Nom de l'étudiant : _____

FACULTE DES ARTS ET DES SCIENCES
DEPARTEMENT D'INFORMATIQUE ET DE RECHERCHE OPERATIONNELLE

**Algorithmes d'apprentissage IFT6266 H13**

Examen final - 6 mai 2013

DIRECTIVES PÉDAGOGIQUES :
- Documentation permise. Ordinateur non permis.
- Répondre directement sur le questionnaire. Vous pouvez utiliser l'arrière des pages aussi si vous en avez besoin.
- Soyez brefs et précis dans vos réponses.
- **Si vous manquez de temps : l'important est de montrer que vous avez compris le problème, plutôt que les détails de la réponse.**
- Échanger des informations lors d'un examen (ou autres formes de tricherie) est du **plagiat**, qui est passible de sanctions allant jusqu'à l'exclusion du programme.
- Suggestion : lisez tout en premier et commencez par les questions qui vous semblent les plus faciles.

1. Comment est-ce qu'un réseau de neurones récurrent peut être utilisé pour *représenter une distribution de probabilité* sur des séquences ? Une fois entraîné, comment peut-il être utilisé pour *générer* une séquence de la distribution apprise ? Dans votre réponse, considérez un détail important : comment arrêter la génération, i.e., décider que la séquence générée est terminée.

2. Pour un réseau de neurones, quelle fonction de perte serait raisonnable pour des cibles binaires, si on interprète la perte comme moins une log-vraisemblance ? Si on voit la sortie d'un réseau de neurones comme les paramètres d'une distribution pour la variable de sortie, quelle est cette distribution ? Et si la cible est une variable réelle, avec une erreur quadratique, i.e., quelle est l'interprétation en terme de log-vraisemblance de l'erreur quadratique ? Y a-t-il aussi une interprétation probabiliste de la régularisation L1 ou L2 des paramètres ?

3. Qu'est-ce que l'arrêt prématuré (*early stopping*) ? Comment peut-on l'interpéter comme une forme de régularisation ? (une réponse qualitative suffit). Peut-on appliquer l'arrêt prématuré pendant l'entraînement des RBMs (par PCD ou CD) ? Pourquoi ?

4. Expliquer le principe d'une recherche sur une grille (*grid search*) pour les hyper-paramètres ? et une recherche aléatoire (*random search*) ? Pourquoi est-ce qu'une recherche aléatoire est souvent beaucoup plus efficace ?

5. Quel est le principe derrière l'application de l'algorithme de rétro-propagation du gradient (back-propagation) dans le cas de systèmes tels que les réseaux recurrents ? Un dessin et quelques explications sur la procédure générale suffirait. Appliquez ce principe pour dériver les équations du calcul de la rétro-propagation du gradient, pour calculer le gradient de la perte totale $L$ par rapport aux paramètres $\theta = (\alpha, \beta, \omega, W, V, c, R, \gamma)$ dans le système récurrent multi-échelle ci-bas, avec la série temporelle $y_t$ :

$$L = \sum_t (y_t - \hat{y}_t)^2 \tag{1}$$

$$\hat{y}_t = \text{sigm}(\alpha + \sum_i \beta_i Z_{ti}) \sum_i \omega_i h_{ti} \qquad \forall t \tag{2}$$

$$h_{ti} = \tanh(Z_{ti} + \sum_j W_{ij} h_{t-1,j} + \sum_k V_{ik} y_{t-k}) \qquad \forall i, t \tag{3}$$

$$Z_{ti} = \frac{(t \mod 10)}{10} z_{\lfloor t/10 \rfloor, i} + \frac{10 - (t \mod 10)}{10} z_{\lfloor t/10 \rfloor - 1, i} \qquad \forall i, t \tag{4}$$

$$z_{si} = \tanh(\eta_{si} + c_i + \sum_j R_{ij} z_{s-1,j} + \sum_{k,l} \gamma_{i,k} h_{10(s-1),k}) \qquad \forall i, s \tag{5}$$

où $s$ avance 10 fois plus lentement que $t$, i.e. si $t \in \{1, \ldots, T\}$, alors $s \in \{1, \ldots, T/10\}$, et $\text{sigm}(u) = 1/(1 + \exp(-u))$.

6. Considérez une RBM binaire-binaire ordinaire. Quelle est la fonction de partition ? Expliquez comment on pourrait utiliser une MCMC (chaîne de Markov Monte-Carlo) pour en estimer le gradient.

7. Quelle est l'interprétation probabiliste d'un auto-encodeur débruitant (*denoising auto-encoder*) ? Comment pourrait-on échantillonner de la distribution apprise par un auto-encodeur débruitant ?

8. Dans le cas d'une machine de Boltzmann complètement observée (pas d'unités cachées), peut-on tirer avantage du fait que toutes les variables sont observées pour éviter d'avoir à échantillonner des exemples négatifs (typiquement avec une chaîne de Markov) ? Justifier votre réponse mathématiquement.