1. $E(s_1, s_2) = Js_1s_2 + h_1s_1 + h_2s_2$ with $s_1, s_2 \in \{0,1\}$

Given the Gibbs distribution $Pr(S_1 = s_1, S_2 = s_2) = \dfrac{\exp\{-\beta E(s_1,s_2)\}}{Z}$.

Z is partition function $\sum_{s_1}^{s_2} \exp(-\beta E(s_1, s_2))$

Substitute energy function:

$$Z = \sum_{s_1}^{s_2} \exp(-\beta(-Js_1s_2 + h_1s_1 + h_2s_2))$$

$$= \sum_{0}^{s_1} \exp(\beta h_1 s_1) \sum_{0}^{s_2} \exp(\beta h_2 s_2) \sum_{s_1}^{s_2} \exp(\beta J s_1 s_2)$$

$$\boxed{= (1+\exp(\beta h_1))(1+\exp(\beta h_2))(1+2\exp(\beta J))}$$

$$\mathcal{E}(s_1 s_2) = \sum_{s_1}^{s_2} s_1 s_2 \, Pr(S_1 = s_1, S_2 = s_2)$$

$$= \sum_{s_1}^{s_2} s_1 s_2 \frac{\exp(\beta J s_1 s_2 - \beta h_1 s_2 - \beta h_2 s_2)}{1+\exp(\beta h_2)} (1+\exp(\beta h_2))(1+2\exp(\beta J))$$

$$\boxed{= \frac{\exp(\beta J) + \exp(2\beta J)}{1+\exp(\beta h_1)} (1+\exp(\beta h_2))(1+2\exp(\beta J))}$$

2. The KL divergence between p and q is given by

$$D(p\|q) = \sum x p(x) \log(p \tfrac{x}{q}(x))$$

$$L(p,\lambda) = D(p\|q) + \sum i \lambda i (\sum x p(x) g i(x) - a i)$$

$$\log(p \tfrac{x}{q}(x)) = -1 - \lambda_1 g_1(x) - \lambda_2 g_2(x) - \dots - \lambda_k g_k(x)$$

$$p(x) = \exp(-1 - \lambda_1 g_1(x) - \lambda_2 g_2(x) - \dots - \lambda_k g_k(x)) q(x)$$

$$\sum x p(x) g i(x) = a i, \; i = 1, 2, \dots$$

$$\sum x \exp(-1 - \lambda_1 g_1(x) - \lambda_2 g_2(x) - \dots - \lambda_k g_k(x)) q(x) g i(x) = a i, \; i = 1, 2, \dots$$

3. We know that $\nabla\varphi(y) - \nabla\varphi \dots$ is the gradient of φ at x.

- We use identity matrix $(\nabla_2 p(x))(y-x)^T(y-x) = \nabla\varphi(x) \cdot \nabla\varphi(x)^T$

$$H = \nabla\varphi(x) \cdot \nabla\varphi \frac{x^T}{(y-x)^T(y-x)}$$

$$u^T \cdot H \cdot u = u^T \cdot \nabla\varphi(x) \cdot \nabla\varphi \frac{x^T}{(y-x)^T(y-x)} \cdot u = (u^T \cdot \nabla\varphi(x)) \cdot \frac{\nabla\varphi(x)^T u}{(y-x)^T(y-x)}$$

$$= \frac{(u^T \cdot \nabla\varphi(x))^2}{(y-x)^T(y-x)} \geq 0$$

$$v^T \cdot H \cdot v = \frac{(v^T \cdot \nabla\varphi(x))^2}{(y-x)^T(y-x)} \geq 0$$

These 2 in conjunction show that the Bregman divergence is convex w.r.t x and y assuming that $\varphi(x)$ is 3 times differentiable.

4. $P(X=k) = p_k$

$H(X) = -\sum p_k \cdot \log_2(p_k)$ ⟵ Entropy

$p_k = \frac{N_k}{N}$

Probability of observing the scores of N students: $P(X_1 = x_1, X_2 = x_2, \dots X_N$

⟶ From this we find the negative log-likelihood of the data:

$L(\theta) = -\log(P(X_1 = x_1, X_2 = x_2, \dots X_N = x_N)$

$= -(-N \cdot \sum (\frac{N_k}{N}) \cdot \log(N_k) + N \cdot \log(N))$

$= N \cdot (-\sum (\frac{N_k}{N}) \cdot \log(N_k)) - N \cdot \log(N)$

$= -N \cdot (\sum (\frac{N_k}{N}) \cdot \log(N_k)) + H(X) \cdot N$

We find that the negative log-likelihood of the data is equal to the negative of entropy multiplied by all students in class. We then add constant factor $N \cdot \log(N)$. They are each related through this factor which is dependent on the total number of students in class. Thus, maximizing the entropy gives the most accurate prediction of the probability distribution of the midterm scores.

$= x_N) = \prod p_k^{N_k}$