur $\boldsymbol{x}_i^{\max}$ représente.
   Réponse : La fonction $M(\boldsymbol{x}_i)$ définie dans (3) mesure la densité de points situés dans un voisinage de $\boldsymbol{x}_i$ (plus cette quantité est grande, plus il y a de points dans un voisinage de $\boldsymbol{x}_i$). Le vecteur $\boldsymbol{x}_i^{\max}$ est le pixel maximisant cette densité, c'est-à-dire le point contenant le plus de voisins dans l'image.

5. (1 pt) Expliquer la signification de SAM$(\boldsymbol{x}, \boldsymbol{y})$ défini dans (7) et donner l'intervalle des valeurs possibles de cette quantité.
   Réponse : Si $\boldsymbol{x} = [x(1), ..., x(d)]^T$ et $\boldsymbol{y} = [y(1), ..., y(d)]^T$ sont deux pixels de l'image, le coefficient de corrélation entre ces deux vecteurs est

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{d} x_i y_i}{\sqrt{\sum_{i=1}^{d} x_i^2}\sqrt{\sum_{i=1}^{d} y_i^2}}.$$

   C'est une quantité à valeurs dans $]-1, +1[$. La fonction $\cos^{-1}$ est une fonction décroissante de $]-1, 1[$ dans $]\pi, 0[$. La quantité $\cos^{-1}[\rho(\boldsymbol{x}, \boldsymbol{y})]$ détermine l'angle entre les deux vecteurs $\boldsymbol{x}$ et $\boldsymbol{y}$. Plus cette valeur est petite, plus les vecteurs $\boldsymbol{x}$ et $\boldsymbol{y}$ sont proches (au sens d'une distance angulaire). La valeur de SAM telle qu'elle est définie dans l'article prend ses valeurs dans l'intervalle $]1-\pi, 1[$ et est d'autant plus grande que l'angle entre les deux vecteurs $\boldsymbol{x}$ et $\boldsymbol{y}$ est faible. C'est une mesure de similarité angulaire entre $\boldsymbol{x}$ et $\boldsymbol{y}$.

6. (1 pt) Quelle est la règle de décision de la méthode one-class SVM permettant de déterminer si un vecteur $\boldsymbol{x}$ est une anomalie ou pas (lorsqu'on n'utilise pas de noyau et pas de variable de relaxation ξ_i) ?
Réponse : L'idée de la méthode one-class SVM est décider qu'un vecteur $\boldsymbol{x}$ est une anomalie si ce vecteur est proche de l'origine avec une frontière de séparation linéaire (hyperplan). On décidera donc que $\boldsymbol{x}$ est une anomalie si
$$\boldsymbol{w}^T\boldsymbol{x} - \rho \leq 0.$$

7. (1 pt) La méthode one-class SVM avec relaxation consiste à résoudre le problème suivant
$$\boxed{\begin{array}{l}\text{minimize } \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{with the constraints } \boldsymbol{w}^T\boldsymbol{x}_i \geq 1-\xi_i,\ \xi_i \geq 0, \forall i\end{array}}$$
Quel est le rôle des variables de relaxation ξ_i (appelées "slack variables" dans l'article) ? Que signifie $\xi_i = 0$? Expliquer comment la méthode se comporte lorsque $C = 0$ ou lorsque C prend une forte valeur.
Réponse : Les variables ξ_i permettent à certains vecteurs $\boldsymbol{x}_i$ de ne pas respecter la contrainte $\boldsymbol{w}^T\boldsymbol{x}_i \geq 1$. Lorsque $\xi_i = 0$, le vecteur $\boldsymbol{x}_i$ vérifie la contrainte, ce qui n'est pas le cas quand $\xi_i > 0$. Lorsque la valeur de C est grande, on a beaucoup de vecteurs $\boldsymbol{x}_i$ qui vérifient la contrainte. Lorsque $C = 0$, tous les vecteurs $\boldsymbol{x}_i$ peuvent violer la contrainte. En pratique, il faut tester plusieurs valeurs de C et garder une valeur qui donne de bons résultats de détection d'anomalies.

8. (1pt) Les auteurs proposent de résoudre le problème
$$\boxed{\begin{array}{l}\text{Minimize } \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n\nu}\sum_{i=1}^{n}\xi_i - \rho \\ \text{with the constraints } \boldsymbol{w}^T\phi(\boldsymbol{x}_i) \geq \rho - \xi_i,\ \xi_i \geq 0, \forall i, \rho \geq 0\end{array}}$$
Quel sont les rôles de la fonction ϕ et du paramètre ν ?
Réponse : La fonction ϕ projette les données dans un espace de grande dimension dans lequel on espère que les données seront linéairement séparables. Le paramètre est une borne supérieure du nombre d'anomalies dans la base d'apprentissage. Par exemple, si $\nu = 0.01$, on a un maximum de 1% de vecteurs $\boldsymbol{x}_i$ détectés comme anomalies.

9. (1 pt) Comment reconnaît-on les vecteurs supports à partir de la solution du problème dual (13) ?
Réponse : Ce sont les vecteurs vérifiant $\alpha_i > 0$.

10. (1 pt) Dans la partie contenant les expérimentations, les auteurs parlent de la vérité terrain de l'image ROSIS (ground truth of ROSIS university. Expliquer ce qu'est cette vérité terrain.
Réponse : La vérité terrain indique la véritable classe de chaque pixel de l'image. Cette vérité terrain peut être obtenue par analyse de l'image par exemple par un photo-interprète.

11. (1 pt) Expliquer l'influence de la valeur de σ^2 dans l'équation (19).
Réponse : La valeur de σ^2 permet de régler la régularité de la frontière entourant le domaine des éléments normaux. Comme montré en cours, plus $\gamma = \frac{1}{\sigma^2}$ est faible, plus la frontière est régulière.

12. (1 pt) Résumer comment la classification de l'image a été effectuée pour les pixels avec et sans vérité terrain.
Réponse : Dans une première étape, on effectue un clustering avec subtractive clustering sur les données associées à la vérité terrain. Le nombre de classes de ce clustering est déterminé avec la méthode SV-PWSD. Pour les pixels restants, les auteurs proposent de classifier un pixel de classe inconnue à partir de la classe majoritairement représentées parmi ses k plus proches voisins.

13. (1 pt) Rappeler le principe de la méthode k-means (KM) qui est utilisée comme méthode de l'état de l'art.
Réponse : La méthode k-means est une méthode itérative qui nécessite tout d'abord de choisir un représentant de chaque classe. On forme ensuite des clusters à l'aide de la règle du plus proche voisins. Les points les plus proches du représentant du cluster i sont affectés au $i$ème cluster. Puis on calcule les barycentres des différentes classes ainsi formées qui deviennent les représentants des classes et on recommence l'opération jusqu'à ce que les groupes de points ne changent plus.