Examen de Science des Données, PHYS-231, 2023/24

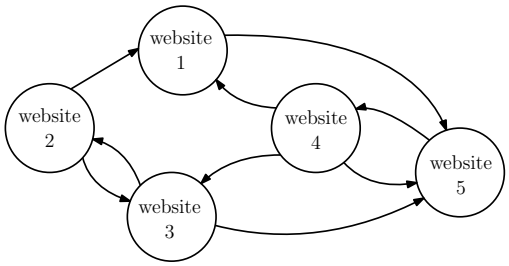
Nom/Sciper:

Instructions :

— Durée de l'examen : 3 heures, 26.01.2024 de 9h15 à 12h15, salle CM1105, CM1120.
— Matériel permis : 2 pages (i.e. une page recto-verso ou deux pages recto) de notes personnelles, papier, matériel pour écrire.
— Les problèmes peuvent être résolus dans n'importe quel ordre.
— Écrivez votre nom et prénom sur **toutes** les feuilles additionnelles que vous rendez. Vous pouvez utilisez la dernière page si vous avez besoin de plus d'espace.
— Le nombre total de points est 75.

1 PageRank & graphes [8 points]

1. (2 points) Donnez la matrice d'adjacence $A \in \{0,1\}^{5\times5}$ du réseau de sites web donné ci-dessous. $A_{ij} = 1$ si le site web j a un lien vers le site web i, sinon $A_{ij} = 0$. Les indices i et j indiquent respectivement les lignes et les colonnes de A.



2. (1 point) Quel site web a le plus grand degré sortant (*out-degree*) ? Quel site web a le plus grand degré entrant (*in-degree*) ?

3. (1 point) À quelle condition une matrice est-elle stochastique par colonnes (*column-stochastic*) ?

4. (1 point) Écrivez la version stochastique par colonne (*column-stochastic*) de la matrice d'adjacence A donnée ci-dessous, où les colonnes doivent être normalisées par le degré sortant (*out-degree*) du site web correspondant.

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix}$$

5. (2 points) Pour rappel, la matrice de Google G pour *PageRank* est définie comme

$$G = (1 - \epsilon)S + \epsilon I \tag{1}$$

où I est la matrice dont chaque élément vaut $1/n$ et $\epsilon \in (0, 1)$. Expliquez pourquoi nous ajoutons le second terme de la somme et pourquoi nous avons besoin de la pondération en $1 - \epsilon$ et ϵ.

6. (1 point) Vous lancez l'algorithme *PageRank* sur le sous-ensemble de Wikipédia de toutes les pages web comprenant le mot-clé « physique », et vous obtenez une liste des valeurs *PageRank* pour ces pages. La page web considérée comme la plus pertinente selon *PageRank* est-elle celle ayant la plus petite ou la plus grande valeur ?

2 SVD [8 points]

1. (2 points) Soit une matrice $X \in \mathbb{C}^{n \times d}$. Définissez la décomposition en valeurs singulières (SVD) de X.

2. (2 points) Soit une matrice $X \in \mathbb{C}^{n \times d}$. Définissez ce que sont les vecteurs singuliers à gauche et à droite (*left and right singular vectors*). Quel est leur lien avec la décomposition en valeurs singulières de X ?

3. (2 points) Soit une matrice $X \in \mathbb{C}^{n \times d}$. Quelle est la meilleure approximation de rang k de cette matrice, c'est-à-dire celle qui minimise l'erreur quadratique moyenne (c'est-à-dire la norme de Frobenius de la difference entre X et son approximation) ?

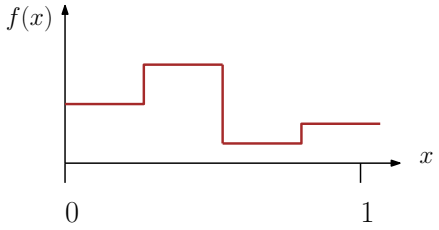
4. (2 points) Supposez que la matrice $X \in \mathbb{R}^{n \times d}$ représente un jeu de données. Le jeu de données contient n points, chacun de dimension d. Soit $X = U\Sigma V^*$ la décomposition en valeurs singulières (SVD) de X. Quelle est l'interprétation des k premiers vecteurs singuliers à droite (*right singular vectors)* ?

3 Descente de gradient [4 points]

1. (1 point) Écrivez l'équation pour une itération de l'algorithme de descente de gradient qui a pour but de minimiser une fonction $L(w)$ où $w \in \mathbb{R}^d$.

2. (2 points) Considérez le taux d'apprentissage (*learning rate*) dans l'algorithme de descente de gradient. Décrivez un inconvénient à prendre un taux d'apprentissage très petit. Décrivez un inconvénient à prendre un taux d'apprentissage très grand.

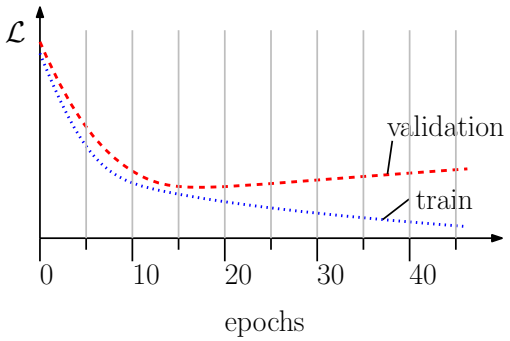
3. (1 point) Pourquoi n'est-ce pas une bonne idée d'utiliser la fonction suivante comme fonction de perte (*loss function*) pour l'optimisation basée sur la descente de gradient ?



4 Erreur d'entraînement, de validation et de test [6 points]

1. (2 points) Quand nous avons parlé de sur-apprentissage (*overfitting*), nous avons décrit l'utilisation des jeux de données de validation, d'entraînement et de test (*validation, training and test sets*). Il est important que les échantillons du jeu de données soient divisés de manière aléatoire pour créer les jeux de validation, d'entraînement et de test. Quels problèmes pourraient survenir si nous prenions le premier tiers du jeu de données pour l'entraînement, le second tiers pour la validation et le dernier tiers pour le test ?

2. (1 point) Vous entraînez un modèle de classification pour classifier des chats et des chiens en utilisant la descente de gradient. Vous observez les courbes de perte (*loss curves*) suivantes pour le jeu de données d'entraînement et de validation au fil du temps :



Le modèle fait-il du sur-apprentissage (*overfitting*) à l'époque (*epoch*) 45 ? Justifiez votre réponse.

3. (2 points) Suite à la question précédente, vous enregistrez les paramètres du modèle à chaque époque durant l'entraînement. De ces paramètres enregistrés, vous voulez choisir ceux qui sont le meilleur pour classer les chats et les chiens. Expliquez ce qu'est l'arrêt précoce (*early stopping*), et choisissez l'époque sur le graphique (à une précision de ~ 5 époques) que vous devriez choisir pour faire un bon arrêt précoce.

4. (1 point) La précision d'entraînement (*training accuracy*) et de validation (*validation accuracy*) pour le modèle choisi dans la partie précédente sont respectivement de 98% et 95%. Devriez-vous utiliser l'un de ces nombres pour décrire comment votre modèle généralise à de nouvelles images non vues auparavant ?

5 Formule de Bayes [4 points]

1. (2 points) Une urne A contient cinq balles : une noire, deux blanches, une verte et une rose. Une urne B contient cinq cents balles : deux cents noires, cent blanches, 50 jaunes, 40 cyans, 30 ocres, 25 vertes, 25 argentées, 20 dorées et 10 pourpres. [Un cinquième des balles de A sont noires ; deux cinquièmes de celles de B sont noires.] Une des urnes est sélectionnée au hasard, l'urne A avec probabilité p et l'urne B avec probabilité $1 - p$, et une balle est tirée de l'urne sélectionnée. La balle tirée est noire. Quelle est la probabilité que l'urne choisie soit l'urne A ?

2. (2 points) Les habitants d'une île disent la vérité un tiers du temps. Ils mentent avec probabilité 2/3. Une habitante de l'île, nommée Alice, vous fait une déclaration. Vous demandez à un autre habitant de l'île, nommé Bob, « Alice a-t-elle dit la vérité ? » Bob répond « Oui. » Quelle est la probabilité qu'Alice ait dit la vérité ?

6 Propagation d'incertitude et probabilités [7 points]

1. (3 points) Prouvez l'inégalité de Chebyshev, qui s'énonce ainsi : soit $\rho(x)$ la fonction de densité de probabilité (p.d.f) d'une variable aléatoire X qui a une moyenne et une variance finie. Alors

$$\text{Proba}\left(|X - \mathbb{E}(X)| \geq l\sigma_X\right) \leq \frac{1}{l^2} \text{ avec } l \in \mathbb{R}, l > 0 \tag{4}$$

où σ_X est l'écart-type (*standard deviation*) de X. Indice : vous pouvez utiliser l'inégalité de Markov qui stipule que pour X positive et $\forall a > 0$ on a $\text{Proba}(X \geq a) \leq \frac{\mathrm{E}(X)}{a}$, sans la prouver.

2. Nous avons discuté de deux formules de propagation d'erreur qui sont souvent utilisées en physique expérimentale :

$$\sigma_G = \sum_{i=1}^{k} \left| \left. \frac{\partial G(X_1, ..., X_k)}{\partial X_i} \right|_{X_j = \mathbb{E}(X_j)\,\forall j} \right| \sigma_{X_i} \tag{7}$$

et

$$\sigma_G^2 = \sum_{i=1}^{k} \left[\left. \frac{\partial G(X_1, ..., X_k)}{\partial X_i} \right|_{X_j = \mathbb{E}(X_j)\,\forall j} \right]^2 \sigma_{X_i}^2 \tag{8}$$

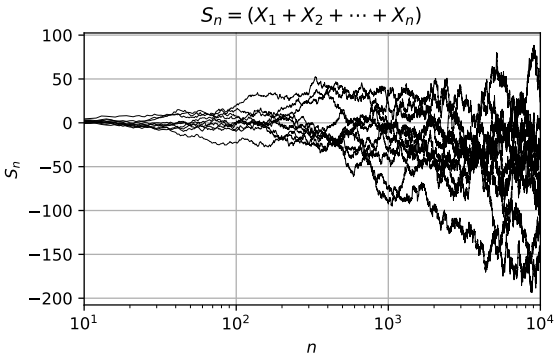
(i) (1 point) Quelle est la principale hypothèse qui doit être vérifiée pour que les deux formules ci-dessus soient valides ?

(ii) (2 points) Quelle est la principale hypothèse qui permet de décider laquelle des deux formules utiliser ?

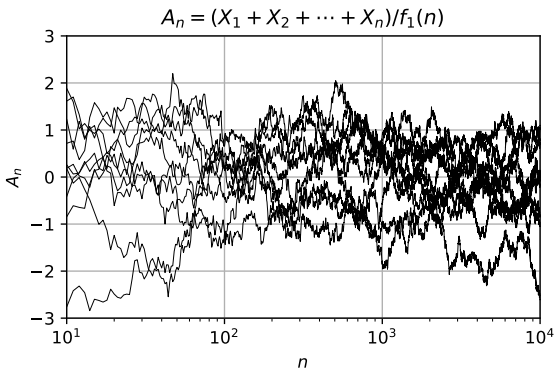
(iii) (1 point) Laquelle des deux formules donne un plus petit σ_G, en particulier quand k est grand ?

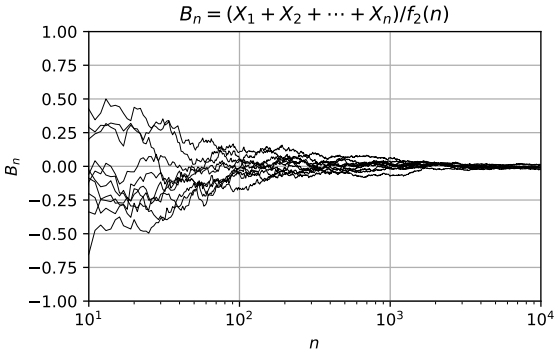
7 Théorème central limite et loi des grands nombres [6 points]

Nous échantillonnons $X_i \sim \mathcal{N}(0,1)$ i.i.d d'une distribution normale. Nous définissons $S_n = X_1 + X_2 + \cdot + X_n$. Comme vous pouvez le voir ci-dessous, c'est une marche aléatoire.



Un ami vous montre deux autres graphiques, où sont représentées les convergences de la loi des grands nombres et du théorème central limite lorsque n croît vers l'infini. Il se souvient avoir tracé des version mises à l'échelle, à savoir $A_n = S_n/f_1(n)$ et $B_n = S_n/f_2(n)$ pour chaque loi. Cependant, il a oublié quelles étaient les fonctions $f_1(n) : \mathbb{R} \to \mathbb{R}$ et $f_2(n) : \mathbb{R} \to \mathbb{R}$.





1. (4 points) Pour chaque figure de votre ami, expliquez si le comportement de convergence est lié à la loi des grands nombres ou au théorème central limite. Pour votre explication, définissez les fonctions $f_1(n)$ et $f_2(n)$ que votre ami a pu utiliser en fonction de n.

2. (2 points) Le théorème central limite stipule que la somme de nombreuses variables aléatoires est distribuée selon une loi gaussienne. Quelles sont les deux hypothèses sur les variables aléatoires pour que cela soit vrai ?

8 Maximum de vraisemblance pour la distribution de Laplace [6 points]

Soit x une variable aléatoire distribuée d'après la distribution de Laplace

$$\rho(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}} \tag{9}$$

où $\mu, b \in \mathbb{R}$, $b > 0$.

1. (3 points) Considérez que nous observons n échantillons indépendants x_i, $i = 1, \ldots, n$ de cette distribution. Utilisez la méthode du maximum de vraisemblance (*maximum likelihood method*) pour estimer la constante μ.

2. (3 points) Considérez que on observe n échantillons indépendants x_i, $i = 1, \ldots, n$ de cette distribution. Utilisez la méthode du maximum de vraisemblance (*maximum likelihood method*) pour estimer la constante b.

9 Prédiction de succès ou d'échec [12 points]

Considérons le modèle suivant pour prédire si un étudiant μ réussira ou non un examen. Supposons que nous disposons d'une base de données de n étudiants précédents, et pour chacun d'eux nous avons collecté des informations : la durée de leur révision, le nombre de cours auxquels ils ont assisté en présentiel, leurs notes des années précédentes dans d'autres cours, le nombre d'heures de sommeil avant l'examen, etc. Nous avons rassemblé toutes ces informations dans un vecteur $\vec{X}_\mu$ à d dimensions. Une matrice $X \in \mathbb{R}^{n\times d}$ contient les informations de tous les étudiants passés.

Nous savons également quels étudiants ont réussi l'examen, que nous notons $y_\mu = +1$, et lesquels ont échoué, dénoté comme $y_\mu = -1$. Nous observons ainsi la matrice $X \in \mathbb{R}^{n\times d}$ et le vecteur $y \in \mathbb{R}^n$.

Nous supposons de plus qu'il existe un vecteur de paramètres $\vec{w}^* \in \mathbb{R}^d$ (qui ne nous est pas connu) tel que la probabilité qu'un étudiant donné μ réussisse l'examen ($y_\mu = 1$) ou non ($y_\mu = -1$) est donnée par

$$P_{\text{examen}}(y_\mu|\vec{w}^*, \vec{X}_\mu) = \frac{\exp\left(y_\mu \sum_{i=1}^d X_{\mu i} w_i^*\right)}{\exp\left(\sum_{i=1}^d X_{\mu i} w_i^*\right) + \exp\left(-\sum_{i=1}^d X_{\mu i} w_i^*\right)}\,. \tag{16}$$

La réussite ou non des étudiants sont considérées comme des variables aléatoire indépendantes, toutes conditionnelles au même vecteur $\vec{w}^*$.

1. (2 points) Tracez la probabilité que l'étudiant μ réussisse l'examen en fonction du paramètre $z_\mu = y_\mu \sum_{i=1}^d X_{\mu i} w_i^*$.

2. (2 points) Écrivez la quantité qui doit être maximisée pour l'estimation du maximum de vraisemblance (*maximum likelihood estimation*) des paramètres w.

3. (2 points) Maximiser la vraisemblance revient à minimiser une fonction de perte (*loss function*), de manière similaire à ce qui a été fait en cours lorsque nous avons décrit les dérivations probabilistes de la perte des moindres carrés (*least square loss*). Écrivez la fonction de perte obtenue dans le cas présent. Indice : la fonction de perte devrait être une somme sur les étudiants μ.

4. (2 points) Écrivez la fonction de perte utilisée dans la régression logistique que nous avons vue en cours.

5. (2 points) Montrez que la fonction de perte du problème de réussite à l'examen ci-dessus est un cas particulier de la régression logistique.

6. (1 point) Quel algorithme utiliseriez-vous pour minimiser cette fonction de perte ? Donnez uniquement le nom de l'algorithme.

7. (1 point) Une fois que le minimiseur $\hat{\vec{w}}$ de la fonction de perte est obtenu, nous pouvons prédire si un nouvel étudiant pour lequel nous avons les données $\vec{X}_{\text{new}} \in \mathbb{R}^d$ réussira ou non l'examen. Écrivez le prédicteur correspondant, c'est-à-dire une fonction de $\vec{X}_{\text{new}} \to \pm 1$.

10 Échantillonnage par chaînes de Markov Monte-Carlo - triangles [8 points]

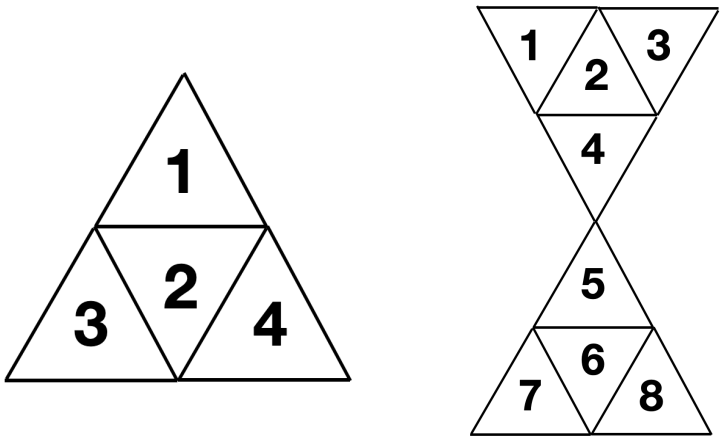


FIGURE 1 – Grilles triangulaires.

1. (2 points) Considérez la grille triangulaire figure 1 à gauche. On souhaite échantillonner les cellules 1, 2, 3 et 4 uniformément en utilisant une chaîne de Markov, où toutes les transitions entre paires des cellules sont permises. On donne les probabilités de transition suivantes : $p(2 \to 1) = p(2 \to 4) = p(2 \to 3) = 1/3$. Écrivez un exemple de chaîne de Markov qui satisfasse la condition de bilan détaillé (*detailed balance*) et qui échantillonne uniformément les cellules. Si une telle chaîne n'existe pas expliquez pourquoi.

2. (2 points) Considérez la grille triangulaire figure 1 à gauche. On souhaite échantillonner les cellules 1, 2, 3 et 4 uniformément en utilisant une chaîne de Markov. On donne les probabilités de transition suivantes : $p(2 \to 1) = p(2 \to 4) = p(2 \to 3) = 1/3$ et $p(1 \to 4) = p(4 \to 3) = p(1 \to 3) = 1/2$. Écrivez un exemple de chaîne de Markov qui satisfasse la condition de bilan détaillé (*detailed balance*) et qui échantillonne uniformément les cellules. Si une telle chaîne n'existe pas expliquez pourquoi.

3. (2 points) Considérez la grille figure 1 à droite. On souhaite échantillonner les cellules 1, 2, 3, 4, 5, 6, 7 et 8 uniformément en utilisant une chaîne de Markov. On donne que les probabilités de transition sont nulles entre les cellules qui ne sont pas voisines (c'est-à-dire nulles entre toutes sauf (1,2), (3,2), (2,4), (5,6), (6, 7) et (6, 8)). Écrivez un exemple de chaîne de Markov qui satisfasse la condition de bilan détaillé (*detailed balance*) et qui échantillonne uniformément les cellules. Si une telle chaîne n'existe pas expliquez pourquoi.

4. (2 points) Considérez la grille figure 1 à droite. On souhaite échantillonner les cellules 1, 2, 3, 4, 5, 6, 7 et 8 uniformément en utilisant une chaîne de Markov. On donne les probabilités de transition suivantes : $p(2 \to 1) \geq 1/4$, $p(2 \to 3) \geq 1/4$, $p(2 \to 4) \geq 1/4$. Écrivez un exemple de chaîne de Markov qui satisfasse la condition de bilan détaillé (*detailed balance*) et qui échantillonne uniformément les cellules. Si une telle chaîne n'existe pas expliquez pourquoi. Dans ce cas toutes les transitions sont permises, non seulement celles entre voisins (e.g. $p(6 \to 3)$ peut être positive).

11 Méthode de Monte-Carlo par chaînes de Markov [6 points]

1. (3 points) Soit une chaîne de Markov sur un espace d'états X définie par la probabilité de transition $p(a \to b)$ d'aller d'un état $a \in X$ à un état $b \in X$ à chaque pas de temps. Soit $\pi(a)$ une distribution de probabilité sur X qui vérifie

$$\pi(a)p(a \to b) = \pi(b)p(b \to a) \, \forall a, b \in X \, . \tag{19}$$

Comment cette condition s'appelle-t-elle ? Prouvez que π est la distribution stationnaire de la chaîne, c'est-à-dire

$$\sum_a \pi(a)p(a \to b) = \pi(b) \, . \tag{20}$$

2. (3 points) Lequel des algorithmes suivants utiliseriez-vous pour échantillonner uniformément l'intérieur du disque de dimension d défini par $D_d = \{x \in \mathbb{R}^d \text{ tel que } ||x|| \leq 1\}$?

   (a) Échantillonnage direct : génère uniformément des points dans $[-1, 1]^d$ et rejette ceux hors de D_d.

   (b) MCMC : suit une marche aléatoire dans D_d et à chaque pas de temps rejette le mouvement s'il amène hors de D_d gardant la position actuelle comme échantillon.

   Justifiez brièvement en distinguant le cas de basse dimension $d \lessapprox 10$ de celui de haute dimension $d \gtrapprox 10$.