IFT6269-A2016	**Hwk 5**	Name:
Prof: Simon Lacoste-Julien	Due date: Dec 19, 2016	Student id:

As usual, please hand in on paper form your derivations and answers to the questions. You can use any programming language for your source code (submitted on Studium as per the website instructions). All the requested figures should be printed on paper with clear titles that indicate what the figures represent.

1. **Cautionary tale about importance sampling (10 points)**
Suppose that we wish to estimate the normalizing constant Z_p for an un-normalized Gaussian $\tilde{p}(x) = \exp(-\frac{1}{2\sigma_p^2}x^2)$; i.e. we have $p(\cdot) \sim \mathcal{N}(0, \sigma_p^2)$ with $p(x) = \tilde{p}(x)/Z_p$. Given N i.i.d. samples $x^{(1)}, \ldots, x^{(N)}$ from a standard normal $q(\cdot) \sim \mathcal{N}(0, 1)$, consider the importance sampling estimate:

$$\hat{Z} = \frac{1}{N}\sum_{i=1}^{N} \frac{\tilde{p}(x^{(i)})}{q(x^{(i)})}.$$

   (a) Show that $\hat{Z}$ is an unbiased estimator of Z_p.

   (b) Letting $f(x) := \tilde{p}(x)/q(x)$, show that $\mathrm{var}(\hat{Z}) = \frac{1}{N}\mathrm{var}(f(X)))$ whenever $\mathrm{var}(f(X))$ is finite.

   (c) For what values of σ_p^2 is this variance actually finite?

2. **Gibbs sampling and mean field variational inference (30 points)**
Consider the Ising model with binary variables $X_s \in \{0, 1\}$ and a factorization of the form:

$$p(x; \eta) = \frac{1}{Z_p} \exp\left(\sum_{s \in V} \eta_s x_s + \sum_{\{s,t\} \in E} \eta_{st} x_s x_t\right).$$

   We consider the 7×7 2D grid as shown in Figure 1 (note that we used toroidal (donut-like) boundary conditions to make the problem symmetric). We will consider approximate inference methods to approximate the node marginal moments $\mu_s := p(X_s = 1)$ in this model.

   (a) Derive the Gibbs sampling updates for this model. Implement the algorithm (with cyclic sequential traversal of the nodes) for $\eta_{st} = 0.5$ for all edges, and $\eta_s = (-1)^s$ for all $s \in \{1, \ldots, 49\}$ (using the node ordering of Figure 1). Run a burn-in period of 1000 epochs (where one epoch amounts to updating each node once). For each of the 5000 subsequent epochs, collect a sample vector, and use the 5000 samples to form Monte Carlo estimates $\hat{\mu}_s$ of the moments $\mathbb{E}[X_s]$ at each node.[1] Output a 7×7 matrix of the estimated moments. Repeat the experiment 10 times and output a 7×7 matrix of the empirical standard deviation of your estimate at each node (this gives an idea of the variability of your estimates).

[1] Note that I said in class that every update of a node yields a *different* sample in theory, and that one should normally use *all* the available samples (after sufficient mixing) for a Monte Carlo estimate, i.e. here it would be 49×5000 samples. But note that using all these samples would give almost the exact same estimates, only differing from the boundary conditions during the first and last epoch...

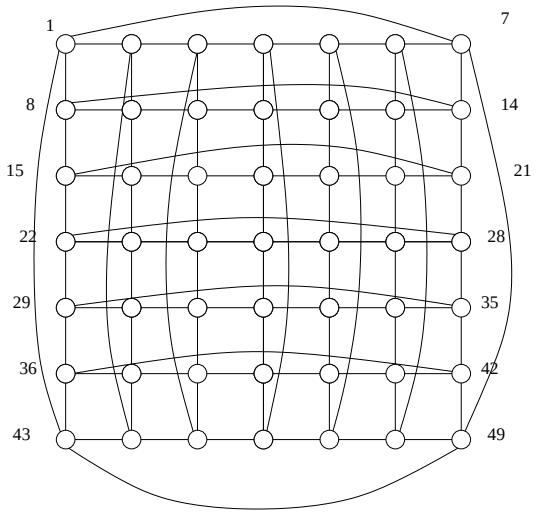

Figure 1: The undirected graphical model considered.

(b) Derive the naive mean field updates (based on a fully factorized approximation), where we use the notation $q(X_s = 1) = \tau_s$, and implement them for the same model. More specifically, do cyclic coordinate descent on $KL(q||p)$, sequentially updating the parameter $\tau_s \in [0, 1]$ for $s = 1, \ldots, 49$. Derive the expression for $KL(q||p) - \log(Z_p)$ and plot it as a function of the number of epochs both for debugging purpose and monitor progress. Let $d(\tau, \tau') := \frac{1}{49} \sum_{s=1}^{49} |\tau_s - \tau'_s|$ be the average ℓ_1 distance between two parameters. Use $d(\tau^{(k-1)}, \tau^k) < 0.001$ as a stopping criterion for convergence (where k counts the number of epochs). Compute $d(\hat{\tau}_s, \hat{\mu}_s)$ between the mean field estimated moments $\hat{\tau}_s$ and the Gibbs estimates $\hat{\mu}_s$. Is the mean field a good approximation here? Try different initializations – does it get stuck in different local minima?