1 Problem 1

For this problem, I will call the random variables A, B, C and D for simpler notation. The graph structure $A - B - C - D - A$ implies that the distribution can be factorized as

$$p(a,b,c,d) \propto \phi_{AB}(a,b)\,\phi_{BC}(b,c)\,\phi_{CD}(c,d)\,\phi_{DA}(d,a) \quad . \tag{1}$$

Therefore, in order for $p(0,1,1,0) = 0$, at least one of

$$\phi_{AB}(0,1), \quad \phi_{BC}(1,1), \quad \phi_{CD}(1,0), \quad \text{or} \quad \phi_{DA}(0,0) \tag{2}$$

must be equal to zero. This is clearly not true since

$$(i)\,p(0,1,1,1) \ne 0, \quad (ii)\,p(1,1,1,0) \ne 0 \quad \text{and} \quad (iii)\,p(0,0,0,0) \ne 0 \quad . \tag{3}$$

Statement (i) implies that none of

$$\phi_{AB}(0,1), \quad \phi_{BC}(1,1), \quad \phi_{CD}(1,1), \quad \text{or} \quad \phi_{DA}(1,0) \tag{4}$$

can equal zero. Similarly, statement (ii) implies that none of

$$\phi_{AB}(1,1), \quad \phi_{BC}(1,1), \quad \phi_{CD}(1,0), \quad \text{or} \quad \phi_{DA}(0,1) \tag{5}$$

can equal zero. And finally, statement (iii) implies that none of

$$\phi_{AB}(0,0), \quad \phi_{BC}(0,0), \quad \phi_{CD}(0,0), \quad \text{or} \quad \phi_{DA}(0,0) \tag{6}$$

are zero. Therefore, by this contradicting example, the distribution cannot be factorized over the given graph structure.

2 Problem 2

(a) A probability distribution over three discrete random variables A, B and C is parameterized as

$$p(a,b,c) \propto \exp\left(-\epsilon_1(a,b) - \epsilon_2(b,c)\right) \quad . \tag{7}$$

If we redefine

$$\epsilon_1'(a, B=b^i) \leftarrow \epsilon_1(a, B=b^i) + \lambda^i \quad \text{and} \quad \epsilon_2'(B=b^i, c) \leftarrow \epsilon_2(B=b^i, c) - \lambda^i \tag{8}$$

then the new distribution is

$$\begin{aligned} p(a,b^i,c) &\propto \exp\left(-\epsilon_1'(a,b^i) - \epsilon_2'(b^i,c)\right) & (9)\\ &= \exp\left(-\epsilon_1(a,b^i) - \lambda^i - \epsilon_2(b^i,c) + \lambda^i\right) & (10)\\ &= \exp\left(-\epsilon_1(a,b^i) - \epsilon_2(b^i,c)\right) & (11) \end{aligned}$$

which is equivalent to the original parameterization. Therefore, any symmetric reparameterization of the energy functions will leave the distribution unchanged.

(b) We would like to find w'_{ij} and u'_i such that

$$p_{\mathrm{Ising}} \propto \exp\left(-\sum_{i<j\in E} w'_{ij}\, z_i\, z_j - \sum_i u'_i\, z_i\right) \tag{12}$$

for $z_i = \pm 1$ is equivalent to

$$p_{\mathrm{Boltzmann}} \propto \exp\left(-\sum_{i,j\in E} w_{ij}\, x_i\, x_j - \sum_i u_i\, x_i\right) \tag{13}$$

for $x_i \in \{0, 1\}$. This can be easily achieved with the mapping $x_i \leftarrow (z_i + 1)/2$. Substituting this into Equation (13), we find

$$\begin{aligned} p_{\mathrm{Boltzmann}} &\propto \exp\left(-\sum_{i,j\in E} \frac{w_{ij}}{4}\,(z_i\, z_j + z_i + z_j + 1) - \sum_i \frac{u_i}{2}\,(z_i + 1)\right) && (14)\\ &= \exp\left(-\sum_{i,j\in E} \frac{w_{ij}}{4}\, z_i\, z_j - \sum_{i,j\in E} \frac{w_{ij}}{4}\,(z_i + z_j) - \sum_i \frac{u_i}{2}\, z_i - \sum_{i,j\in E} \frac{w_{ij}}{4} - \sum_i \frac{u_i}{2}\right) && (15)\end{aligned}$$

The last two terms in this equation are constant so they can be absorbed into the partition function, leaving

$$p_{\mathrm{Boltzmann}} \propto \exp\left(-\sum_{i,j\in E} \frac{w_{ij}}{4}\, z_i\, z_j - \sum_i \left[\frac{u_i}{2} + \sum_{j\in E_i} \frac{w_{ij}}{4}\right] z_i\right) \tag{16}$$

where E_i is the set of edges containing node i. Therefore, the correct mapping between the two distributions is

$$w'_{ij} = \frac{w_{ij}}{4} \quad \text{and} \quad u'_i = \frac{u_i}{2} + \sum_{j\in E_i} \frac{w_{ij}}{4} \quad . \tag{17}$$

3 Problem 3

For a simple (non-pairwise) distribution on 3 random variables A, B and C, factored according to

$$p(a, b, c) \propto \phi_{ABC}(a, b, c) \quad , \tag{18}$$

we can introduce a new variable D to convert it to the pairwise

$$p(a, b, c, d) \propto \phi_{AD}(a, d)\, \phi_{BD}(b, d)\, \phi_{CD}(c, d) \tag{19}$$

where

$$\begin{aligned} p(a, b, c) &= \sum_i p(a, b, c, D = d^i) && (20)\\ \phi_{ABC}(a, b, c) &\propto \sum_i \phi_{AD}(a, D = d^i)\, \phi_{BD}(b, D = d^i)\, \phi_{CD}(c, D = d^i) \quad . && (21)\end{aligned}$$

To determine the forms of the pairwise potentials, first, we can assert that D assumes a tuple value $(d_1, d_2, \ldots)$ with one entry for each connected node (i.e. three in this example for A, B, and C). Then, the potentials will be

$$\phi_{AD}(a, d) = \phi_{ABC}(a, d_2, d_3) \tag{22}$$

with similar forms for the other edges.

Following this example, the general procedure for *one* particular non-pairwise potential will be

$$\phi_X(\boldsymbol{x}) \to \prod_{i=1}^{N} \phi_{X_i,Y}(x_i, Y) \tag{23}$$

where Y is an N-tuple and

$$\phi_{X_i,Y}(x_i, Y) = \phi_X(y_1, \ldots, y_{i-1}, x_i, y_{i+1} \ldots, y_N) \quad . \tag{24}$$

4 Problem 4: Exponential Families

(a)

1. A multivariate Gaussian with identity covariance in K dimensions is part of the exponential family:

$$\begin{aligned} \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, I) &= \frac{1}{(2\pi)^{K/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{K}[x_i - \mu_i]^2\right) & (25)\\ &= \frac{1}{(2\pi)^{K/2}} \exp\left(\sum_{i=1}^{K}\mu_i x_i - \frac{1}{2}\sum_{i=1}^{K}x_i^2 - \frac{1}{2}\sum_{i=1}^{K}\mu_i^2\right) \quad . & (26) \end{aligned}$$

   Therefore, setting $\boldsymbol{f}(\boldsymbol{x}) = (x_1, x_1^2, \ldots, x_K, x_K^2)^T$, $\boldsymbol{\eta} = (\mu_1, -1/2, \ldots, \mu_K, -1/2)^T$, $\ln Z = \sum \mu_i^2/2 + K\ \ln(2\pi)/2$ and $h(\boldsymbol{x}) = 1$ puts this distribution in the correct form.

2. The Dirichlet distribution in K dimensions is

$$D(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \frac{1}{Z(\boldsymbol{\alpha})} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \tag{27}$$

   where

$$Z(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma\left(\prod_{i=1}^{K}\alpha_i\right)} \quad . \tag{28}$$

   This can be rewritten as

$$D(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \exp\left(\sum_{i=1}^{K}[1 - \alpha_i]\ln\theta_i - \ln Z(\boldsymbol{\alpha})\right) \quad . \tag{29}$$

   This is clearly a member of the exponential family with $\boldsymbol{f}(\boldsymbol{\theta}) = (\ln\theta_1, \ldots, \ln\theta_K)^T$, $\boldsymbol{\eta} = (1 - \alpha_1, \ldots, 1 - \alpha_K)^T$ and $h(\boldsymbol{\theta}) = 1$.

3. The log-normal distribution is parameterized as

$$\begin{aligned}\mathcal{L}(y;0,\sigma^2) &= \left|\frac{\mathrm{d}\ln y}{\mathrm{d}y}\right| \mathcal{N}(\ln y;0,\sigma^2) &(30)\\ &= \frac{1}{y\sqrt{2\pi\sigma^2}}\exp\left(-\frac{[\ln y]^2}{2\,\sigma^2}\right) \quad . &(31)\end{aligned}$$

Setting $f(y) = (\ln y)^2$, $\eta = -1/2\sigma^2$, $h(y) = y^{-1}$ and $Z = 1/\sqrt{2\pi\sigma^2}$ shows that this is also a member of the exponential family.

4. The Boltzmann distribution can be written as

$$\begin{aligned}p(\boldsymbol{x}) &= \frac{1}{Z}\exp\left(\sum_i u_i\,x_i + \sum_{i,j\in E} w_{ij}\,x_i\,x_j\right) &(32)\\ &= \frac{1}{Z}\exp\left(\sum_i u_i\,x_i + \sum_{i,j\in E} w_{ij}\,x_i\,x_j - \ln Z\right) \quad . &(33)\end{aligned}$$

Therefore, we can set $\boldsymbol{\eta} = \{\boldsymbol{u},\boldsymbol{w}\}$ where $\boldsymbol{w} = \{w_{ij},\,\forall\,(i,j)\in E\}$ and $\boldsymbol{f}(\boldsymbol{x}) = \{\boldsymbol{x},\boldsymbol{\xi}\}$ where $\boldsymbol{\xi} = \{x_i\,x_j,\,\forall\,(i,j)\in E\}$ to show that this distribution can also be written in the form of a member of the exponential family.

(b) For a continuous distribution, the partition function is given by

$$Z(\boldsymbol{\eta}) = \int h(\boldsymbol{x})\,\exp\left(\boldsymbol{\eta}\cdot\boldsymbol{f}(\boldsymbol{x})\right)\,\mathrm{d}\boldsymbol{x} \tag{34}$$

and the gradient with respect to $\boldsymbol{\eta}$ is

$$\nabla_\eta Z(\boldsymbol{\eta}) = \int \boldsymbol{f}(\boldsymbol{x})\,h(\boldsymbol{x})\,\exp\left(\boldsymbol{\eta}\cdot\boldsymbol{f}(\boldsymbol{x})\right)\,\mathrm{d}\boldsymbol{x} \quad . \tag{35}$$

Therefore, since

$$\begin{aligned}\nabla_\eta \ln Z(\boldsymbol{\eta}) &= \frac{1}{Z(\boldsymbol{\eta})}\nabla_\eta Z(\boldsymbol{\eta}) &(36)\\ &= \int \boldsymbol{f}(\boldsymbol{x})\,p(\boldsymbol{x};\boldsymbol{\eta}) &(37)\end{aligned}$$

where

$$p(\boldsymbol{x};\boldsymbol{\eta}) = h(\boldsymbol{x})\,\exp\left(\boldsymbol{\eta}\cdot\boldsymbol{f}(\boldsymbol{x}) - \ln Z(\boldsymbol{\eta})\right) \quad , \tag{38}$$

the result is immediately clear

$$\nabla_\eta \ln Z(\boldsymbol{\eta}) = \int \boldsymbol{f}(\boldsymbol{x})\,p(\boldsymbol{x};\boldsymbol{\eta}) = E_{p(x;\eta)}\left[\boldsymbol{f}(\boldsymbol{x})\right] \quad . \tag{39}$$

(c) For the multivariate Gaussian in example 1, the log-partition function becomes

$$\ln Z = \frac{1}{2} \sum_{i=1}^{K} \left[\frac{\mu_i^2}{\sigma_i^2} + \ln \sigma_i^2 + \ln 2\pi \right] \tag{40}$$

when we intorduce a diagonal covariance tensor $\mathbf{\Sigma} = \mathrm{Diag}(\sigma_1^2, \sigma_2^2, \ldots)$. Also, $\boldsymbol{\eta} \leftarrow \{\mu_i/\sigma_i^2, -1/2\sigma_i^2\} = \{\alpha_i, \beta_i\}$ so the log partition function can be re-written

$$\ln Z = \sum_i \left[-\frac{\alpha_i^2}{4\beta_i} + \ln\left(-\frac{1}{2\beta_i} \right) + \ln 2\pi \right] \tag{41}$$

Therefore, the derivative of $\ln Z$ with respect to $\alpha_i = \mu_i/\sigma_i^2$ is

$$\frac{\mathrm{d} \ln Z}{\mathrm{d}\alpha_i} = -\frac{\alpha_i}{2\beta_i} = \mu_i = E[x_i] \tag{42}$$

as expected and

$$\frac{\mathrm{d} \ln Z}{\mathrm{d}\beta_i} = \frac{\alpha_i^2}{4\beta_i^2} - \frac{1}{2\beta_i} = \mu_i^2 + \sigma_i^2 = E[x_i^2] \quad . \tag{43}$$

This verifies the claim that the gradient of the log partition function gives the expectaion values of $\boldsymbol{f}(\boldsymbol{x})$.

(d) Since

$$p(Y = 1 | \boldsymbol{x}; \boldsymbol{\alpha}) = \frac{1}{1 + e^{-\boldsymbol{\alpha} \cdot \boldsymbol{x}}} \tag{44}$$

for $\boldsymbol{x} = (1, x_1, x_2, \ldots, x_n)$, $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_n)$ and binary Y,

$$p(Y = 1 | \boldsymbol{x}; \boldsymbol{\alpha}) = 1 - \frac{1}{1 + e^{-\boldsymbol{\alpha} \cdot \boldsymbol{x}}} = \frac{e^{-\boldsymbol{\alpha} \cdot \boldsymbol{x}}}{1 + e^{-\boldsymbol{\alpha} \cdot \boldsymbol{x}}} \quad . \tag{45}$$

Therefore,

$$p(Y = y | \boldsymbol{x}; \boldsymbol{\alpha}) = \frac{e^{(1-y)\, \boldsymbol{\alpha} \cdot \boldsymbol{x}}}{Z(\boldsymbol{\alpha}, \boldsymbol{x})} \tag{46}$$

where

$$Z(\boldsymbol{\alpha}, \boldsymbol{x}) = 1 + e^{-\boldsymbol{\alpha} \cdot \boldsymbol{x}} \quad . \tag{47}$$

Therefore, setting $\boldsymbol{f}(y, \boldsymbol{x}) = (1 - y)\, \boldsymbol{x}$ and $h(\boldsymbol{x}, \boldsymbol{y}) = 1$, we see that this conditional distribution is part of the exponential family.

5 Problem 5: Conjugacy and Prediction

(a) The Dirichlet distribution is

$$\mathrm{Dir}(\boldsymbol{\theta} | \boldsymbol{\alpha}) \propto \prod_k \theta_k^{\alpha_k - 1} \tag{48}$$

and the Multinomial distribution is

$$\mathrm{Mult}(\boldsymbol{x} | \boldsymbol{\theta}) \propto \prod_k \theta_k^{x_k} \quad . \tag{49}$$

Therefore, the posterior on $\boldsymbol{\theta}$ is

$$p(\boldsymbol{\theta}|\boldsymbol{x};\boldsymbol{\alpha}) \propto \left(\prod_k \theta_k^{x_k}\right)\left(\prod_k \theta_k^{\alpha_k-1}\right) = \prod_k \theta_k^{\alpha_k-1+x_k} \propto \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}+\boldsymbol{x}) \quad . \tag{50}$$

Therefore, the posterior given a single "observation" $\boldsymbol{x}$ is given by a Dirichlet with new hyperparameters. Given multiple, independent observations, this becomes

$$p(\boldsymbol{\theta}|\{\boldsymbol{x}\};\boldsymbol{\alpha}) \propto \prod_{i=1}^{N} p(\boldsymbol{\theta}|\boldsymbol{x}^{(i)};\boldsymbol{\alpha}) = \prod_{i=1}^{N} \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}+\boldsymbol{x}^{(i)}) \quad . \tag{51}$$

Now, to give the result in the notation of the problem, since $\boldsymbol{x}^{(i)}$ is zero everywhere except in one component where it equals one, the posterior can be written as a Dirichlet distribution with hyperparameters $\boldsymbol{\alpha}'$ given by

$$\alpha'_k = \alpha_k + \sum_{i=1}^{N} \begin{cases} 1, & \text{if } \boldsymbol{x}_k^{(i)} = 1 \\ 0, & \text{otherwise} \end{cases} \tag{52}$$

.

(b) The joint posterior on $\boldsymbol{x}_\mathrm{new}$ and $\boldsymbol{\theta}$ is

$$p(\boldsymbol{x}_\mathrm{new},\boldsymbol{\theta}|\{\boldsymbol{x}^{(i)}\};\boldsymbol{\alpha}) = p(\boldsymbol{x}_\mathrm{new}|\boldsymbol{\theta})\,p(\boldsymbol{\theta}|\{\boldsymbol{x}^{(i)}\};\boldsymbol{\alpha}) \tag{53}$$

and this is given by

$$\mathrm{Mult}(\boldsymbol{x}_\mathrm{new}|\boldsymbol{\theta})\,\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}') \sim \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}'') \quad . \tag{54}$$

Integrating over $\boldsymbol{\theta}$, we get

$$p(\boldsymbol{x}_\mathrm{new}|\{\boldsymbol{x}^{(i)}\};\boldsymbol{\alpha}) = \int \mathrm{d}\boldsymbol{\theta}\,p(\boldsymbol{x}_\mathrm{new},\boldsymbol{\theta}|\{\boldsymbol{x}^{(i)}\};\boldsymbol{\alpha}) = \frac{\prod_k \Gamma\left(\alpha'_k + x_{\mathrm{new},k}\right)}{\Gamma\left(\sum_k [\alpha'_k + x_{\mathrm{new},k}]\right)} \quad . \tag{55}$$

6 Problem 6: Kullback-Leibler divergence

(a) For a convex function $f(x)$, Jensen's inequality is

$$E[f(x)] \ge f(E[x]) \quad . \tag{56}$$

For our problem, we can define $y = q/p$ and $f(x) = -\log x$. Therefore, $E_p[y] = \sum p \frac{q}{p} = \sum q = 1$ and

$$E_p[f(y)] = -\sum p \log\frac{q}{p} = \sum p \log\frac{p}{q} \ge -\log\left(E_p[y]\right) = -\log(1) = 0 \quad . \tag{57}$$

This proves that $D(p||q) \ge 0$. If $p = q$ then

$$D(p||q) = \sum p\,\log\frac{p}{q} = \sum p\,\log 1 = \sum 0 = 0 \quad . \tag{58}$$

If $p \neq q$ then $D(p||q) > 0$ since equality in Jensen's inequality holds only when $f(x)$ is not strictly convex (i.e. only when $p = q$). Therefore, $D(p||q) = 0$ if and only if $p = q$.

(b) The K–L divergence can be rewritten as

$$D(p||q) = \sum p\,\log\frac{p}{q} = \sum(p\log p - p\log q) = -H(p) - \sum p\log q \quad . \tag{59}$$

Then, we can choose the uniform distribution $q = 1/k$ and find

$$D(p||q) = -H(p) + \log k\sum p = -H(p) + \log k \geq 0 \rightarrow \log k \geq H(p) \tag{60}$$

with equality when $p = q = 1/k$.