SHORT SUMMARY:

1. Mutual information is a convex function relative to the input probability vector (proved in the class)

2. Space of vector probability is also a convex space ($\Sigma Pi = 1$ is a plane in n dimensions, also the convexity is proved in Gallager's book page 83)

3. the maximum point could be found by following the direction of mutual information's gradient starting from an arbitrary point in the feasible space of probability vectors.

4. our iterations are in the form of "$x_{k+1} = x_k +$ learning_rate*jacobian"

5. Derivatives of mutual information with respect to input vector weights are calculated in "Gallager's book page 91,92":

   $\partial I(X;Y)/\partial P_i = I(x = k ;Y)-\log 2(e)$ (P_i is the ith weith of input vector)

6. $I(x = k\,; Y) = \sum_j P(j|k) \log \frac{P(j|k)}{\sum_i P_i P(j|i)}$ (Gallager's book page 91)
   (P(j|k) is the element on the kth row and the jth column of the transition matrix of channel)