Due Date : March 27th 23 :00, 2023

Instructions

- *Pour toutes les questions, montrez votre travail !*
- *Utiliser un système de rédaction de documents tel que LaTeX.*
- *Soumettez vos réponses par voie électronique via le système de notation du cours*
- *Les TAs pour ce devoir sont :* ***Dinghuai Zhang*** *(IFT6135B) et* ***Ghait Boukachab*** *(IFT6135A).*

Question 1 (8-9-8)**.** Cette question est à propos des fonctions d'activation et des problèmes de disparition/explosion des gradients (*vanishing/exploding gradients*) dans les réseaux neuronaux récurrents (RNNs). Soit $\sigma : \mathbb{R} \to \mathbb{R}$ une fonction d'activation. Quand le paramètre est un vecteur, on applique σ élément par élément (*element-wise*). Considérez l'unité récurrente suivante :

$$\boldsymbol{h}_t = \boldsymbol{W}\sigma(\boldsymbol{h}_{t-1}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}$$

1.1 Montrez que l'application de la fonction d'activation de cette manière donne une réccurence qui est équivalente à la façon usuelle de l'appliquer : $\boldsymbol{g}_t = \sigma(\boldsymbol{W}\boldsymbol{g}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b})$ (c.-à-d. exprimez $\boldsymbol{g}_t$ en fonction de $\boldsymbol{h}_t$). Plus formellement, utilisez le principe d'induction mathématique. Dans cette question, vous n'avez qu'à prouver le pas d'induction ; assumant que votre expression est vraie au temps $t-1$.

Réponse : On doit prouver que c'est valide au temps $t-1$. On pose $\boldsymbol{g}_{t-1} = \sigma(\boldsymbol{h}_{t-1})$ pour prouver que $\boldsymbol{g}_t = \sigma(\boldsymbol{h}_t)$. Puisque $\boldsymbol{h}_t = \boldsymbol{W}\sigma(\boldsymbol{h}_{t-1}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}$ alors, $\sigma(\boldsymbol{h}_t) = \sigma(\boldsymbol{W}\sigma(\boldsymbol{h}_{t-1}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b})$. Ainsi,

$$\begin{aligned}
\boldsymbol{g}_t &= \sigma(\boldsymbol{W}\boldsymbol{g}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}), && (1)\\
&= \sigma(\boldsymbol{W}\sigma(\boldsymbol{h}_{t-1}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}), && (2)\\
\boldsymbol{g}_t &= \sigma(\boldsymbol{h}_t). && (3)
\end{aligned}$$

Avec $t = 0$, $\boldsymbol{g}_0 = \sigma(\boldsymbol{h}_0)$ et donc les deux manières sont équivalentes.

*1.2 On note par $||\boldsymbol{A}||$ la norme L_2 d'une matrice $\boldsymbol{A}$ ($||\boldsymbol{A}|| := \max_{\boldsymbol{x}:||\boldsymbol{x}||=1} ||\boldsymbol{A}\boldsymbol{x}||$). Supposons que $\sigma(x)$ a une dérivée bornée, c.-à-d. $|\sigma'(x)| \leq \gamma$ pour un certain $\gamma > 0$ et pour tout x. On dénote par $\lambda_1(\cdot)$ la plus grande valeur propre d'une matrice symétrique. Montrez que si la plus grande valeur propre des poids est majorée par $\frac{\delta^2}{\gamma^2}$ pour un certain $0 \leq \delta < 1$, alors les gradients de l'état caché (*hidden state*) vont disparaître au cours du temps, i.e.

$$\lambda_1(\boldsymbol{W}^\top \boldsymbol{W}) \leq \frac{\delta^2}{\gamma^2} \quad \Longrightarrow \quad \left|\left|\frac{\partial \boldsymbol{g}_T}{\partial \boldsymbol{g}_0}\right|\right| \to 0 \text{ as } T \to \infty$$

Vous pouvez utiliser les propriétés suivantes de l'opérateur de norme L_2 :

$$||\boldsymbol{A}\boldsymbol{B}|| \leq ||\boldsymbol{A}||\,||\boldsymbol{B}|| \qquad \text{and} \qquad ||\boldsymbol{A}|| = \sqrt{\lambda_1(\boldsymbol{A}^\top \boldsymbol{A})}$$

Réponse : Avec $\boldsymbol{g}_t = \sigma_t(\boldsymbol{W}\boldsymbol{g}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b})$, évaluons $\frac{\partial \boldsymbol{g}_T}{\partial \boldsymbol{g}_0}$

$$\frac{\partial \boldsymbol{g}_T}{\partial \boldsymbol{g}_0} = \frac{\partial \sigma_T}{\partial \boldsymbol{g}_{T-1}} \frac{\partial \sigma_{T-1}}{\partial \boldsymbol{g}_{T-2}} \cdots \frac{\partial \sigma_1}{\partial \boldsymbol{g}_0}, \tag{4}$$

$$= \prod_{t=0}^{T-1} \frac{\partial \sigma_{t+1}}{\partial \boldsymbol{g}_t}. \tag{5}$$

Évaluons $\frac{\partial \sigma_{t+1}}{\partial \boldsymbol{g}_t}$ avec $\boldsymbol{g}_t = \sigma_t(\boldsymbol{W}\boldsymbol{g}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b})$,

$$\frac{\partial \sigma_{t+1}}{\partial \boldsymbol{g}_t} = \text{diag}(\sigma'_{t+1})\boldsymbol{W}. \tag{6}$$

En substituant le résultat précédent dans (4),

$$\frac{\partial \boldsymbol{g}_T}{\partial \boldsymbol{g}_0} = \prod_{t=0}^{T-1} \text{diag}(\sigma'_{t+1})\boldsymbol{W}. \tag{7}$$

Ainsi,

$$\left|\left|\frac{\partial \boldsymbol{g}_T}{\partial \boldsymbol{g}_0}\right|\right| = \left|\left|\prod_{t=0}^{T-1} \text{diag}(\sigma'_{t+1})\boldsymbol{W}\right|\right|, \tag{8}$$

$$= \prod_{t=0}^{T-1} \left|\left|\text{diag}(\sigma'_{t+1})\boldsymbol{W}\right|\right|, \tag{9}$$

$$\leq \prod_{t=0}^{T-1} ||\text{diag}(\sigma'_{t+1})||||\boldsymbol{W}||, \tag{10}$$

$$\leq \prod_{t=0}^{T-1} |\gamma|\sqrt{\lambda_1} = \prod_{t=0}^{T-1} |\gamma||\frac{\delta}{\gamma}|, \tag{11}$$

$$\left|\left|\frac{\partial \boldsymbol{g}_T}{\partial \boldsymbol{g}_0}\right|\right| \leq \prod_{t=0}^{T-1} |\delta| = |\delta|^T. \tag{12}$$

Puisque $0 \leq \delta < 1$,

$$\lim_{T\to\infty} \left|\left|\frac{\partial \boldsymbol{g}_T}{\partial \boldsymbol{g}_0}\right|\right| = \lim_{T\to\infty} |\delta|^T = 0. \tag{13}$$

La norme du gradient $\left|\left|\frac{\partial \boldsymbol{g}_T}{\partial \boldsymbol{g}_0}\right|\right|$ tend vers 0 lorsque $T \to \infty$. C.Q.F.D.

1.3 Que pensez-vous qu'il se passerait aux gradients de l'état caché si la condition de la question précédente était inversée, c.-à-d. que la plus grande valeur propre était plus grande que $\frac{\delta^2}{\gamma^2}$? Cette condition est-elle *nécessaire* ou *suffisante* pour que le gradient explose ? (Répondez en une ou deux phrases).

Réponse : La valeur propre la plus grande du gradients de l'état caché doit avoir une plus grande que $\frac{\delta^2}{\gamma^2}$ i.e. c'est *nécessaire*, cependant, ce n'est pas *suffisant* puisque la condition est une borne inférieure.

Question 2 (8-8-9)**.** Dans cette question, vous allez démontrer qu'une estimation du premier moment du gradient, en utilisant une moyenne courante exponentielle (*exponential moving average*), est équivalente à l'utilisation de momentum (*SGD with momentum*), et est biaisée par un facteur multiplicatif. Le but de cette question est pour vous de considérer la relation entre différents schémas d'optimisation, et de pratiquer l'étude de l'effet (en termes de biais/variance) de l'*estimation* d'une quantité.

Soit $\boldsymbol{g}_t$ un échantillon non-biaisé (*unbiased sample*) du gradient au temps t et $\Delta\boldsymbol{\theta}_t$ la mise-à-jour à faire. Initialisez $\boldsymbol{v}_0$ au vecteur de zeros.

1. Pour $t \geq 1$, considérez les règles de mise-à-jour suivantes :
   - SGD avec momentum :
   $$\boldsymbol{v}_t = \alpha\boldsymbol{v}_{t-1} + \epsilon\boldsymbol{g}_t \qquad \Delta\boldsymbol{\theta}_t = -\boldsymbol{v}_t$$
   où $\epsilon > 0$ et $\alpha \in (0, 1)$.
   - SGD avec moyenne courante de $\boldsymbol{g}_t$:
   $$\boldsymbol{v}_t = \beta\boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t \qquad \Delta\boldsymbol{\theta}_t = -\delta\boldsymbol{v}_t$$
   où $\beta \in (0, 1)$ et $\delta > 0$.

   Exprimez les deux règles de mise-à-jour de façon récursive ($\Delta\boldsymbol{\theta}_t$ en fonction de $\Delta\boldsymbol{\theta}_{t-1}$). Montrez que ces deux règles de mise-à-jour sont équivalentes ; c.-à-d., exprimez (α, ϵ) en fonction de (β, δ).

   Réponse :

   SGD avec momentum : avec $\Delta\boldsymbol{\theta}_{t-1} = -\boldsymbol{v}_{t-1}$,
   $$\begin{aligned}
   \Delta\boldsymbol{\theta}_t &= -\boldsymbol{v}_t, && (14)\\
   &= -(\alpha\boldsymbol{v}_{t-1} + \epsilon\boldsymbol{g}_t), && (15)\\
   &= -(\alpha(-\Delta\boldsymbol{\theta}_{t-1}) + \epsilon\boldsymbol{g}_t), && (16)\\
   \Delta\boldsymbol{\theta}_t &= \alpha\Delta\boldsymbol{\theta}_{t-1} - \epsilon\boldsymbol{g}_t. && (17)
   \end{aligned}$$

   SGD avec moyenne mobile de $\boldsymbol{g}_t$: avec $\Delta\boldsymbol{\theta}_{t-1} = -\delta\boldsymbol{v}_{t-1}$,
   $$\begin{aligned}
   \Delta\boldsymbol{\theta}_t &= -\delta\boldsymbol{v}_t, && (18)\\
   &= -\delta(\beta\boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t), && (19)\\
   &= -\delta(\beta(-\frac{1}{\delta}\Delta\boldsymbol{\theta}_{t-1}) + (1-\beta)\boldsymbol{g}_t), && (20)\\
   \Delta\boldsymbol{\theta}_t &= \beta\Delta\boldsymbol{\theta}_{t-1} - \delta(1-\beta)\boldsymbol{g}_t. && (21)
   \end{aligned}$$

   Terme à terme, $\beta\Delta\boldsymbol{\theta}_{t-1} = \alpha\Delta\boldsymbol{\theta}_{t-1}$ et $\epsilon\boldsymbol{g}_t = \delta(1-\beta)\boldsymbol{g}_t$ implique $\beta = \alpha$ et $\epsilon = \delta(1-\beta)$ respectivement.

2. Déroulez la règle de mise-à-jour de la méthode avec la moyenne courante, c.-à-d., exprimez $\boldsymbol{v}_t$ comme une combinaison linéaire de $\boldsymbol{g}_i$'s $(1 \leq i \leq t)$.

   Réponse : La règle de mise-à-jour déroulé pour SGD avec moyenne mobile :
   $$\begin{aligned}
   \boldsymbol{v}_t &= \beta\boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t, && (22)\\
   &= \beta(\beta\boldsymbol{v}_{t-2} + (1-\beta)\boldsymbol{g}_{t-1}) + (1-\beta)\boldsymbol{g}_t, && (23)\\
   &= \beta(\beta(\beta\boldsymbol{v}_{t-3} + (1-\beta)\boldsymbol{g}_{t-2}) + (1-\beta)\boldsymbol{g}_{t-1}) + (1-\beta)\boldsymbol{g}_t, && (24)\\
   &\vdots, && (25)\\
   \boldsymbol{v}_t &= \beta^t\boldsymbol{v}_0 + \sum_{i=1}^{t}\beta^{t-i}(1-\beta)\boldsymbol{g}_i. && (26)
   \end{aligned}$$

SGD avec momentum, $\beta = \alpha$:

$$\boldsymbol{v}_t = \alpha^t \boldsymbol{v}_0 + \sum_{i=1}^{t} \alpha^{t-i}(1-\alpha)\boldsymbol{g}_i. \tag{27}$$

3. Supposons que $\boldsymbol{g}_t$ a une distribution stationnaire indépendante de t. Déterminez un estimateur de $\mathbb{E}[\boldsymbol{g}_t]$ en utilisant $\mathbb{E}[\boldsymbol{v}_t]$.
Réponse : Avec $\boldsymbol{v}_t = \beta^t \boldsymbol{v}_0 + \sum_{i=1}^{t} \beta^{t-i}(1-\beta)\boldsymbol{g}_i$,

$$\mathbb{E}[\boldsymbol{v}_t] = \beta^t \mathbb{E}[\boldsymbol{v}_0] + \sum_{i=1}^{t} \beta^{t-i}(1-\beta)\mathbb{E}[\boldsymbol{g}_i], \tag{28}$$

$$= \beta^t \mathbb{E}[\boldsymbol{v}_0] + \mathbb{E}[\boldsymbol{g}_t] \sum_{i=1}^{t} \beta^{t-i}(1-\beta). \tag{29}$$

où on a supposé que $\boldsymbol{g}_t$ est indépendante de t.
En isolant pour $\mathbb{E}[\boldsymbol{g}_t]$,

$$\mathbb{E}[\boldsymbol{g}_t] = \frac{\mathbb{E}[\boldsymbol{v}_t] - \beta^t \mathbb{E}[\boldsymbol{v}_0]}{\sum_{i=1}^{t} \beta^{t-i}(1-\beta)}. \tag{30}$$

Question 3 (8-8-9)**.** L'équation suivante est le développement de Taylor du second ordre d'une fonction f au point x_0 :

$$\hat{f}_{x_0}(x) = f(x_0) + (x - x_0)^T g + \frac{1}{2}(x - x_0)^T H (x - x_0), \tag{31}$$

avec $g = \frac{\partial f}{\partial x}(x_0)$ et $H = \frac{\partial^2 f}{\partial^2 x}(x_0)$. ici $f(x) \in \mathbb{R}$, $x, x_0, g \in \mathbb{R}^n, H \in \mathbb{R}^{n \times n}$.

3.1 Supposons que nous commencions à partir de x_0 et que nous effectuions une descente de gradient en une étape avec le gradient g et un taux d'apprentissage égal à ϵ, quelle est la valeur de $\hat{f}_{x_0}(\cdot)$ après la mise à jour ?
Réponse : Pour trouver x_1 à partir de x_0 dans une descente de gradient, on exécute $x_1 = x_0 - \epsilon g$,

$$\hat{f}_{x_0}(x_1) = f(x_0) + (x_0 - \epsilon g - x_0)^T g + \frac{1}{2}(x_0 - \epsilon g - x_0)^T H (x_0 - \epsilon g - x_0), \tag{32}$$

$$= f(x_0) + (-\epsilon g)^T g + \frac{1}{2}(-\epsilon g)^T H (-\epsilon g), \tag{33}$$

$$= f(x_0) + (-\epsilon g)^T g + \frac{1}{2}(-\epsilon g)^T H (-\epsilon g), \tag{34}$$

$$\hat{f}_{x_0}(x_1) = f(x_0) - \epsilon g^T g + \frac{1}{2}\epsilon^2 g^T H g. \tag{35}$$

3.2 Analysez les termes que vous avez obtenus à la question précédente. Expliquez sous quelles conditions de ϵ la descente par gradient fonctionnerait (c'est-à-dire qu'elle réduirait la valeur de la fonction cible $\hat{f}_{x_0}(\cdot)$).
Réponse : Le terme $g^T g$ pointe dans la direction opposé minimisant $\hat{f}_{x_0}(\cdot)$, et $g^T H g$ pointe dans la direction minimisant $\hat{f}_{x_0}(\cdot)$ (d'où la soustraction et l'addition). ϵ contrôle la *vitesse* du pas du gradient et de l'Hessienne. Le terme $f(x_0)$ représente la valeur initiale de la fonction avant la décente de gradient.

3.3 En fixant le gradient de $\hat{f}_{x_0}(\cdot)$ à zéro. Pourriez-vous dériver un nouvel algorithme d'optimisation basé sur ceci ?

Réponse : Oui, ce serait un algorithme d'optimisation du second ordre. L'algorithme est donné par $\nabla_x \hat{f}_{x_0}(x) = 0$,

$$\nabla_x \hat{f}_{x_0}(x) = g + H(x - x_0) = 0, \tag{36}$$
$$H^{-1}g + x - x_0 = 0, \tag{37}$$
$$x = x_0 - H^{-1}g. \tag{38}$$

Cet algorithme inclut de l'information sur la courbure de $\hat{f}_{x_0}(x)$. Généralement, on ajoute ϵ pour contrôler la décente.

Question 4 (8-9-8)**.** La normalisation des poids (WN) est une technique de reparamétrisation inspirée de la normalisation par lots (BN), visant à améliorer le conditionnement du gradient. Au lieu d'avoir un paramètre de poids régulier (noté w) pour chaque neurone, c'est-à-dire $y = \sigma(w^T x + b)$, nous découplons le vecteur de poids en deux termes :

$$w = \frac{g}{\|u\|} u \tag{39}$$

où $g \in \mathbf{R}$ est un facteur d'échelle et u est normalisé par la valeur euclidienne $\|u\|$. Cette méthode a des effets similaires à ceux de l'implémentation de BN, mais présente un coût de calcul plus faible.

4.1 Considérons le modèle le plus simple, où nous n'avons qu'une seule couche de sortie conditionnée par une caractéristique d'entrée x. Supposons que x est blanchi pour être distribué indépendamment avec une moyenne de zéro et une variance unitaire. Une opération BN standard est définie par l'équation (8.35) du livre sur l'apprentissage profond. Montrez que dans ce cas simple, WN est équivalent à BN (en ignorant les termes d'échelle et de décalage appris pour BN et WN) qui normalise la caractéristique transformée linéairement $w^T x + b$.

Réponse : Avec BN en ignorant le terme de décalage,on a que $y = w^T \left(\frac{x-\mu}{\sigma}\right) + b$ et avec WN en ignorant le terme de échelle, on a que $y = \left(\frac{v}{||v||}\right)^T x + b$. En les égalisant, on trouve,

$$w^T \left(\frac{x-\mu}{\sigma}\right) + b = \left(\frac{v}{||v||}\right)^T x + b, \tag{40}$$
$$w^T \left(\frac{x-\mu}{\sigma}\right) = \left(\frac{v}{||v||}\right)^T x. \tag{41}$$

Puisque la moyenne de x est zéro, $\mu = 0$,

$$w^T \left(\frac{x}{\sigma}\right) = \left(\frac{v}{||v||}\right)^T x. \tag{42}$$

Avec $\sigma = ||v||$,

$$w^T x = v^T x. \tag{43}$$

Puisque $w^T x = v^T x$ alors dans ce cas simple WN est équivalent à BN. C.Q.F.D.

4.2 Montrez que le gradient d'une fonction de perte L par rapport aux nouveaux paramètres u peut être exprimé sous la forme $sW^* \cdot \nabla_w L$, où s est un scalaire et W^* est la matrice de projection du complément orthogonal. Remarque : W^* projette le gradient loin de la direction de w, qui est généralement (empiriquement) proche d'un vecteur propre dominant de la covariance du gradient. Cela permet de conditionner le paysage sur lequel nous voulons optimiser.

Réponse : Le gradient de L est donné par,

$$\nabla_u L(u) = \nabla_u L\left(w\left(g\frac{u}{||u||}\right)\right), \tag{44}$$

$$= \nabla_u w \nabla_w L, \tag{45}$$

puisque w est fonction de u et par la règle de dérivé en chaîne.
La partie gradient $\nabla_u w$ est donné par,

$$\nabla_u w = \nabla_u\left(g\frac{u}{||u||}\right), \tag{46}$$

$$= g\nabla_u\left(\frac{u}{||u||}\right). \tag{47}$$

Par la règle de dérivé de fractions,

$$\nabla_u w = g\left(\frac{(\nabla_u u)||u|| - u^T(\nabla_u||u||)}{||u||^2}\right). \tag{48}$$

Avec,

$$\nabla_u||u|| = \nabla_u\sqrt{u\cdot u}, \tag{49}$$

$$= \frac{u}{\sqrt{u\cdot u}}. \tag{50}$$

$$= \frac{u}{||u||}. \tag{51}$$

Alors,

$$\nabla_u w = g\left(\frac{(\nabla_u u)||u|| - u^T(\nabla_u||u||)}{||u||^2}\right), \tag{52}$$

$$= g\left(\frac{(I)||u|| - u^T\left(\frac{u}{||u||}\right)}{||u||^2}\right), \tag{53}$$

$$\nabla_u w = \frac{g}{||u||}\left(I - \frac{u^T u}{||u||^2}\right). \tag{54}$$

Avec $s = \frac{g}{||u||}$ et $W^* = \left(I - \frac{u^T u}{||u||^2}\right)$, le gradient $\nabla_u L(u)$ devient,

$$\nabla_u L(u) = \nabla_u w \nabla_w L, \tag{55}$$

$$= \frac{g}{||u||}\left(I - \frac{u^T u}{||u||^2}\right)\nabla_w L, \tag{56}$$

$$\nabla_u L(u) = sW^*\nabla_w L. \tag{57}$$

C.Q.F.D.

4.3 Les chercheurs ont découvert que pour la normalisation du poids, la norme des paramètres ($\|u\|$) ne cesse d'augmenter. Montrez que $\|u\|$ devient égal ou supérieur après une étape de mise à jour du gradient à l'aide du théorème de Pythogore et de l'équation de mise à jour du gradient de la question précédente.

Réponse : La décente est donné par,

$$u' = u - \epsilon \nabla_u L. \tag{58}$$

Sa norme,

$$||u'||^2 = ||u - \epsilon \nabla_u L||^2, \tag{59}$$
$$= ||u||^2 + ||\epsilon \nabla_u L||^2 - 2\epsilon u \nabla_u L. \tag{60}$$

Puisque $u \perp \nabla_u L$ i.e. $u \perp W^*$,

$$||u'||^2 = ||u||^2 + ||\epsilon \nabla_u L||^2, \tag{61}$$
$$= ||u||^2 \left(1 + \frac{||\epsilon \nabla_u L||^2}{||u||^2}\right). \tag{62}$$

Avec $\frac{||\epsilon \nabla_u L||^2}{||u||^2} = c^2$,

$$||u'||^2 = ||u||^2 \left(1 + c^2\right). \tag{63}$$

Ainsi,

$$||u'|| = \sqrt{||u||^2 (1 + c^2)}, \tag{64}$$
$$||u'|| = ||u|| \sqrt{1 + c^2}. \tag{65}$$

Puisque $\sqrt{1 + c^2} > 1$ alors $||u'||$ ne peut qu'augmenter. C.Q.F.D.