INF8225-TP1

Arnaud Emmanuel Robert

Février 2020

1 Partie 1

a) L'évènement H est entièrement determiné par P (Pluie) et A (Arroseur). Afin d'obtenir la probabilité $Pr(H = 1)$, on conditione sur P et A selon la loi de probabilité totale:

$$Pr(H = 1) = \sum_{i=0}^{1}\sum_{j=0}^{1} Pr(H = 1|P = i, A = j) * Pr(P = i, A = j) = 0.272 \quad (1)$$

b) Lorsque seul W est observé, il existe une relation de dépendance condionnelle entre H et W. On applique à nouveau la loi de probabilité totale sur $Pr(H = 1|W = 1)$ et on s'aperçoit que les termes de la forme $Pr(H = 1|P = i, A = j, W = 1)$ peuvent être simplifiés en $Pr(H = 1|P = i, A = j)$ pusique qu'en cas d'observation de P et A, il n'existe plus de dépendance entre W et H. On utilise également la formule de Bayes pour obtenir $Pr(P = i|W = 1)$. On obtient alors:

$$\begin{aligned}
Pr(H = 1|W = 1) &= \sum_{i=0}^{1}\sum_{j=0}^{1} Pr(H = 1|P = i, A = j, W = 1) * Pr(P = i, A = j|W = 1) \\
&= \sum_{i=0}^{1}\sum_{j=0}^{1} Pr(H = 1|P = i, A = j) * Pr(P = i|W = 1) * Pr(A = j) \\
&= \sum_{i=0}^{1}\sum_{j=0}^{1} Pr(H = 1|P = i, A = j) * \frac{Pr(W = 1|P = i)Pr(P = i)}{Pr(W = 1)} * Pr(A = j) \\
&= 0.5955
\end{aligned}$$

c) On peut calculer cette valeur de manière similaire au calcul en b). Il est également possible de procéder en remarquant que si W=0, alors il ne peut y avoir eu de pluie et donc P=0 (car Pr(W=1—P=1) = 1). On a alors:

$$Pr(H = 1|W = 0) = Pr(H = 1|P = 0) = 0.09 \quad (2)$$

d) Comme P a été observé, les événements H et W sont conditionnellement indépendants. De fait:

$$Pr(H = 1|P = 0, W = 1) = Pr(H = 1|P = 0) = 0.09 \tag{3}$$

e) On procède de manière similaire à b) en utilisant la loi de probabilité totale et en simplifiant via les indépendances. On peut également partir de $Pr(H = 1|W = 1)$ et appliquer la formule de Bayes.

$$\begin{aligned} Pr(W = 1|H = 1) &= \sum_{i=0}^{1} Pr(W = 1|P = i, H = 1) * Pr(P = i|H = 1) \\ &= \sum_{i=0}^{1} Pr(W = 1|P = i) * \frac{Pr(H = 1|P = i)Pr(P = i)}{Pr(H = 1)} \\ &= 0.788 \end{aligned}$$

f) Comme W est conditionnellement indépendant de A, on obtient le même résultat que la question précédente.

$$Pr(W = 1|H = 1, A = 1) = Pr(W = 1|H = 1) = 0.788 \tag{4}$$

2 Partie 2

2.1 Questions

2.1.1 Question a)

La Figure 1 présente les courbes d'apprentissage d'entraînement, validation et test pour 12 cas d'entraînements avec différents paramètres pour le taux d'apprentissage et la taille des minibatch. Sur chaque image on obtient également la valeur de "accuracy" sur l'ensemble de test, ainsi que la valeur final de validation loss.

Toutes les courbes sont présentées en échelle logarithmique et partagent les mêmes échelles sur les deux axes afin de pouvoir facilement comparer les cas entre eux. Certaines valeurs très proches de 0 (et donc très négatives en échelle logarithmique) pour les courbes d'entraînement ont été omises afin de préserver une échelle d'ordonnée qui permettent la comparaison entre les modèles. De fait, certaines des courbe d'entraînement sont parfaitement lisses vers les dernières epochs.

Sur la Figure 1 on s'aperçoit que le fait de combiner un taux d'apprentisage élevé avec une valeur de batch faible donne lieu à une convergence chaotique. En effet, la matrice de paramètres est alors grandement (du fait du taux d'apprentissage) et régulièrement (du fait de la taille des minibatch) modifiée lors de chaque mise à jour. A l'inverse, un taux trop faible pour une taille de batch trop grande ne permet pas de s'approcher d'une valeur de validation idéale. La matrice de paramètre n'est en fait que trop peu régulièrement modifiée (car les batch sont très grand) et l'ampleur de la mise à jour est trop faible (car le taux est faible).

Il convient alors de trouver un compromis entre ces deux valeurs. Dans notre cas, un taux d'apprentissage de 0.01 et une taille de minibatch de 20 semble donner les meilleurs résultats.

2.1.2 Question b)

La Figure 2 permet de comparer les performances d'un modèle utilisant ADAM et d'un autre utilisaut des paramètre fixes optimaux. Les échelle en ordonnée n'est cette fois pas logarithmique contrairement à la question a).

En utilisant ADAM, on s'aperçoit que le modèle va overfitter plus rapidement. De fait, ADAM permet à l'entrainement de converger plus rapidement. Cependant , les bienfaits de l'utilisation de ADAM sur ce problème ne sont pas parfaitements visibles. En effet, le moment où la courbe de validation devient plate semble être atteint relativement au même moment.

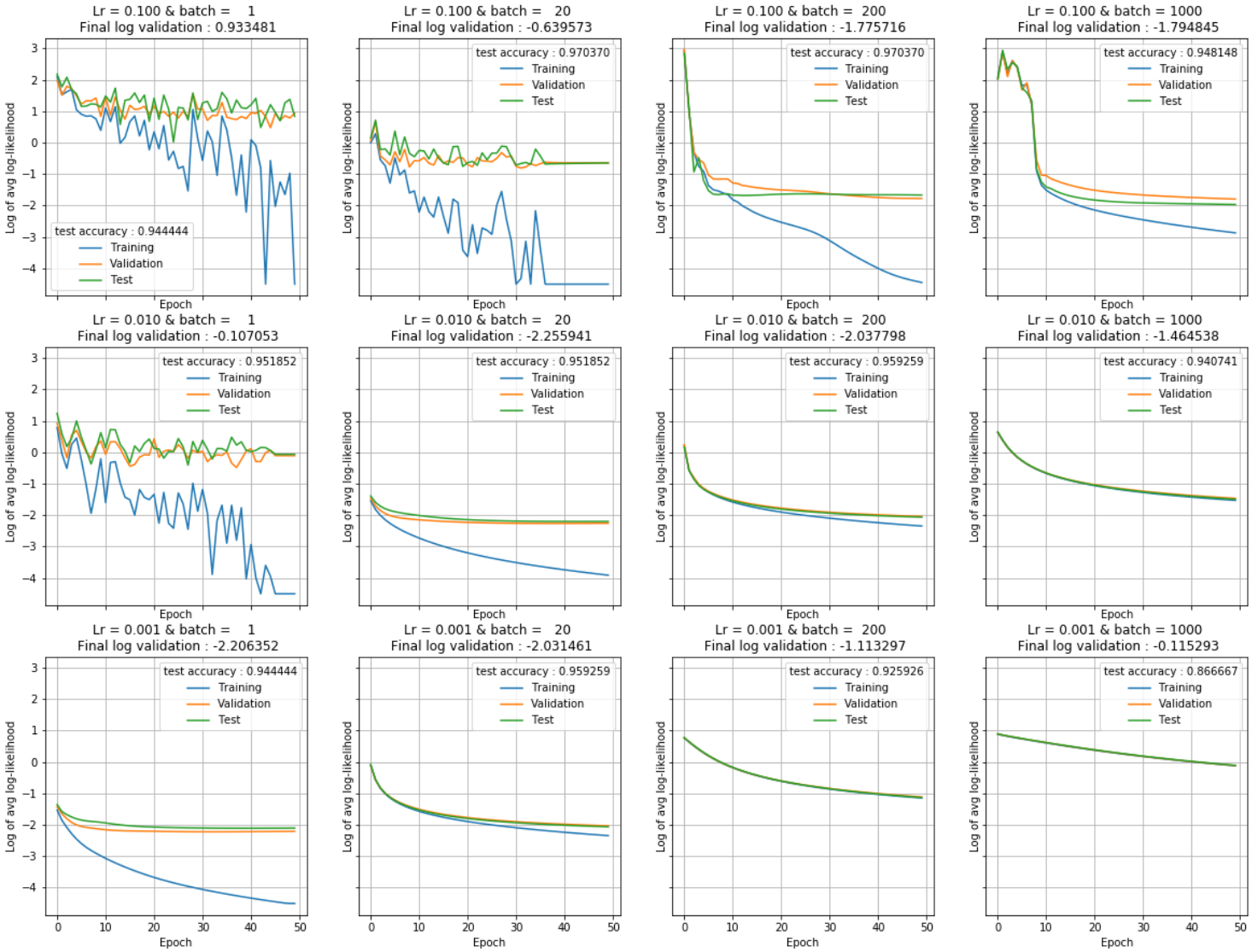

Figure 1: Validation, training and test average log-likelihoods (log scale) with different parameters for learning rate and batch size.

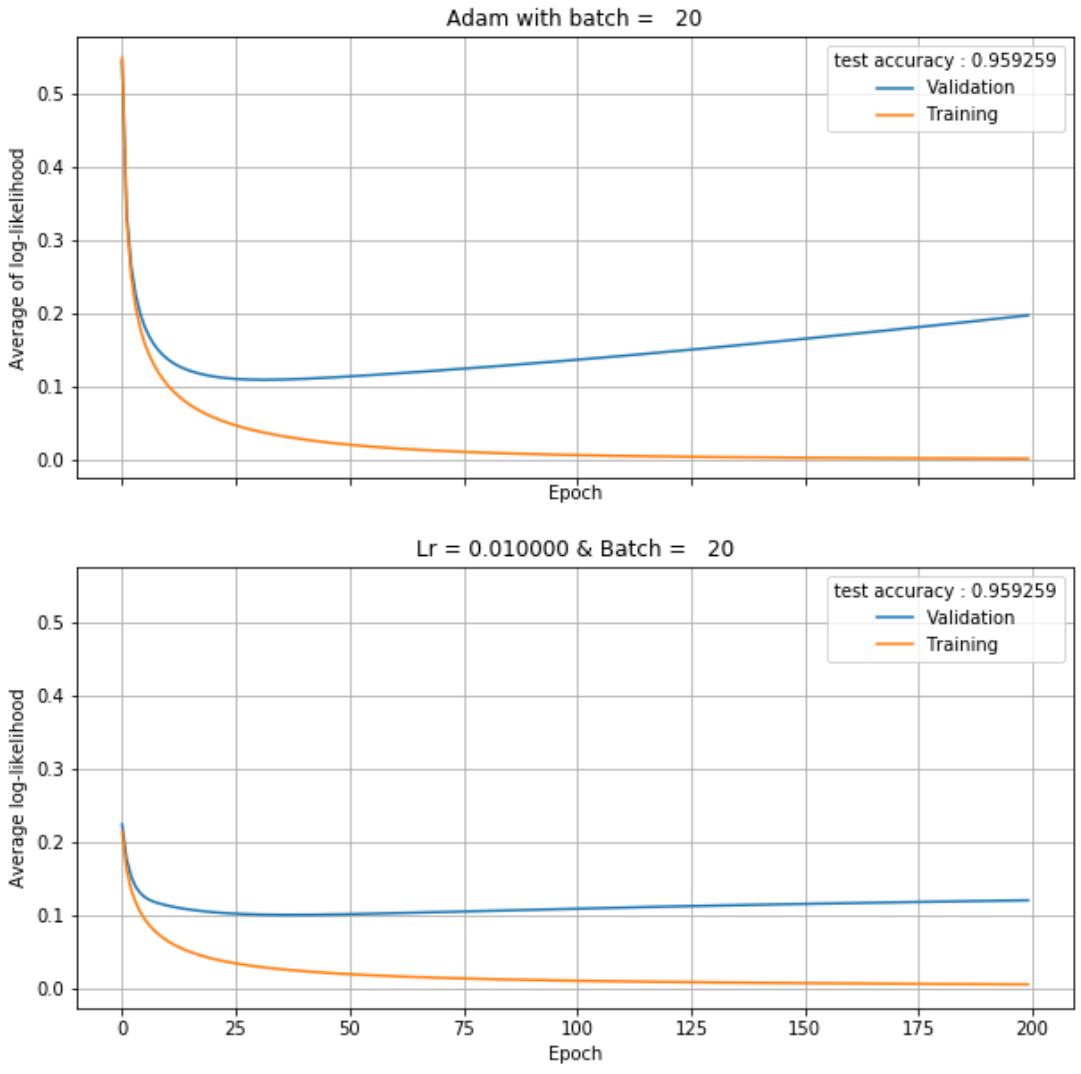

Figure 2: Comparison of losses with and without using ADAM