Honor code: This take home exam is under the honor code. Ethical behavior is important in science as it is difficult to double check whether you are manipulating your results. Please start on the right foot and do not deceive my trust in you! This exam is **open book** but **closed internet**! This means that you are allowed to use your class notes (you can use the class website), you can use probability textbooks, but you are **not** allowed to search on the internet (or use Wikipedia e.g.). You are also obviously not allowed to discuss with anybody else the questions – it has to be **your own work** without the help of anybody else.

Please submit your solution as a pdf in Studium.

Total: 60 points

A - Short problems

1. (6 points) **Maximum entropy principle.** Let $\mathcal{P}$ be the set of all distributions on $\mathbb{N}$, the set of natural integers $\{0, 1, 2, \ldots\}$. What is the family of distribution $(p_\alpha)_{\alpha \in \mathbb{R}_+} \subseteq \mathcal{P}$, where p_α is the distribution of maximal entropy in $\mathcal{P}$ such that $\mathbb{E}_p[X] = \alpha$? Explain briefly why and gives the formula for the pmf $p_\alpha(n)$ as a function of α and $n \in \mathbb{N}$.

2. (6 points) **Sampling.** Propose a sampling scheme to sample exactly from the conditional distribution $X \mid \|X - y\|_2 \leq 1$ (i.e. X conditioned on the event that $\|X - y\|_2 \leq 1$), where $y \in \mathbb{R}^d$ is a fixed vector and X is a multivariate Gaussian random variable $\mathcal{N}(0, I_d)$. Prove that the proposed sampling scheme indeed yields a variable that has exactly the desired distribution.

B - Factorization and Markov properties

1. (4 points) **UGM from constraints.** Find the unique undirected graphical model on four random variables X_1, X_2, X_3, X_4 that satisfies simultaneously the following conditions (a) and (b):

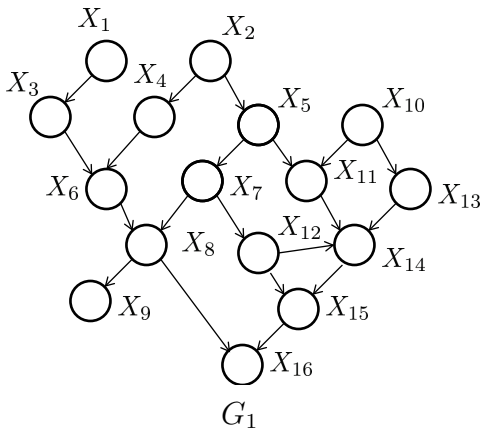
   (a) All distributions that factorize according to the model satisfy the following conditional independence statements:

   $$X_1 \perp\!\!\!\perp X_2 \mid (X_3, X_4), \quad X_3 \perp\!\!\!\perp X_4 \mid (X_1, X_2), \quad X_1 \perp\!\!\!\perp X_4 \mid X_3$$

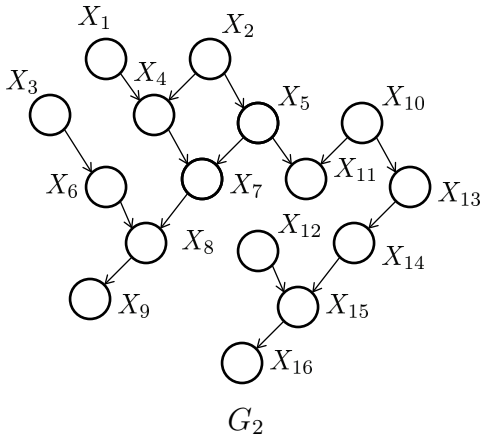
   (b) There exist distributions that factorize according to the model and such that

   $$X_1 \not\perp\!\!\!\perp X_3 \mid X_2, \quad X_2 \not\perp\!\!\!\perp X_4 \mid X_1, \quad X_1 \not\perp\!\!\!\perp X_4.$$

2. (4 points) **DGM.** Consider the two directed graphical models below. Is it possible to have a distribution p on $X_1, \ldots, X_{16}$ such that $p \in \mathcal{L}(G_1)$ and $p \in \mathcal{L}(G_2)$? Justify your answer. HINT: your answer should be quite short!



G_1



G_2

C - EM algorithm (10 points)

NB model of Bernouilli variables.

Let $X \in \{0,1\}^L$ be a random binary vector of length L. We consider the following naive Bayes (NB) latent variable model for X. We have a Bernouilli latent variable $Z \in \{0,1\}$ with $p(z=1) = \pi$. Writing X as $(X_1, ..., X_L)$, we have that given $Z = z$, we posit that the X_l's are *conditionally independent* Bernouilli random variables with probability θ_z (this is a simple NB model). That is, $p(X_l = 1|Z = 1) = \theta_1$ and $p(X_l = 1|Z = 0) = \theta_0$.

Suppose that we have n i.i.d. observations $\{X^{(i)}\}_{i=1}^n$ from this model. We want to estimate the parameters $(\pi, \theta_0, \theta_1)$ by maximum likelihood via the EM algorithm.

1. Draw the full graphical model (with latent variables) of the situation above using the plate notation. The random variables are $X_l^{(i)}$ and $Z^{(i)}$ for $i = 1, \ldots, n$, $l = 1, \ldots, L$.

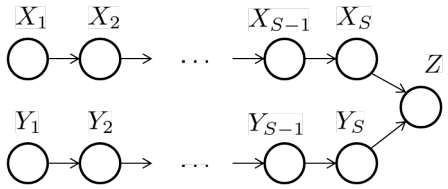
2. Derive the EM algorithm to estimate $(\pi, \theta_0, \theta_1)$. In particular, specify which quantities are computed for the E-step, and which quantities are computed for the M-step. To help the grading, express your solution using the following notation for the relevant quantities:

   - $\tau_i := p(Z^{(i)} = 1|X^{(i)})$
   - $\tau := \sum_{i=1}^n \tau_i$
   - $c_i := \sum_{l=1}^L X_l^{(i)}$

   And as τ_i is a probability between 0 and 1, please express it using the sigmoid function $\sigma(\cdot)$ ($\sigma(y) := \frac{1}{1+\exp(-y)}$).

D - Parallel chains (6 points)

Consider the directed graphical model G below:



Suppose that the variables X_s and Y_s are discrete K-valued for $s = 1, \ldots, S$, and that Z is a binary random variable. Consider the following form for the factors for a specific distribution p in $\mathcal{L}(G)$:

- $p(X_1 = l) = p(Y_1 = l) = \pi_l$ for $l = 1, \ldots, K$.
- $p(X_s = i | X_{s-1} = j) = p(Y_s = i | Y_{s-1} = j) = A_{ij}$ for $i, j = 1, \ldots, K$ and $s = 2, \ldots, S$.
- $p(Z = 1 | X_S = k, Y_S = l)$ takes the value p whenever $k = l$, and q otherwise (i.e. when $k \neq l$).

By using the graph eliminate algorithm (or just clever use of distributivity), give a simple formula for the marginal $p(Z = 1)$ in terms of the matrix A, vector π and scalars q, p and S. Provide brief explanations of how to obtain the result.

E - Metropolized Gibbs sampler (10 points)

We would like to obtain samples from a *strictly positive* distribution p of the random variable $X = (X_1, \ldots, X_n)$ (which we can think of as a distribution from a graphical model of interest). For simplicity, we assume that the X_i are K-valued random variables, with $K \geq 3$. We will use the notation X_{-i} to refer to $X_{\{1,\ldots,n\} \setminus \{i\}}$. We will denote by p_i the conditional distribution of the ith variable given all the others, so that

$$p_i(\, z_i \mid x^t_{-i}) := \mathbb{P}(X_i = z_i \mid X_{-i} = x^t_{-i}).$$

The Metropolized Gibbs sampler is a variant of Gibbs sampling, which takes the form of a Metropolis-Hasting algorithm that updates a single variable X_i at a time: to update the variable X_i, instead of just sampling this variable conditionally on the other variables, it makes a proposal drawn from the transition q_i with

$$q_i\big(z_i, x^t_{-i} | x^t_i, x^t_{-i}\big) := \begin{cases} \dfrac{p_i(z_i \mid x^t_{-i})}{1 - p_i(x^t_i \mid x^t_{-i})} & \text{for } z_i \neq x^t_i \\ 0 & \text{for } z_i = x^t_i, \end{cases} \tag{1}$$

and where the x^t_{-i} part is not affected. The motivation for the variant is that we can accelerate the exploration of the space by taking in consideration the previous value for X_i given by x^t_i in the proposal.

1. Give the acceptance probability $\alpha_i((z_i, x^t_{-i})|(x^t_i, x^t_{-i})$ that arises from this proposal distribution q_i in a Metropolis-Hasting algorithm for the target distribution p (i.e. that ensures that detailed balance equation is satisfied).

2. Consider using a cyclic scan Gibbs sampler that samples each variable X_i in a cyclic order. Does the Markov Chain produced by the Metropolized Gibbs sampler converge to the target distribution p? Explain why (state the explicit conditions used for your conclusion).

3. BONUS (optional): what could go wrong when $K = 2$?

F - Bayesian point estimates: posterior mean *vs.* MAP for exponential family (14 points)

Assume we observe an i.i.d. sample $(x_1, \dots, x_n)$ of realizations of a Bernoulli random variable whose unknown mean parameter we denote by $\mu = \mathbb{E}_\mu[X] = \mathbb{P}_\mu(X = 1)$.

We first consider the Bayesian estimation of μ. We choose as a prior for μ the uniform distribution on $[0, 1]$, and will consider the corresponding MAP estimator.

1. Express the posterior mean estimate $\hat{\mu}_{\text{PM}}$ for μ as a function of $\bar{x} := \frac{1}{n}\sum_{i=1}^n x_i$.

2. What is the maximum a posteriori (MAP) estimate $\hat{\mu}_{\text{MAP}}$ for μ?

We now consider again the same estimators, but we instead work with η, the canonical parameter of the exponential family.

3. We said that the prior $p_\mu(\mu)$ on μ is the uniform distribution on $[0, 1]$. Express the canonical parameter η of a Bernouilli as a function of its moment parameter μ, i.e. $\eta(\mu)$. Given the prior $p_\mu(\mu)$, show that the induced prior distribution $p_\eta(\eta)$ on η (under the mapping $\eta(\mu)$) is $p_\eta(\eta) = \sigma(\eta)(1 - \sigma(\eta)$, where $\sigma(\cdot)$ is the sigmoid function.

4. What is the value of the posterior mean estimate $\int \mu(\eta)\, p(\eta|x_1, \dots, x_n)\, d\eta$ under this prior p_η on η, for $\mu(\eta) = \sigma(\eta)$ the usual moment mapping?

5. For this prior p_η, what is the MAP estimator $\hat{\eta}_{\text{MAP}}$? What is the corresponding moment parameter $\mu(\hat{\eta}_{\text{MAP}})$?

6. Comment on the different estimators obtained.