Student name: _____

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

COURSE TITLE: **Algorithmes d'apprentissage**
COURSE NUMBER: **IFT6266 H12**

PROFESSOR: Yoshua Bengio

FINAL EXAM H12: 30 september 2012 TIME: 9h30 - 11h20
ROOM: Z-210

DIRECTIVES PÉDAGOGIQUES:
- Documentation allowed, but no computer allowed.
- Answer directly on the questionnaire, possibly using back pages.
- Be brief and precise.
- **If you sense that you will not have enough time, concentrate on showing that you have understood how to solve the problem, without wasting time on details.**
- Cheating is serious and will be severely punished (you can be rejected from the university).
- Suggestion: first read all the questions and start with those you find easiest. Gérez votre temps!

---

1. (14 points) Give an argument showing how a distributed representation (e.g. with an RBM) can be up to exponentially more efficient (in terms of number of parameters vs number of examples required) than a local representation (e.g. k-means clustering).

2. (12 points) How could learning of deep representations be useful to perform transfer? (e.g. where most examples are from classes other than the classes of interest).

3. (6 points) Consider a function f returning a random output given an input state $x \in \mathcal{X}$, returning a new state $x' \in \mathcal{X}$, i.e., f outputs a sample from a conditional distribution $Q(x'|x)$, and its successive application would generate a Monte-Carlo Markov chain. Let P represent the asymptotic distribution of that chain (and assume it exists), i.e., a distribution on the space $\mathcal{X}$. What relationship must hold between P and Q?

4. (16 points) Consider training a regularized auto-encoder, with on one hand a sparsity penalty, and on the other hand a denoising training criterion. Indicate for each of the hyper-parameters below the expected effect of variations of the hyper-parameter on (1) training error and (2) test error. Examples of effects: increase, decrease, none, U-shaped curve, $\cap$-shaped curve. Consider for each hyper-parameter the effect on (1) and (2) and the two cases **with** or **without** denoising. Hence there are *FOUR CASES* to consider for each hyper-parameter below (1-with, 1-w/o, 2-with, 2-w/o):

   (a) Number of training iterations.

   (b) Sparsity regularizer (on the hidden units).

   (c) Corruption noise level.

   (d) Number of hidden units.

5. (8 points) To follow-up on the previous question, a question that arises is whether one can use a test error (which one?) to choose these hyper-parameters in the two cases (with and without corruption) studied above. Please explain well your answer.

6. (14 points) What is the link between auto-encoders and principal components analysis? Explain clearly. You may use drawings as well.

7. (14 points) Consider the following variation of the contractive auto-encoder, with a penalty term for the colinearity between weight vectors of simultaneously active hidden units.

$$
\begin{aligned}
h_i(x) &= \text{sigm}(b_i + \sum_j w_{ij} x_j) \\
r_j(x) &= c_j + \sum_i w_{ji} h_i(x) \\
C(x) &= ||r(x) - x||^2 + \alpha ||\frac{\partial h(x)}{\partial x}||_F^2 + \beta \sum_{i \neq j} (\frac{\partial h_i(x)}{\partial x} \cdot \frac{\partial h_j(x)}{\partial x})^2
\end{aligned}
$$

Compute the gradient $\frac{\partial C(x)}{\partial w_{ij}}$ and write an expression as simplified as you can, that one could directly implement in a program.

8. (8 points) Please generalize the binomial-binomial RBM (with observed inputs x) to an energy function defined as follows, with two groups of hidden units h and g having a multiplicative interaction:

$$E = -\sum_{ijk} W_{ijk} x_i h_j g_k$$

and supposing again that all variables are binary. Start by showing that one can do Gibbs by blocks in three steps (one for each group of variables x, g et h), i.e., that the elements of each group are independent given the other two groups, and compute a formula for the probability of a node (say h_j) given the other two groups.

9. (8 points) To follow-up on the previous question, how could one learn (and update) the W's in this model, following one of the recipes known for the regular RBM? This will require providing an update procedure for the W, which should be obtained by looking at the derivative of the energy (or the free energy) with respect to parameters?