Using this MRF, we can see that all $v_1, \dots, v_n$ are conditionally independent given C. This will allow us to use the known appearance models to get $P(C \mid v_1, \dots, v_n)$

If $v_i \perp v_j \mid C$ for all i, j then

$$P(v_1, \dots, v_n \mid C) = \prod_{i=1}^{k} P(v_i \mid C) = \prod_{i=1}^{k} P(v_i \mid C_i)$$

Using Bayes Rule:

$$P(C \mid v_1, \dots, v_n) = \frac{P(C, v_1, \dots, v_n)}{P(v_1, \dots, v_n)} = \frac{P(v_1, \dots, v_n \mid C)\, P(C)}{P(v_1, \dots, v_n)}$$

Now suppose $C = \{C_i = i, C_j = j, C_{\dots}\}$
and we swap C_i and C_j to get $C' = \{C_i = j, C_j = i, C_{\dots}\}$

Then
$$A(C' \mid C) = \min\left(1, \frac{\dfrac{P(v_1, \dots, v_n \mid C') \cdot P(C')}{P(v_1, \dots, v_n)}}{\dfrac{P(v_1, \dots, v_n \mid C) \cdot P(C)}{P(v_1, \dots, v_n)}}\right)$$

$$= \min\left(1, \frac{P(v_1, \dots, v_n \mid C') \cdot P(C')}{P(v_1, \dots, v_n \mid C) \cdot P(C)}\right) \rightarrow P(C) = P(C') = \frac{1}{k!} \rightarrow \frac{P(C')}{P(C)} = 1$$

$$= \min\left(1, \frac{P(v_1, \dots, v_n \mid C_i = j, C_j = i, C_{\dots})}{P(v_1, \dots, v_n \mid C_i = i, C_j = j, C_{\dots})}\right)$$

$$\boxed{= \min\left(1, \frac{P(v_i \mid C_i = j) \cdot P(v_j \mid C_j = i)}{P(v_i \mid C_i = i) \cdot P(v_j \mid C_j = j)}\right)}$$

1) Data Association

- K objects $u_1, \dots, u_K$
- K observations $v_1, \dots, v_K$, $Val(v_i) = \{a_1, \dots, a_L\}$
- K correspondence variables $C_1, \dots, C_K$, where $Val(C_i) = \{1, \dots, K\}$
- known appearance model for each object u_K, $P_K(v_i = a_l \mid C_i = K)$

a) Compute acceptance probability for each MH step

$$A(x' \mid x) = \min\left(1, \frac{P(x')\, Q(x \mid x')}{P(x)\, Q(x' \mid x)}\right)$$

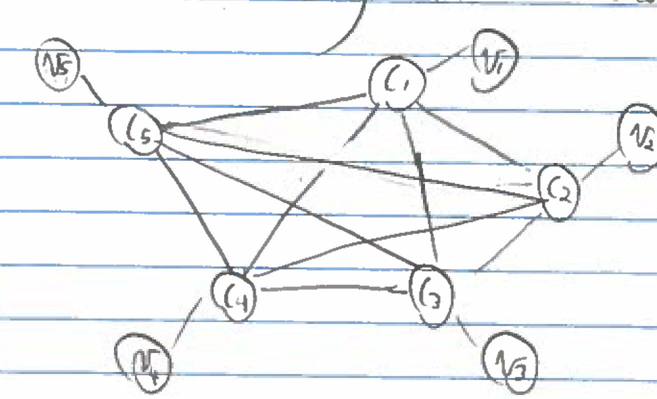
$$= \min\left(1, \frac{P(C' \mid v_1, \dots, v_K) \cdot Q(C \mid C')}{P(C \mid v_1, \dots, v_K) \cdot Q(C' \mid C)}\right)$$

Q models the probability of sampling a new assignment $C' = \{C_1', C_2', \dots, C_K'\}$ given $C = \{C_1, C_2, \dots, C_K\}$
$Q(C' \mid C)$ is 0 if C' is not C w/ exactly two values swapped
$Q(C' \mid C)$ is uniform otherwise

Since $Q(C' \mid C)$ and $Q(C \mid C')$ are equally likely, these cancel to 1 in A.

$$A(C' \mid C) = \min\left(1, \frac{P(C' \mid v_1, \dots, v_K)}{P(C \mid v_1, \dots, v_K)}\right)$$

We can use the following MRF to describe the problem:



Cs are fully connected, V_is are only connected to corresponding C_i

b) We run MH Sampler a long time and collect M Samples $(C_1[m], \dots, C_n[m])$ for $m = 1, \dots, M$ after mixing

$$P(C_i \mid v_1, \dots, v_n) = P(C_i = k \mid v_1, \dots v_n) \quad \text{for } k = 1, \dots, K$$

Let $g(C_i) = \mathbb{1}[C_i = k]$

$$\hat{g}(C_i[1], \dots, C_i[M]) = \frac{1}{M} \sum_{m=1}^{M} g(C_i[m])$$

c) For Gibbs Sampling to work, it must be true that

$$A(c' \mid c) = 1 = \min(1, 1)$$

$$= \frac{P(v_i \mid C_i = j) \cdot P(v_j \mid C_j = i)}{P(v_i \mid C_i = i) \cdot P(v_j \mid C_j = j)}$$

This will not work, because there is no guarantee this quantity $= 1$

b) for C_i

$g(C) = 1$ ... $P(C_i = k)$

$g(t) = 1[C_i = k]$

$$\hat{g}(m_1, \ldots, m_M) = \frac{1}{M} \sum_{m=1}^{M} g(C[m])$$

$P(C_i = k \mid v_1, \ldots, v_N)$ for $k = 1, \ldots, K$

Let $g(C_i) = 1[C_i = k]$

$$\hat{g}(C_i[1], \ldots, C[M]) = \frac{1}{M} \sum_{m=1}^{M} g(C[m])$$

c) For Gibbs Sampling to work, it must be true that

$$A(c' \mid c) = 1 = \min(1, 1) =$$

$$\rightarrow \frac{P(v_1 \mid C_1 = 1) \cdot P(v_2 \mid C_2 = 2)}{P(v_1 \mid C_1 = 2) \cdot P(v_2 \mid C_2 = 1)} = 1$$

This will not work because there is no guarantee this quantity $== 1$

2b)

$$\frac{\partial}{\partial x} f(x,y), \quad \underbrace{f(x,y) = xy}_{= y},$$

$$\frac{\partial}{\partial \theta_i} J(\theta; D) = \frac{1}{|D|} \sum_{x,y \in D} f_i(x,y) - \overset{(1-\alpha)}{\frac{1}{|D|} \sum_{x \in D} E_\theta[f_i(x,y)]} - \overset{(\alpha)}{\frac{1}{|D|} \sum_{y \in D} E_\theta[f_i(x,y)]}$$

if Q = empirical dist:

if $Q = P_\theta(W \mid Z = z)$:

2) Multi-conditional Parameter Learning, Markov Networks

$D = \{(x^1, y^1), \ldots, (x^M, y^M)\}$

model has parameters $\Theta = [\theta_1, \ldots, \theta_n]$, estimated w/:

$$J(\Theta; D) = (1-\alpha)\, \ell_{Y|X}(\Theta; D) + \alpha\, \ell_{X|Y}(\Theta; D)$$

$\ell_{X|Y}(\Theta; D)$ is conditional log-likelihood of D using $P_\Theta(X \mid Y)$ defined by Markov Network w/ parameters Θ. Same w/ $\ell_{Y|X}$

We have n features $f_i(X_i, Y_i)$ where X_i, Y_i are possibly empty subsets of X, Y respectively

a) Write full objective function $J(\theta; D)$ in terms of all f_i and θ_i.

$$\ell_{Y|X}(\Theta; D) = \frac{1}{|D|} \log p(D, \theta) = \frac{1}{|D|} \sum_{x,y \in D} \Theta^T f(x,y) - \frac{1}{|D|} \sum_{x \in D} \log Z(x, \Theta)$$

$$\ell_{X|Y}(\Theta; D) = \frac{1}{|D|} \sum_{x,y \in D} \Theta^T f(x,y) - \frac{1}{|D|} \sum_{y \in D} \log Z(y, \Theta)$$

$$J(\Theta; D) = (1-\alpha)\left[\frac{1}{|D|} \sum_{x,y \in D} \Theta^T f(x,y) - \frac{1}{|D|} \sum_{x \in D} \log Z(x, \Theta)\right]$$
$$+ \alpha\left[\frac{1}{|D|} \sum_{x,y \in D} \Theta^T f(x,y) - \frac{1}{|D|} \sum_{y \in D} \log Z(y, \Theta)\right]$$
$$= (1-\alpha)\left[\frac{1}{|D|} \sum_{x,y \in D} \Theta^T f(x,y)\right] - (1-\alpha)\left[\frac{1}{|D|} \sum_{x \in D} \log Z(x, \Theta)\right] + //$$
$$= 1 \cdot \left[\frac{1}{|D|} \sum_{x,y \in D} \Theta^T f(x,y)\right] - (1-\alpha)\left[\frac{1}{|D|} \sum_{x \in D} \log Z(x, \Theta)\right] - \alpha\left[\frac{1}{|D|} \sum_{y \in D} \log Z(y, \Theta)\right]$$

where $Z(X, \Theta) = \sum_{y \in D} \Theta^T f(x,y)$, $Z(y, \Theta) = \sum_{x \in D} \Theta^T f(x,y)$

b) Derive $\frac{\partial}{\partial \theta_i} \ell(\theta; D)$

$$\frac{\partial}{\partial \theta_i} \ell(\theta; D) = \frac{1}{|D|} \sum_{(x,y) \in D} f_i(x,y) - (1-\alpha) \cdot \left[\frac{1}{|D|} \sum_{x \in D} E_Q[f_i(x,y)] \right]$$

$$- \alpha \cdot \left[\frac{1}{|D|} \sum_{y \in D} E_Q[f_i(x,y)] \right]$$

If Q is the empirical distribution of our dataset $\hat{P}$, or a conditional distribution of the form $P_\theta(W \mid Z = z)$, this E_Q will apply.

M-Step 1:

$$\Theta_c' = \frac{\sum \text{weight}(c)}{\#\text{datapoints}} = \frac{\frac{1}{M \cdot |Val(C)|}}{M \cdot |Val(C)|} = \frac{1}{|Val(C)|^2}$$

$$\Theta_{x_i|c}' = \frac{\overbrace{\#\text{datapoints where } X_i = x_j \text{ and } C = c}^{\text{denote as } t_{ij}} \cdot \text{weight}}{\#\text{datapoints where } C = c \cdot \text{weight}}$$

$$= \frac{t_{ij} \cdot \frac{1}{|Val(C)|}}{M \cdot \frac{1}{|Val(C)|}} = \frac{t_{ij}}{M}, \text{ where } t_{ij} \leq M$$

To show EM has converged, we will run one more E-step:

E-Step 2: $w = \frac{1}{Z} \cdot \frac{1}{|Val(C)|^2} \cdot \prod_{i=1}^{n} \frac{t_{ij}}{M} = \frac{1}{Z} \cdot \frac{1}{|Val(C)|^2} \cdot \frac{1}{M} \cdot \prod_{i=1}^{n} t_{ij}$

$$Z = \sum_{h=1}^{|Val(C)|} \frac{1}{|Val(C)|^2} \cdot \frac{1}{M} \cdot \prod_{i=1}^{n} t_{ij} = \frac{1}{|Val(C)|} \cdot \frac{1}{M} \cdot \prod_{i=1}^{n} t_{ij}$$

$$w = \frac{\frac{1}{|Val(C)|^2} \cdot \frac{1}{M} \cdot \prod_{i=1}^{n} t_{ij}}{\frac{1}{|Val(C)|} \cdot \frac{1}{M} \cdot \prod_{i=1}^{n} t_{ij}} = \frac{|Val(C)|}{|Val(C)|^2} = \frac{1}{|Val(C)|}$$

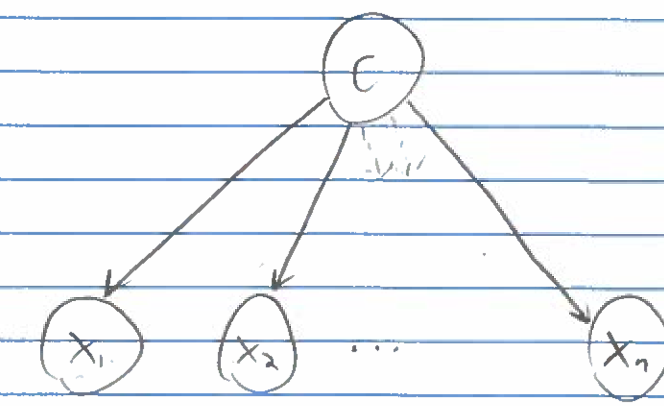
Since the value for all weights is the same as the previous E-step, EM has converged.

The final parameter values are:

$$\Theta_c = \frac{1}{|Val(C)|^2}$$

$$\Theta_{x_i|c} = \frac{t_{ij}}{M}$$

3) Expectation Maximization in a Naive Bayes Model



Class variable C, discrete evidence variables $X_1, \dots, X_n$

CPDs parametrized: $P(C=c) = \Theta_c$, $P(X_i = x \mid C=c) = \Theta_{x_i \mid c}$ for $i = 1, \dots, n$
and for all assignments $x_i \in Val(X_i)$, classes $c \in Val(C)$

Given dataset $D = \{x[1], \dots, x[M]\}$ each $x[m]$ assignment to $X_1, \dots, X_n$
C not observed

E-Step 1: $X[m] = (X_1 = x_1, X_2 = x_2, \dots, X_n = x_n, C = c_1)$
$(\quad '' \quad , C = c_2)$
$\vdots$
$(\quad '' \quad , C = c_{|Val(C)|})$

Expand $x[m]$ into $|Val(C)|$ datapoints, one for each $c \in C$

$$w = \frac{1}{Z} \Theta_c^{\circ} \prod_{i=1}^{n} \Theta_{x_i \mid c}^{\circ} = \frac{1}{Z} \cdot \frac{1}{|Val(C)|} \cdot \prod_{i=1}^{n} \frac{1}{|Val(X_i)|}$$

$$Z = \sum_{j=1}^{|Val(C)|} \frac{1}{|Val(C)|} \cdot \prod_{i=1}^{n} \frac{1}{|Val(X_i)|} = \prod_{i=1}^{n} \frac{1}{|Val(X_i)|}$$

$$w = \frac{\frac{1}{|Val(C)|} \cdot \prod_{i=1}^{n} \frac{1}{|Val(X_i)|}}{\prod_{i=1}^{n} \frac{1}{|Val(X_i)|}} = \frac{1}{|Val(C)|}$$

Weight w is same for all datapoints

M-step:
$$\Theta_{c_1}' = \frac{\sum weight(c_1)}{\# \text{ data}} = \frac{M \cdot \frac{1}{|Val(C)|}}{M \cdot |Val(C)|} = \frac{1}{|Val(C)|^2}$$

$$\Theta_{X_i = x_i \mid C = c_1}' = \frac{n_x M \cdot \frac{1}{|Val(C)|}}{M \cdot \frac{1}{|Val(C)|}} = \frac{n_x}{M}, \quad n_x M \rightarrow n_x t$$

E-Step 2:
$$w = \frac{1}{Z} \cdot \frac{1}{|Val(C)|^2} \cdot \prod_{i=1}^{n} \frac{t_i}{M} = \frac{1}{Z} \cdot \frac{1}{|Val(C)|^3} \cdot \frac{1}{M} \cdot \prod_{i=1}^{n} t$$

$$Z = |Val(C)| \cdot \frac{1}{|Val(C)|^3} \cdot \frac{1}{M} \cdot \prod_{i=1}^{n} t_i =$$
$$= \frac{1}{|Val(C)|} \cdot \frac{1}{M} \cdot \prod_{i=1}^{n} t$$

$$\rightarrow w = \frac{\frac{1}{|Val(C)|^3} \cdot \frac{1}{M} \cdot \prod_{i=1}^{n} t}{\frac{1}{|Val(C)|} \cdot \frac{1}{M} \cdot \prod_{i=1}^{n} t_i} = \frac{|Val(C)|}{|Val(C)|^3} = \frac{1}{|Val(C)|}$$

a) Derive expression for cond. prob pixel (i,j) black given MB

$$E(y,x) = -\eta \sum_{i,j} y_{i,j} x_{i,j} - \beta \sum_{((i,j),(i',j')) \in E} y_{i,j} y_{i',j'}$$

$$p(y,x) = \frac{1}{Z} \exp(-E(y,x)) \quad , \text{ Let } y_{-\ell} = y_{-y_{i,j}}$$

$$p(y_{i,j}=1 \mid y_{m}(i,j)) = p(y_{i,j}=1 \mid y_{-\ell}, X)$$

$$= \frac{p(y_{i,j}=1, y_{-\ell}, X)}{p(y_{-\ell}, X)}$$

$$= \frac{\exp(-E(y_{i,j}=1, y_{-\ell}, X))}{\exp(-E(y_{i,j}=1, y_{-\ell}, X)) + \exp(-E(y_{i,j}=-1, y_{-\ell}, X))}$$

$$= \frac{\exp\left(\eta\left(\sum_{i,j} y_{i,j} x_{i,j} + x_\ell\right) + \beta\left(\sum_{((i,j),(i',j')) \in E} y_{i,j} y_{i',j'} + \sum_{y_m} y_{i',j'}\right)\right)}{// + \exp\left(\eta\left(\sum_{i,j} y_{i,j} x_{i,j} - x_\ell\right) + \beta\left(\sum_{\in E} y_{i,j} y_{i',j'} - \sum_{y_m} y_{i',j'}\right)\right)}$$

$$\eta \sum_{i,j} y_{i,j} x_{i,j} + \eta x_\ell + \beta \sum_{\in E} y_{i,j} y_{i',j'} + \beta \sum_{y_m} y_{i',j'}$$

$$\rightarrow \exp\left(\eta x_\ell + \beta \sum_{y_m} y_{i',j'}\right) \cdot \exp\left(\eta \sum_{i,j} y_{i,j} x_{i,j} + \beta \sum_{\in E} y_{i,j} y_{i',j'}\right)$$

Let $a = \eta x_\ell + \beta \sum_{y_m} y_{i',j'}$, $b = \eta \sum_{i,j} y_{i,j} x_{i,j} + \beta \sum_{\in E} y_{i,j} y_{i',j'}$

Then $p(y_{i,j}=1 \mid y_m(i,j)) = \dfrac{e^a \cdot e^b}{e^a \cdot e^b + e^{-a} \cdot e^b} = \dfrac{e^a \cdot e^b}{e^b \cdot (e^a + e^{-a})}$

$$= \frac{e^a}{e^a + e^{-a}} \cdot \frac{e^{-a}}{e^{-a}} = \frac{1}{1 + e^{-2a}} = \sigma(2a)$$

$$= \sigma\left(2\left(\eta x_\ell + \beta \sum_{y_m} y_{i',j'}\right)\right)$$

CS228: Probabilistic Graphical Models

Homework 4

Luke Jaffe

Due: 03/03/2017
Submitted: 03/03/2017

Problem 4: Programming Assignment

a) **TODO:** rewrite and scan

b)

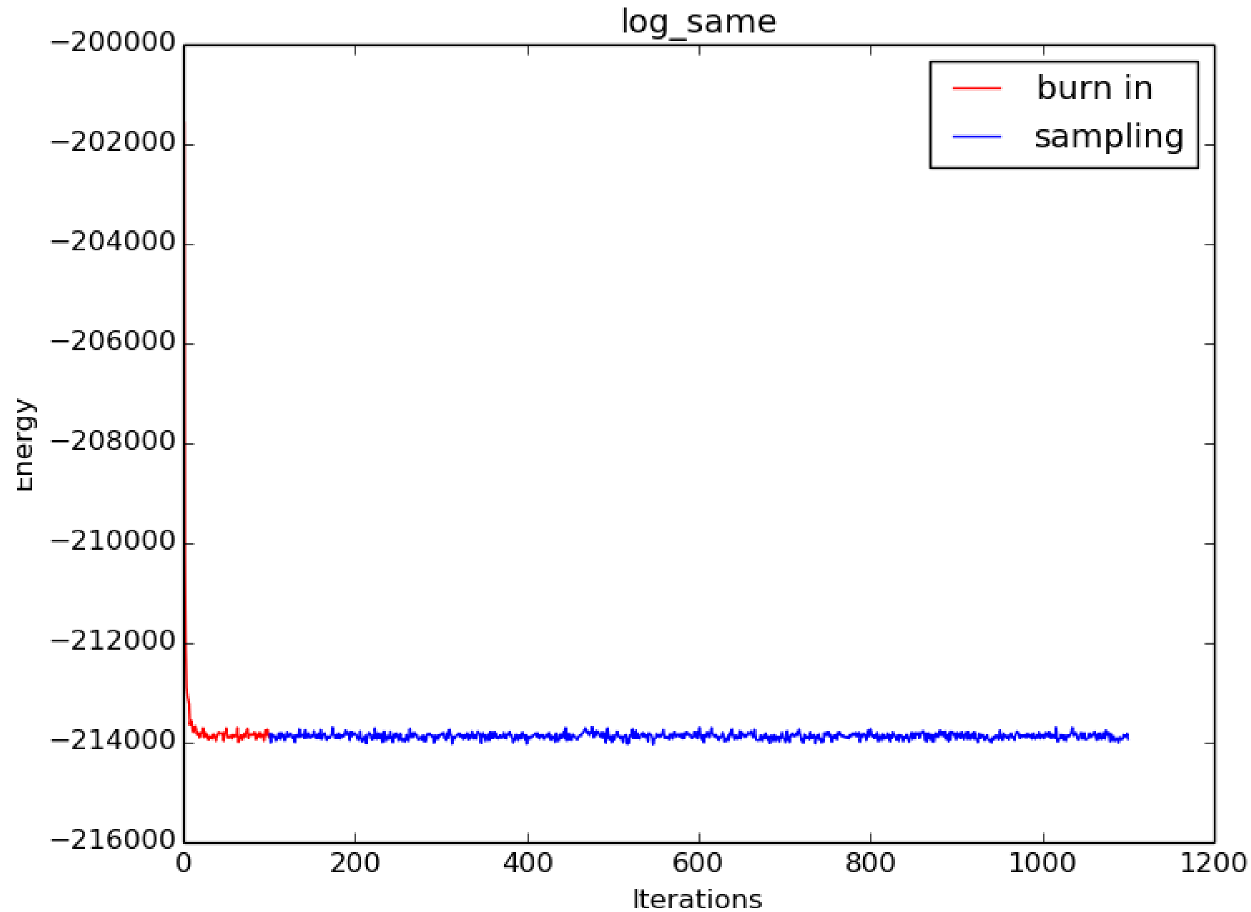
i. Outline a Gibbs sampling algorithm (in pseudocode) that iterates over the pixels in the image and samples each y_{ij} given its Markov Blanket.

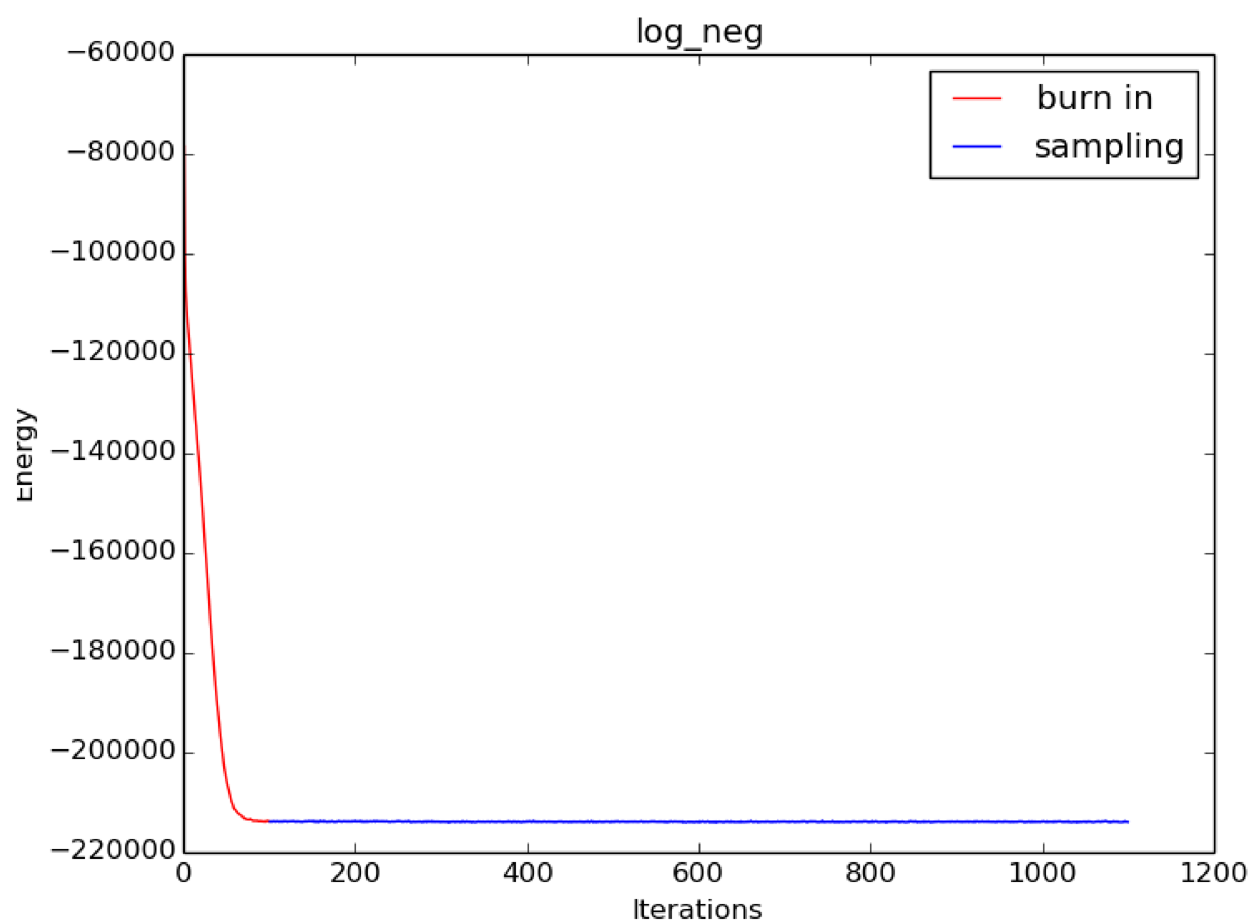
```
given constants eta, beta

function sample(Y, X):
    initialize matrix X to the noisy image
    initialize matrix Y randomly (same size as X)
    for I = 1 to N:
        for j = 1 to M:
            x_term = eta*X[i][j]
            y_term = 0
            for y in markov_blanket(Y, i, j):
                y_term += y
            y_term *= beta
            a = x_term + y_term
            s = sigmoid(2*a)
            Y[i][j] = s

function gibbs(Y, X, B, N):
    for t = 1 to B:
        sample(Y, X)
    for t = 1 to N
        sample(Y, X)
    return Y
```

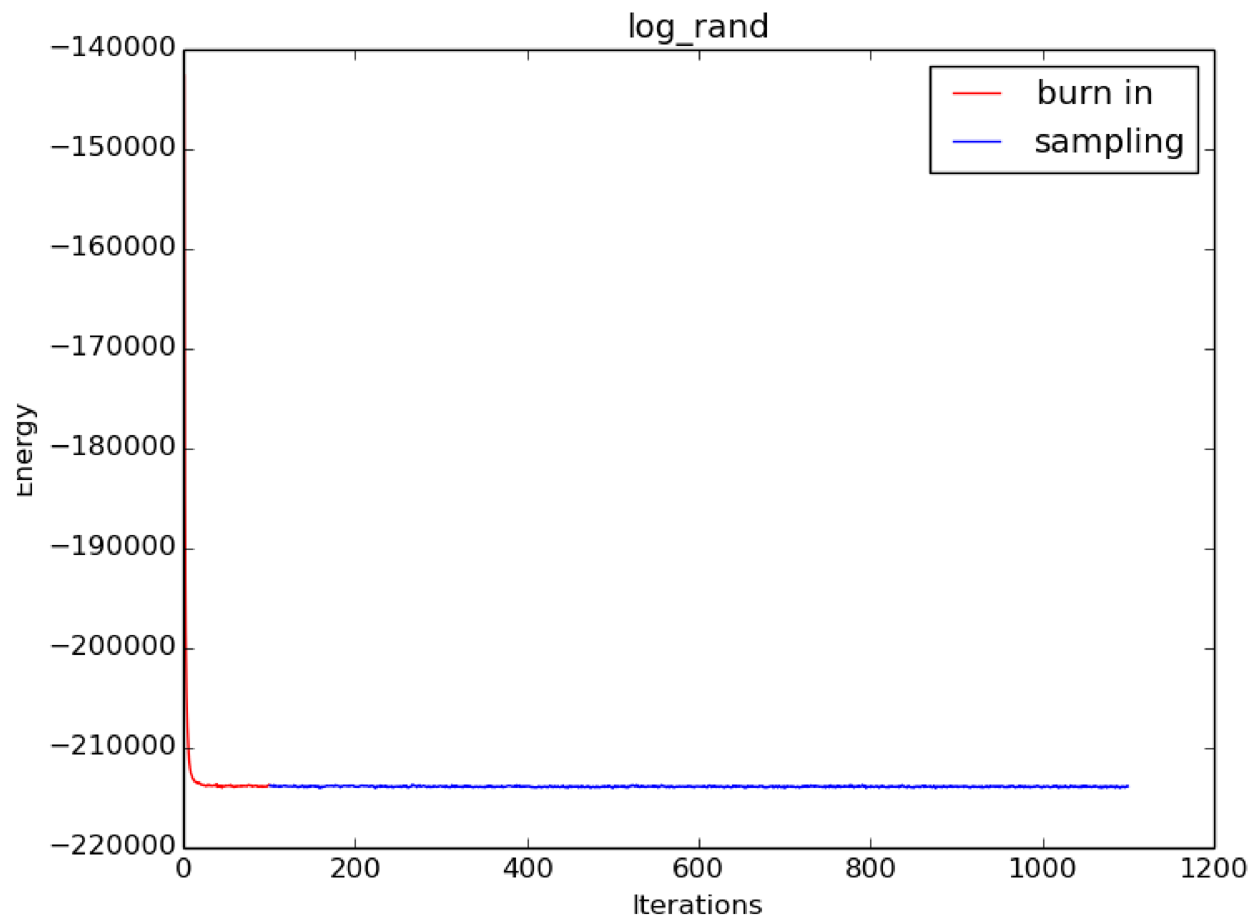
ii. How can we show in our case that the equilibrium distribution is in fact the posterior distribution p(y|x)?

We can show this using the Monte Carol equation.

c)


log_same
burn in
sampling
Energy
Iterations
−200000
−202000
−204000
−206000
−208000
−210000
−212000
−214000
−216000
0
200
400
600
800
1000
1200


log_neg
burn in
sampling
Energy
Iterations
−60000
−80000
−100000
−120000
−140000
−160000
−180000
−200000
−220000
0
200
400
600
800
1000
1200



i. Do all three seem to be converging to the same general region of the posterior, or are some obviously suboptimal?

While all three methods do not converge in exactly the same way, they all converge to the same general region of the posterior after some iterations. The negative initialization may be considered suboptimal since it takes longer to converge than the same or random initialization methods.

ii. Does the burn-in seem to be adequate in length?

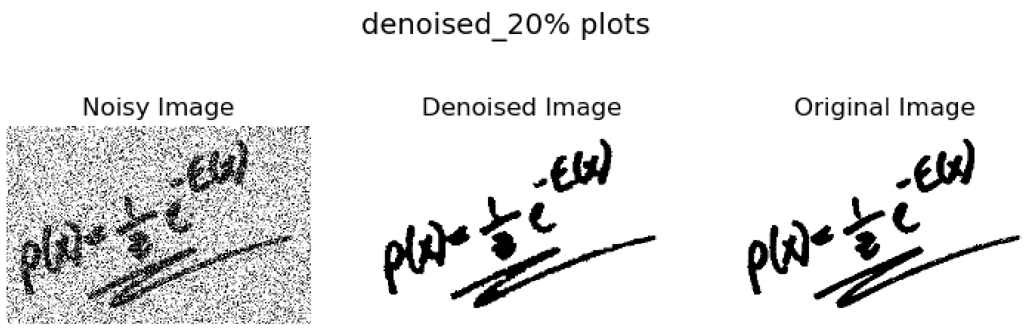
Yes, the burn-in is more than sufficient in length. The same and random initializations seem to converge in less than 20 iterations, while the negative initialization needs 80 or so.

iii. Is there substantial fluctuation from iteration to iteration, indicating that the chain is mixing well, or does it become stuck at particular energies for several iterations at a time?

Yes, there is substantial fluctuation between iterations for all methods. In the plots above, the “same” method may appear to fluctuate the most, but this is just because of the y-axis scale.

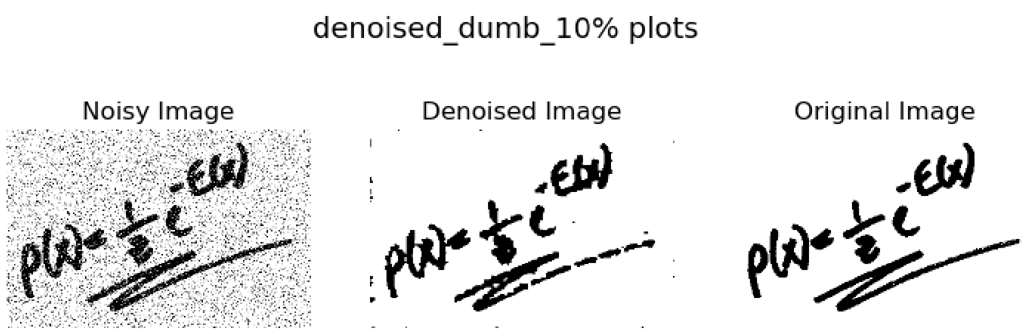
d)

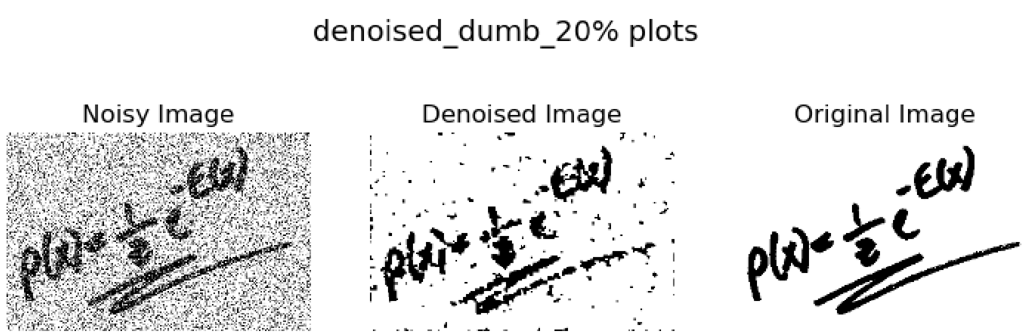
Denoised 10% error: 0.0059
Denoised 20% error: 0.0104



e)

Denoised dumb 10% error: 0.0213
Denoised dumb 20% error: 0.0588

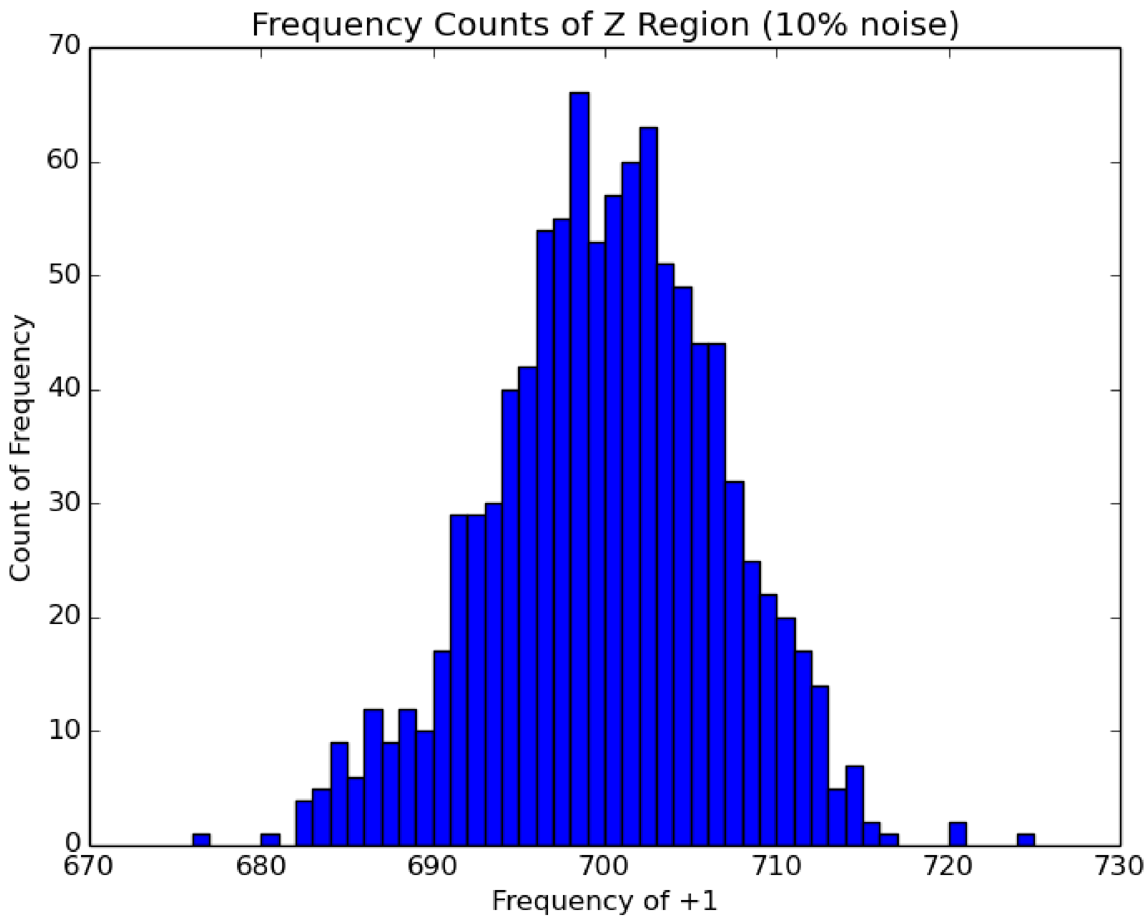


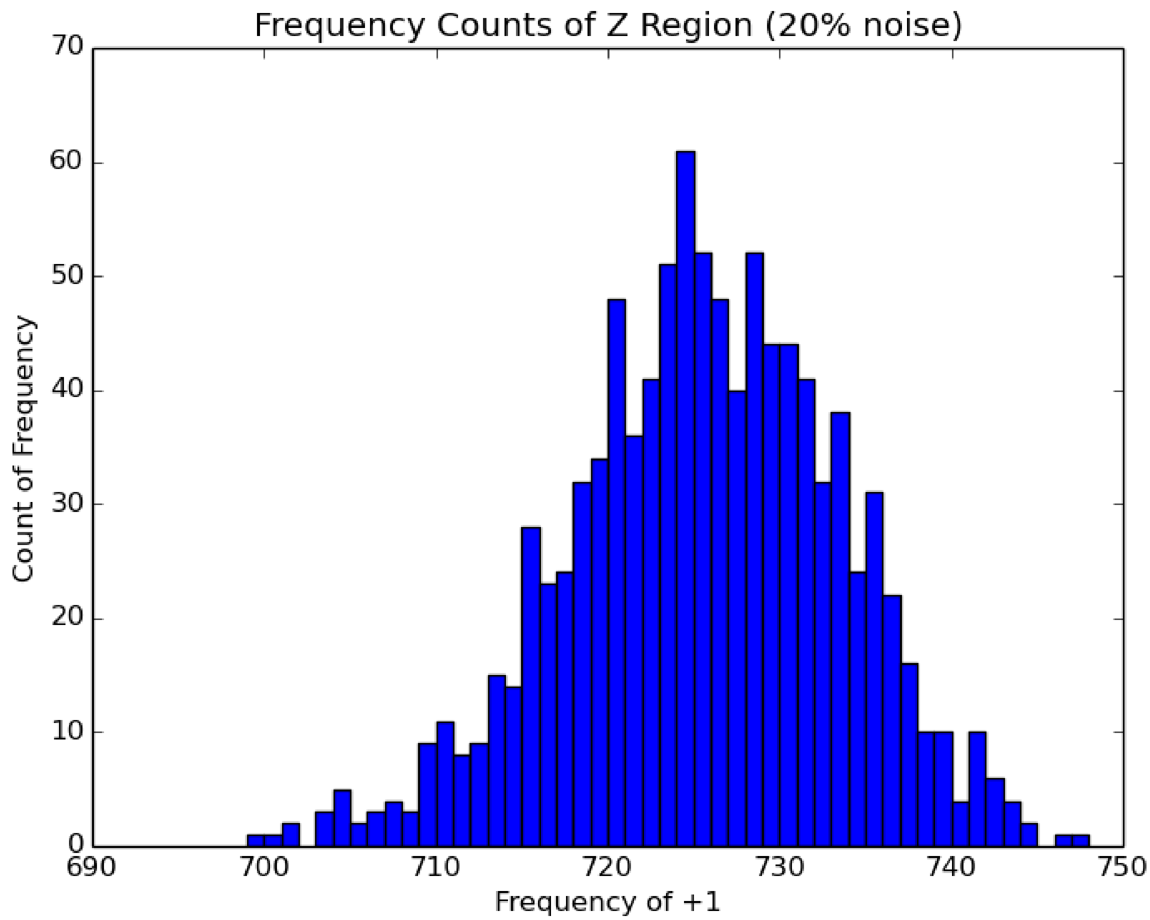


i. Does the Gibbs sampler do better than the trivial algorithm? Why or why not?

Yes, the Gibbs sampler does better than the trivial algorithm. The error is lower in the 10% noise case (0.5% v. 2%) and the 20% noise case (1% v. 6%). The images for the Gibbs sampler are also visibly better denoised. This is because the Gibbs sampler is built on a model which to some degree correctly codifies the spatial locality relationships of the actual imagery, whereas the other method is naive.

f)





i. The distribution of frequencies for the noisier case is slightly wider (greater variance), and has a greater mean (~725 v. ~700).