EXAMEN D'APPRENTISSAGE DU LUNDI 12 MAI 2025.

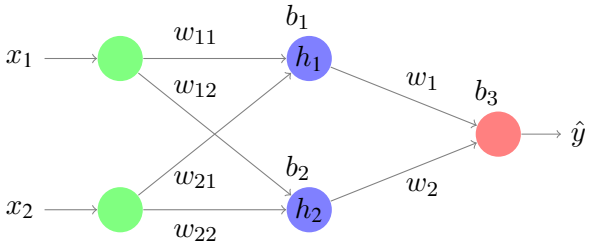
Documents autorisés :
1 feuille A4 Recto/Verso **manuscrite**.

Durée :
1h30 (+30 min tiers temps)

Préambule

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

Considérons le réseau à 2 couches suivant :



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

où σ désigne la fonction sigmoide, f la fonction d'activation de la couche cachée.

On rappelle qu'un tel réseau est entraîné en minimisant une fonction objectif J, qui évalue l'erreur commise par le réseau sur ses prédictions $\hat{y}^{(i)}$ à partir de données $x^{(i)}$ d'un ensemble d'apprentissage en les comparant à des labels $y^{(i)}$ au moyen d'une fonction de coût (par exemple, la MSE). Ici, on peut par exemple considérer que :

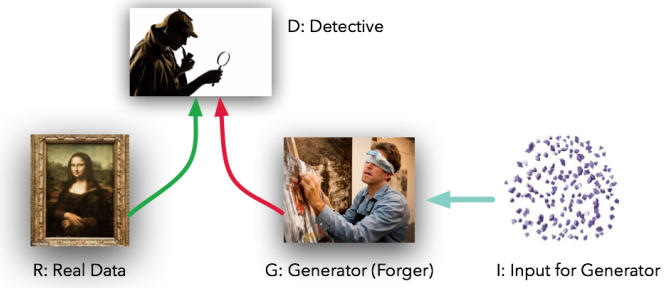
$$J = \frac{1}{m} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2$$

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.

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 :

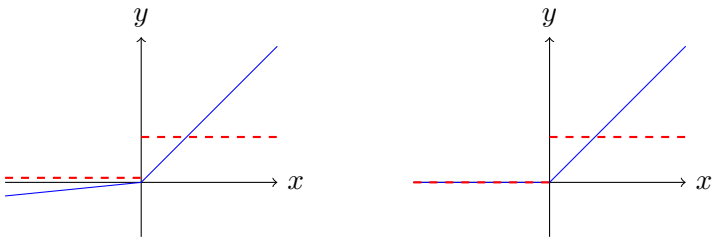
- Le **Générateur** est le réseau principal, sa tâche est d'apprendre à générer des données le plus réalistes possibles. Sur la figure, il apparait comme *Forger*, car tel un faussaire, ce réseau doit générer des données qui ont l'air réelles.
- Le **Discriminateur** est un second réseau créé dans le 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. (1 pt) 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).

   L'article DCGAN (Radford et al.) propose une étude approfondie des différentes architectures possibles pour la construction du générateur et du discriminateur dans le cadre de la génération d'images. 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 (trait plein) utilisées sur les couches cachées du Générateur et du Discriminateur, accompagnées de leurs dérivées respectives (en pointillés) :



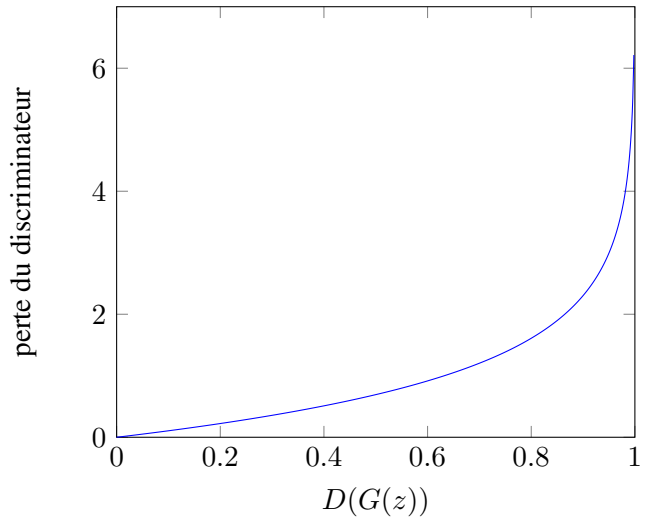
$$\text{leakyReLU}(x) = \begin{cases} \alpha x & \text{si } x < 0 \\ x & \text{sinon} \end{cases} \qquad \text{reLU}(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{sinon} \end{cases}$$

Leaky Rectified Linear Unit *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.

3. (1 pt) Compte tenu de votre réponse précédente, quelle fonction d'activation est utilisée pour quel réseau ? Justifiez votre réponse.

La figure suivante présente l'allure de la fonction de perte (loss) 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.



On rappelle, à titre d'indication pour les 2 prochaines questions, que la mise à jour d'un paramètre β à chaque itération k de la descente de gradient s'effectue à l'aide de la formule

$$\beta^{\{k\}} \leftarrow \beta^{\{k-1\}} - \alpha \frac{\partial J}{\partial \beta}$$

4. (1 pt) 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.

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 pt) D'après vous, et toujours en observant l'allure de la courbe, d'où vient l'instabilité de l'entraînement des GAN ?

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**	$32 \times 32 \times 1$	
Convolution	64 filtres de taille 3×3 **padding = 1** et **stride = 2**		
Vectorisation (Flatten)			
Dense	32 neurones		
Dense (sortie)	... neurone(s)		

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) ?

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.

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 ?

3. (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.

4. (1 pt) Expliquer la signification de SAM$(\boldsymbol{x}, \boldsymbol{y})$ défini dans (7) et donner l'intervalle des valeurs possibles de cette quantité.

5. (1 pt) La méthode one-class SVM avec relaxation consiste à résoudre le problème suivant

$$\boxed{\begin{array}{l}\text{minimize } \dfrac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{with the constraints } \boldsymbol{w}^T\boldsymbol{x}_i \geq 1-\xi_i,\ \xi_i \geq 0, \forall i\end{array}}$$

   Quel est le rôle des variables de relaxation ξ_i (appelées "slack variables" dans l'article) ? Que signifie $\xi_i = 0$?

6. (1 pt) Les auteurs proposent de résoudre le problème

$$\boxed{\begin{array}{l}\text{Minimize } \dfrac{1}{2}\|\boldsymbol{w}\|^2 + \dfrac{1}{n\nu}\sum_{i=1}^{n}\xi_i - \rho \\ \text{with the constraints } \boldsymbol{w}^T\phi(\boldsymbol{x}_i) \geq \rho - \xi_i,\ \xi_i \geq 0, \forall i, \rho \geq 0\end{array}}$$

   Quel sont les rôles de la fonction ϕ et du paramètre ν ?

7. (1 pt) Comment reconnaît-on les vecteurs supports à partir de la solution du problème dual défini par l'équation (13) de l'article ?

8. (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.

9. (1 pt) Expliquer l'influence de la valeur de σ^2 dans l'équation (19).

10. (1 pt) Expliquer comment la première classification de l'image a été effectuée pour les pixels de classe inconnue.