Low-Complexity Quantization of Discrete Memoryless Channels

Jiuyang (Alan) Zhang
Rakuten, Inc

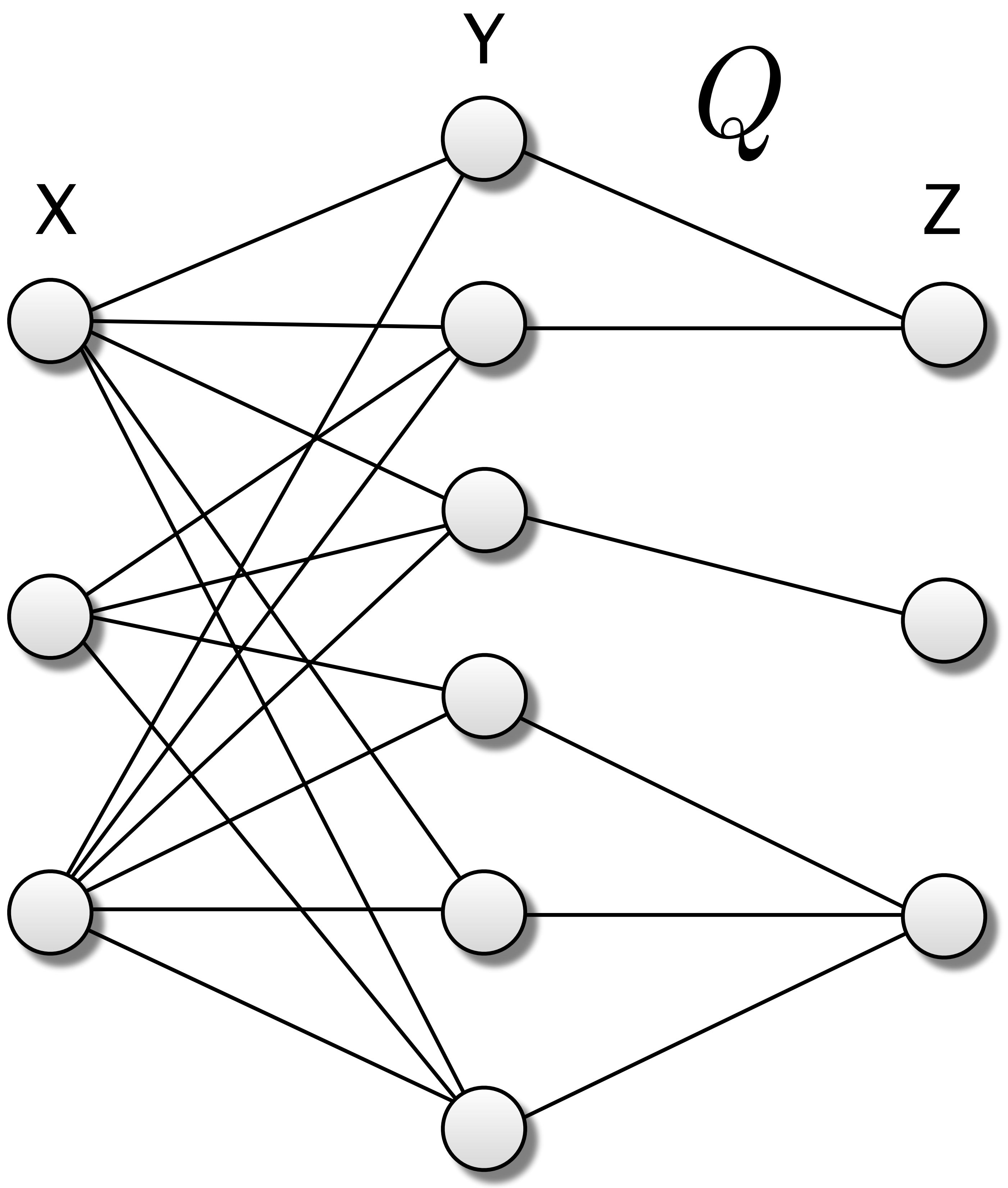
Brian M. Kurkoski
Japan Advanced Institute of Science and Technology

JAPAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY 1990

November 1, 2016
International Symposium on Information Theory and Its Applications
Monterey, California, USA

Introduction

$$\mathsf{X} \xrightarrow{\Pr(\mathsf{Y}|\mathsf{X})} \mathsf{Y} \xrightarrow{Q} \mathsf{Z}$$



Given a joint distribution $p_{\mathsf{XY}}(x, y)$, find the quantizer Q which maximizes mutual information:

$$Q^* = \arg\max_{Q} I(\mathsf{X}; \mathsf{Z})$$

with $K < M$ where:

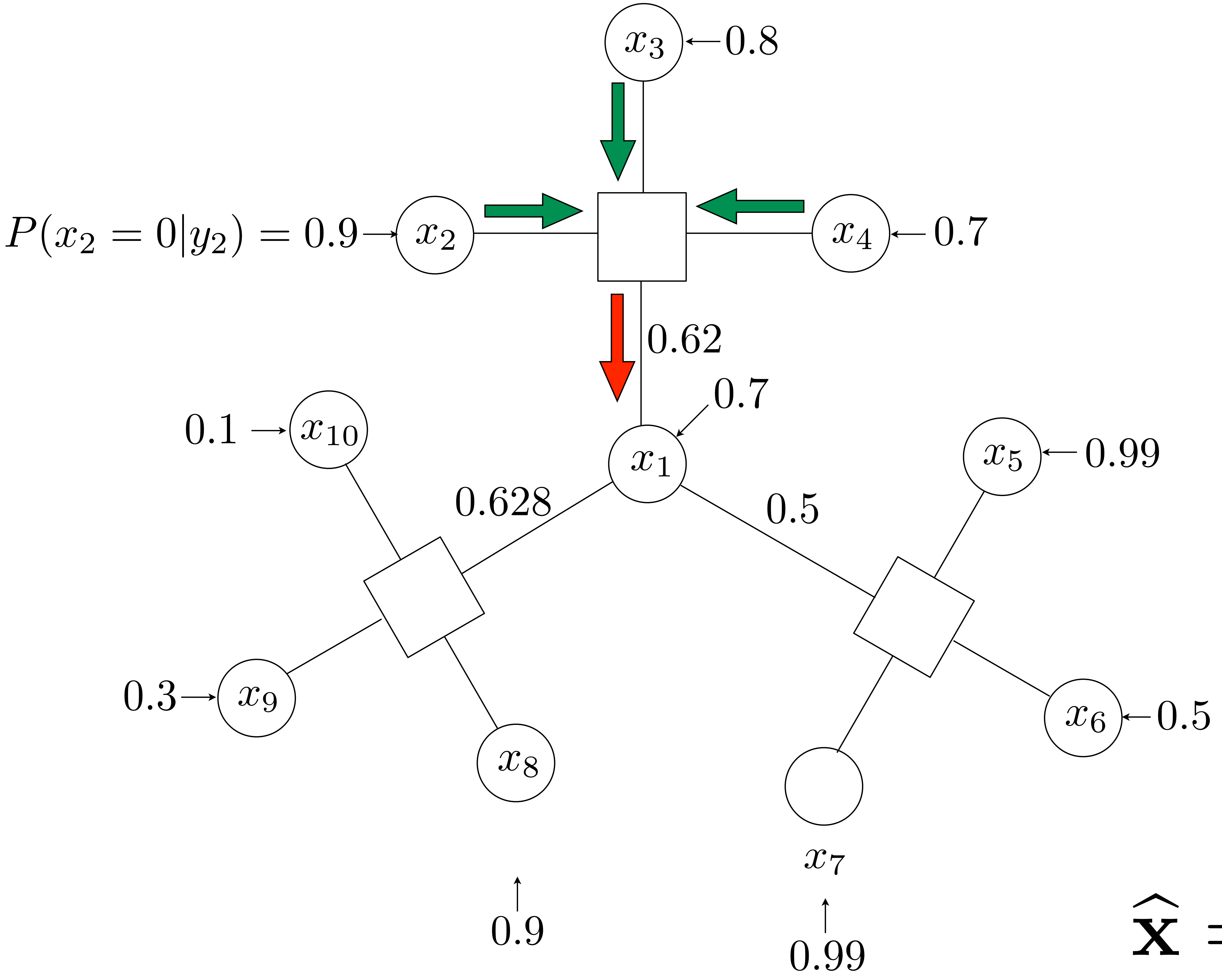
$$M = |\mathcal{Y}|$$
$$K = |\mathcal{Z}|$$

Motivation: Decoder & Detector VLSI

- Information and coding theorists develop elegant decoding algorithms
- Factor-graph decoding algorithms use real numbers
- Algorithms are often implemented in VLSI, particularly for power-sensitive applications like mobile devices, SSDs, etc.
- VLSI has fixed precision, e.g. real numbers might be approximated using 4 to 8 bits
- This fixed-point design is often done in an ad hoc fashion.

Factor Graph Representation of Decoders

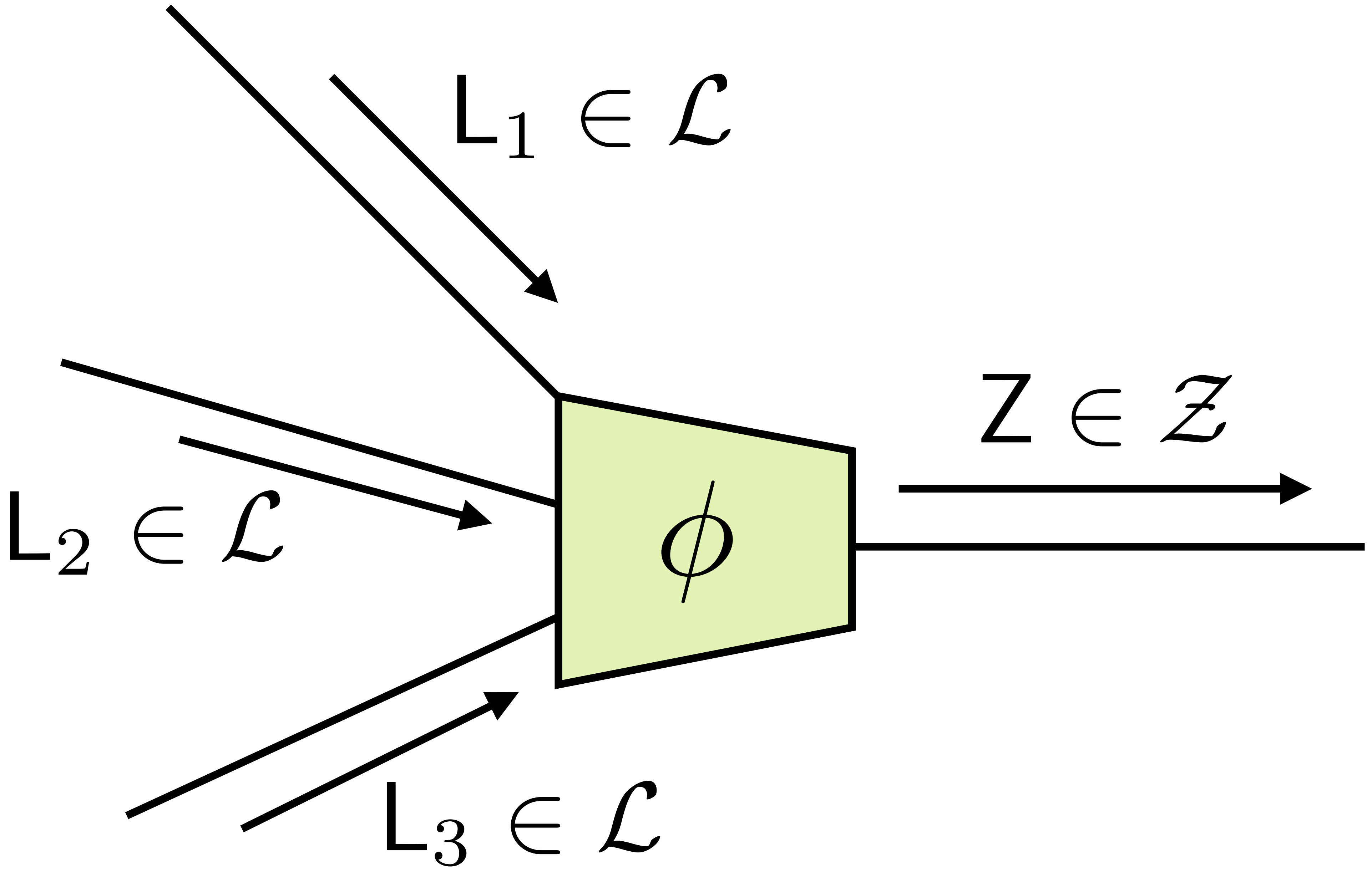
input $\mathbf{y}$
from channel



$\widehat{\mathbf{x}}$ output:
transmitted
codeword
estimate

$$\widehat{\mathbf{x}} = \arg\max_{\mathbf{x}} \Pr(\mathbf{y}|\mathbf{x})$$

Message Passing Decoding on Factor Graphs



$$\phi : \mathcal{L}_1 \times \mathcal{L}_2 \times \mathcal{L}_3 \to \mathcal{Z}$$

Decoding algorithms can be represented as “message passing on a factor graph”

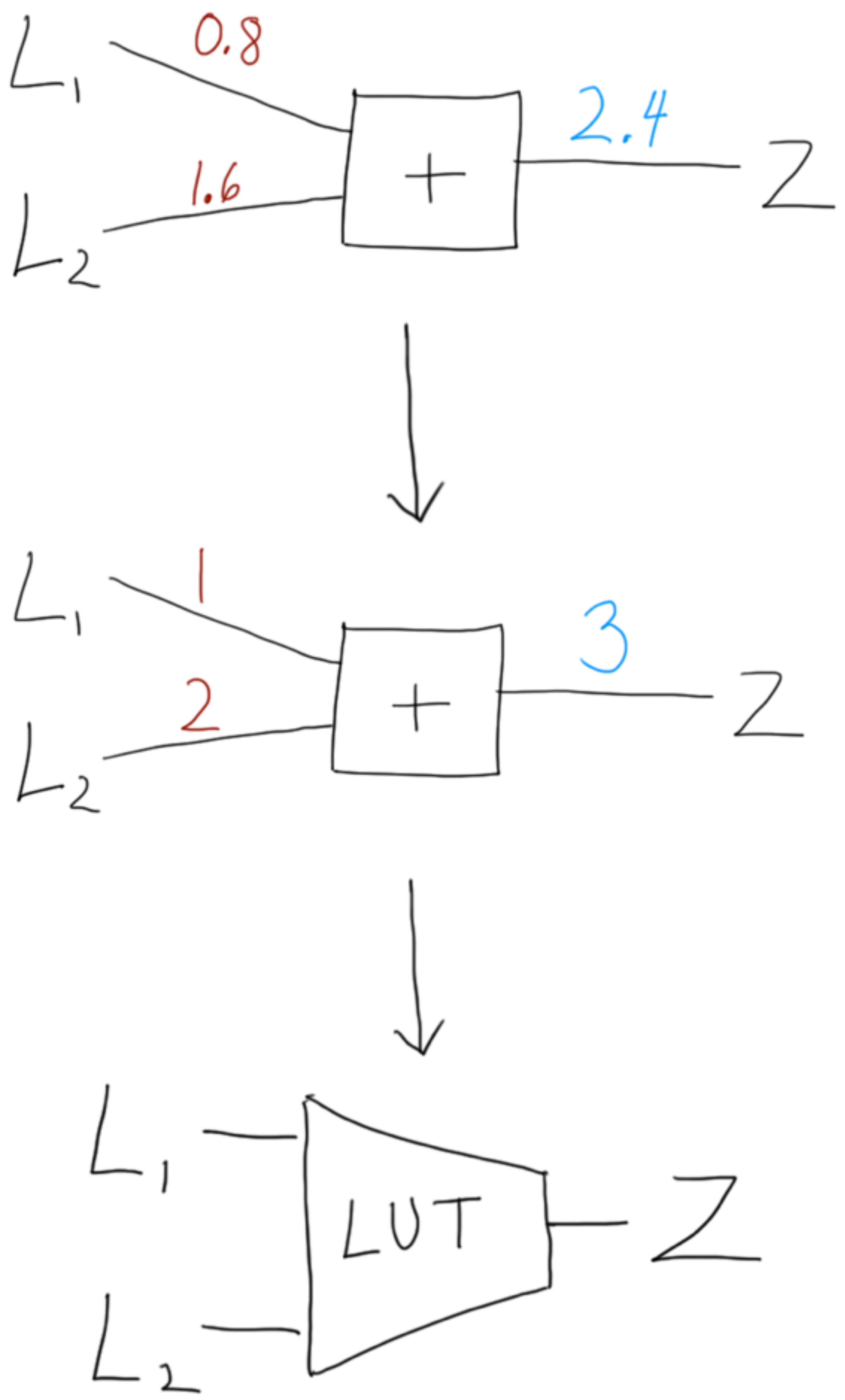
- Edges represent messages
- Nodes represent “local” decoding functions **φ**

Important example:

- messages are probabilities
- decoding functions are derived from basic rules of probability theory

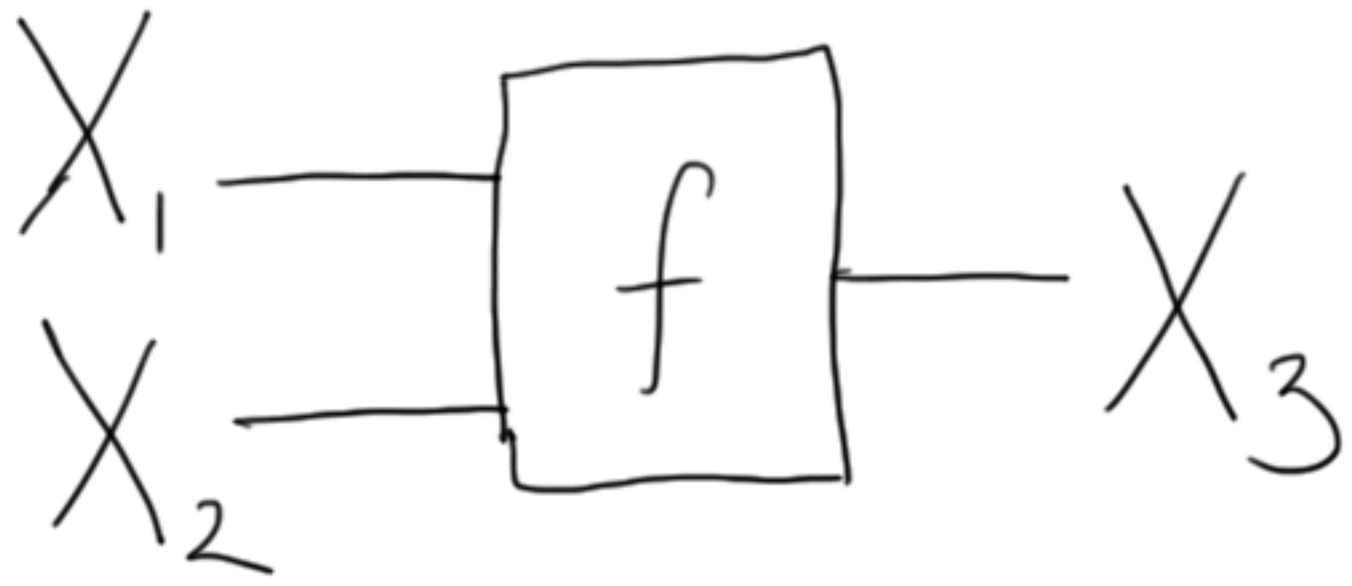
Max-LUT Method

Max-LUT is a method for implementing the node decoder functions for graph-based decoders, using lookup tables that maximize mutual information.



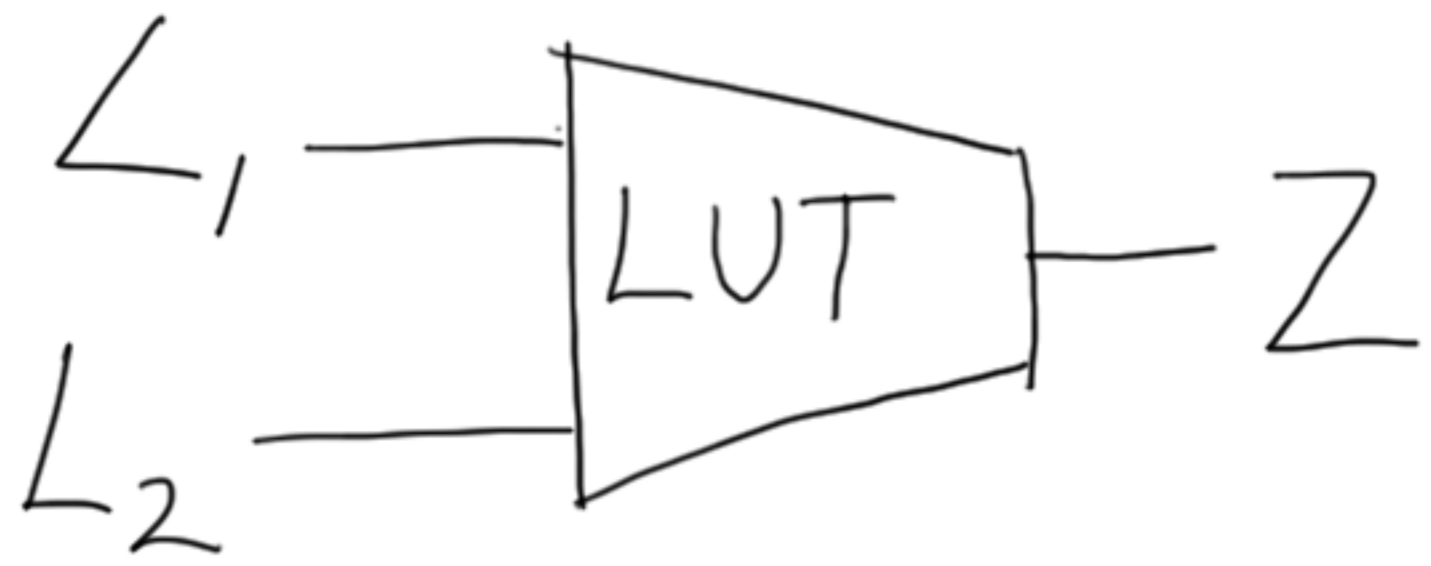
Max-LUT Method: Central Idea

Encoder Side: Code Symbols X_i



- Check node $x_3 = x_1 + x_2$
- Var node $x_1 = x_2 = x_3$
- etc.

Decoder Side



L_i is a noisy version of X_i ,

Z is a noisy version of X_3

Choose LUT to maximize mutual information

$$\max I\big(\mathsf{X}_3; \mathsf{Z}\big) = \max_{\mathrm{LUT}} I\big(\mathsf{X}_3; \mathrm{LUT}(\mathsf{L}_1, \mathsf{L}_2)\big)$$

Max-LUT Step 1: Joint Distribution

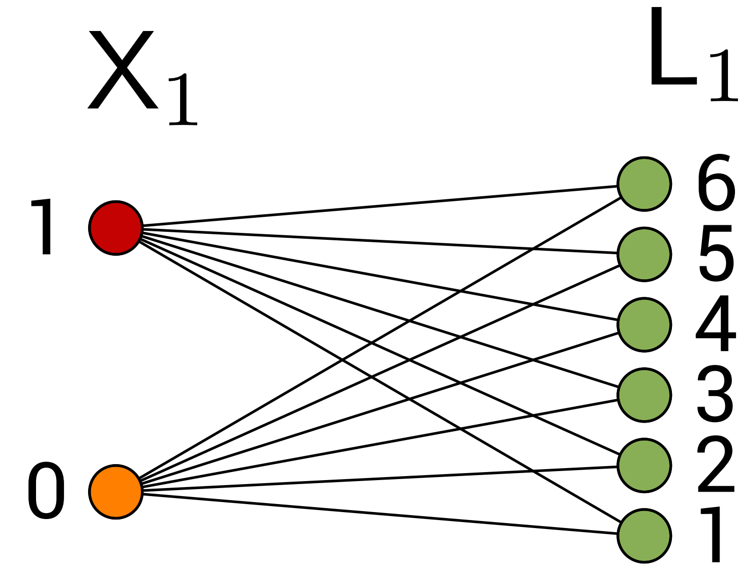
- Step 1: Construct joint distribution

$\Pr(\mathsf{L}_1|\mathsf{X}_1)$

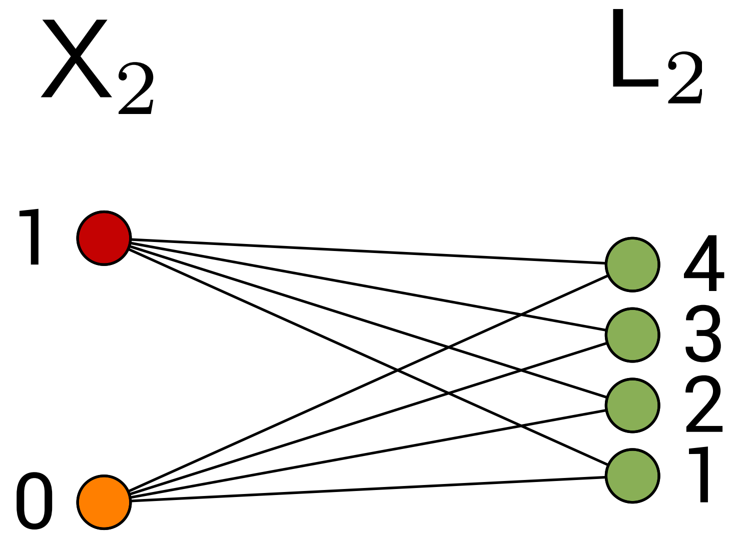
$\mathsf{L}_1 \in \{1,2,3,4,5,6\}$

$\Pr(\mathsf{L}_2|\mathsf{X}_2)$

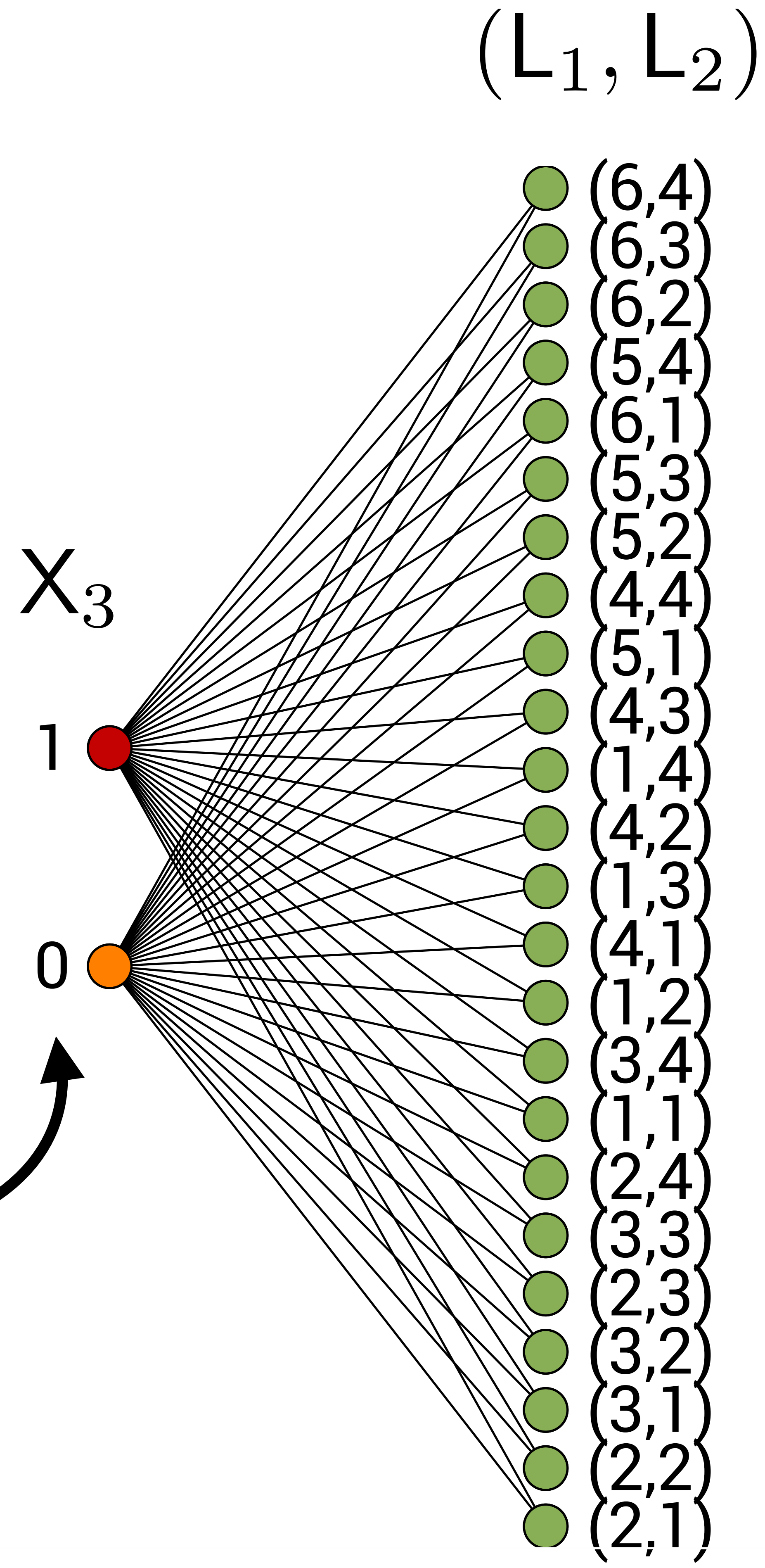
$\Pr(\mathsf{L}_1, \mathsf{L}_2|\mathsf{X}_3)$



$\times$



$=$

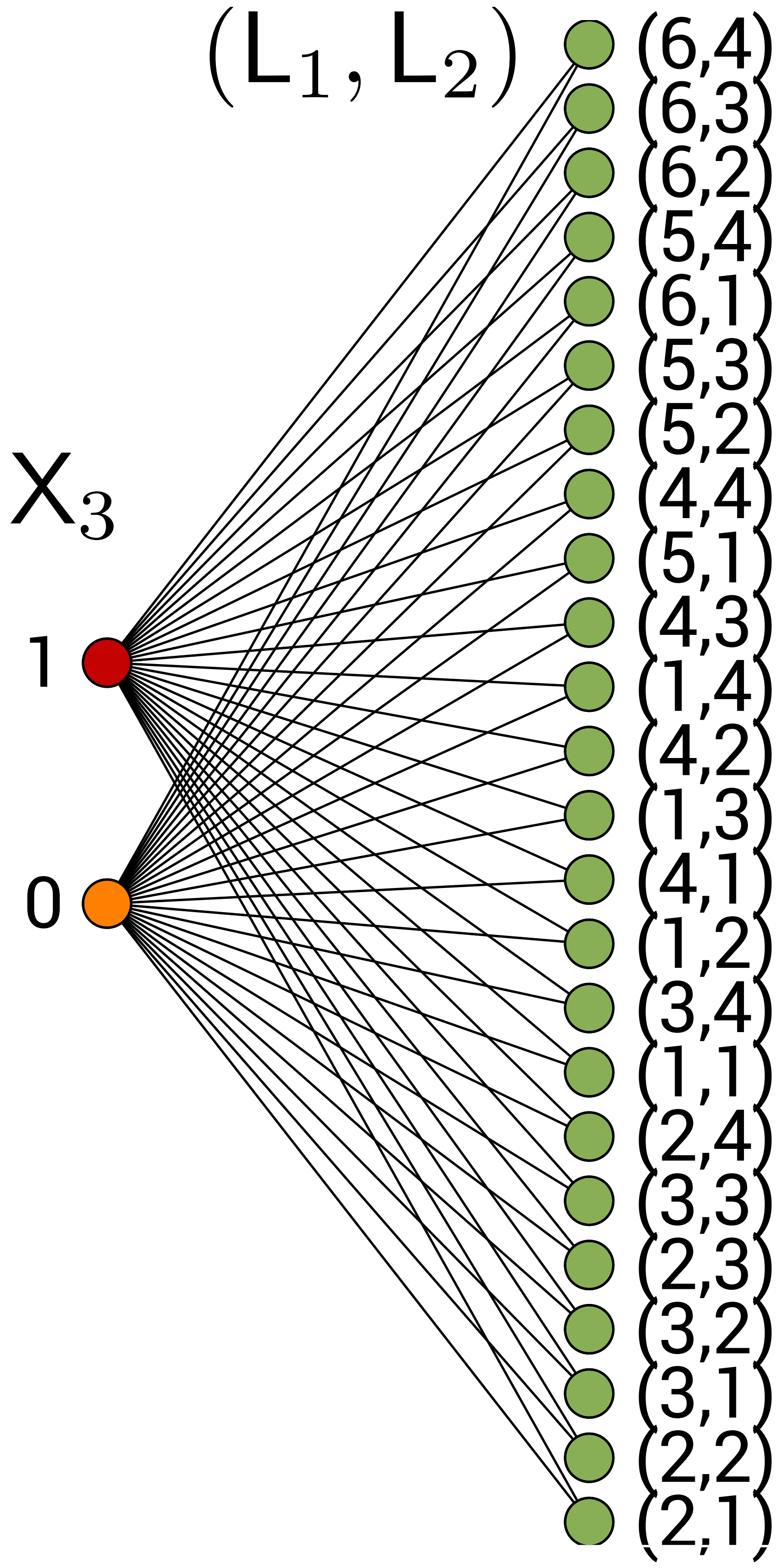


local constraint:

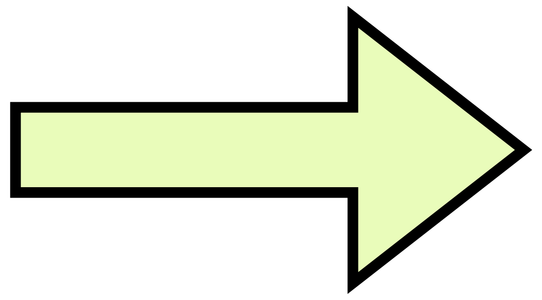
$f(x_1, x_2) = x_3$

Max-LUT Step 2: Quantize

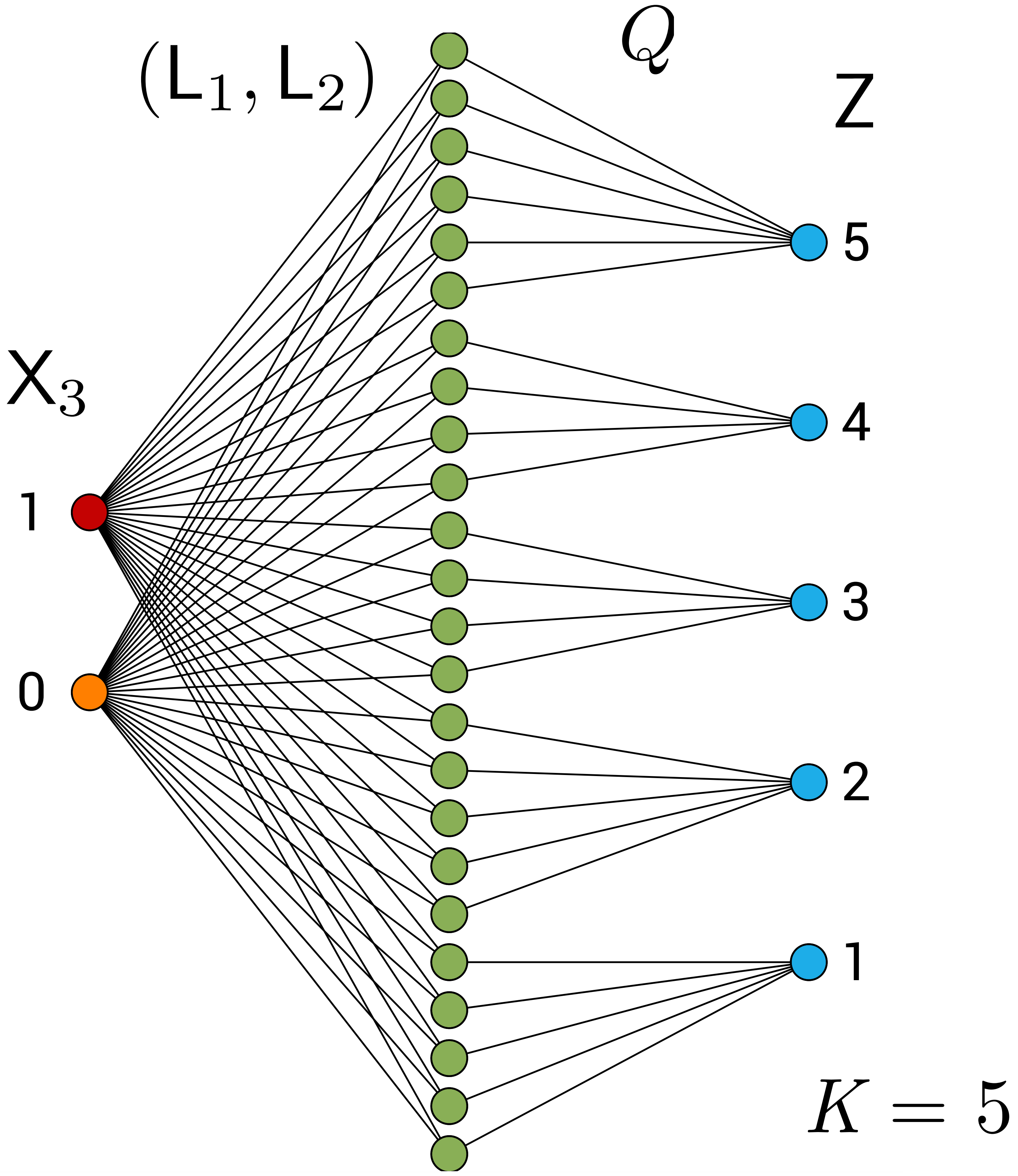
- Too many levels! Reduce to Z with K levels
- Quantizer is a mapping from $(\mathsf{L}_1,\mathsf{L}_2)$ to Z



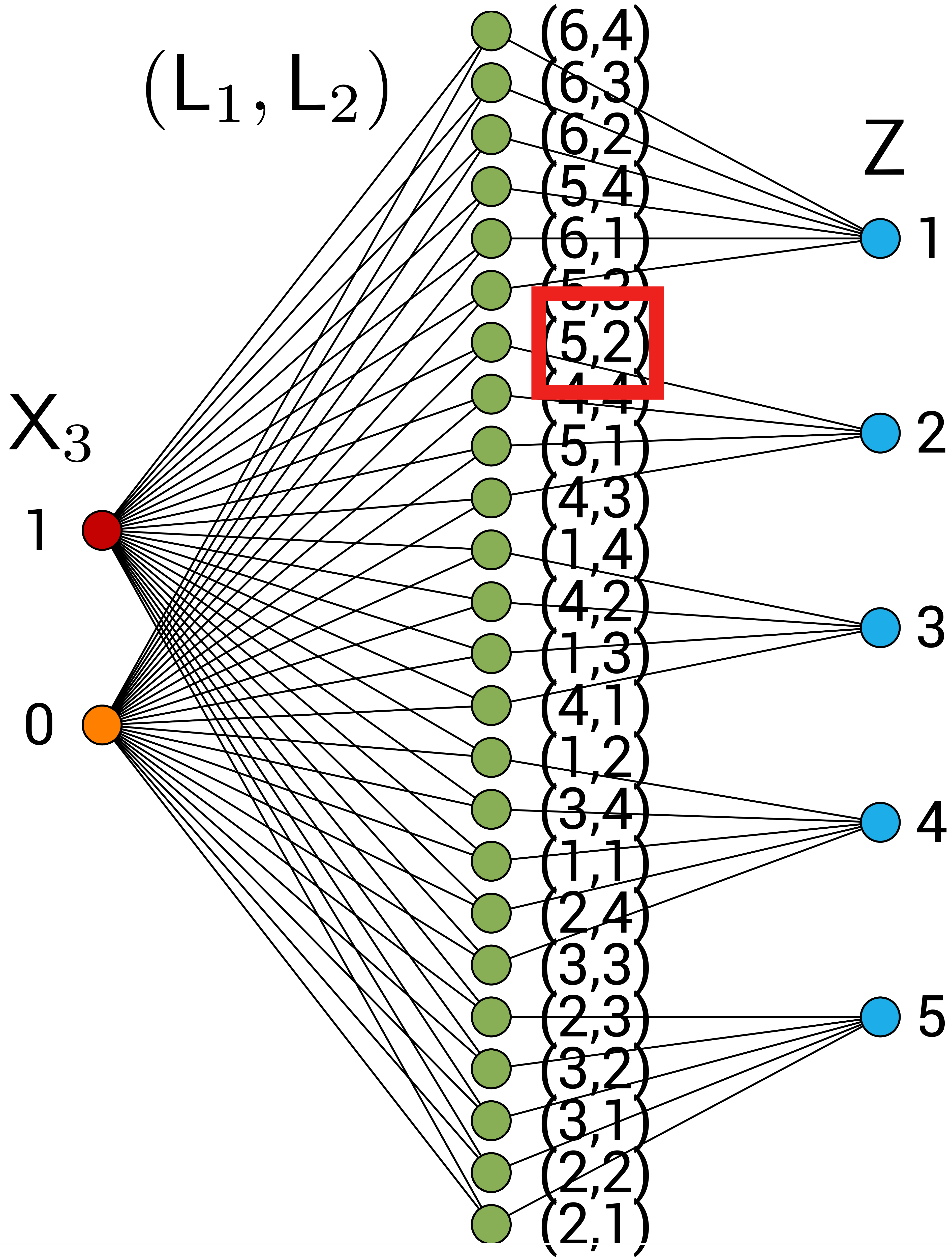
$$\max_{Q} I(\mathsf{X}_3; \mathsf{Z})$$



$$Q : \mathcal{L}_1 \times \mathcal{L}_2 \to \mathsf{Z}$$



Max-LUT Step 3: Lookup Table



L_1 \ L_2	1	2	3	4
1	4	4	3	3
2	5	5	5	4
3	5	5	4	4
4	3	3	2	2
5	2	$\boxed{2}$	1	1
6	1	1	1	1

Lookup table:

$Z = \mathrm{LUT}(L_1, L_2)$

Quantization of DMCs

Global optimization has exponential complexity in general

$J = 2$ binary input, finding optimal quantizer complexity M^3 or better

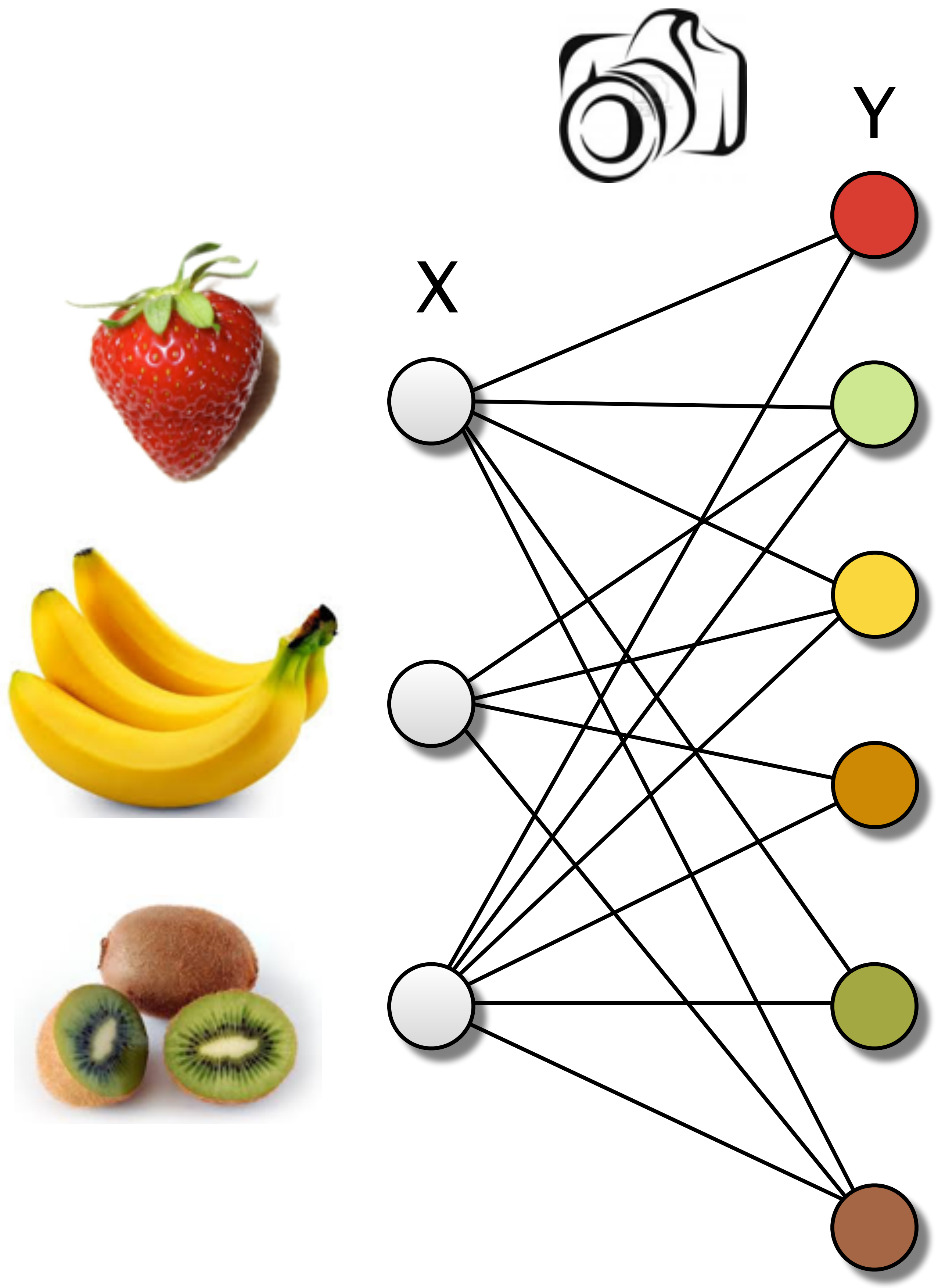
Information theory quantization = machine learning classification

Suboptimal approaches for non-binary inputs

- Greedy combining complexity $M^2(M - K)$
- Information bottleneck method
- **KL-Means algorithm**
  - ✦ Replace K-Means Euclidean distance metric with KL divergence
  - ✦ Quantization in reverse channel Pr[X | Y]
  - ✦ Minimizing KL divergence equivalent to max mutual information
  - ✦ Complexity MKT for T iterations

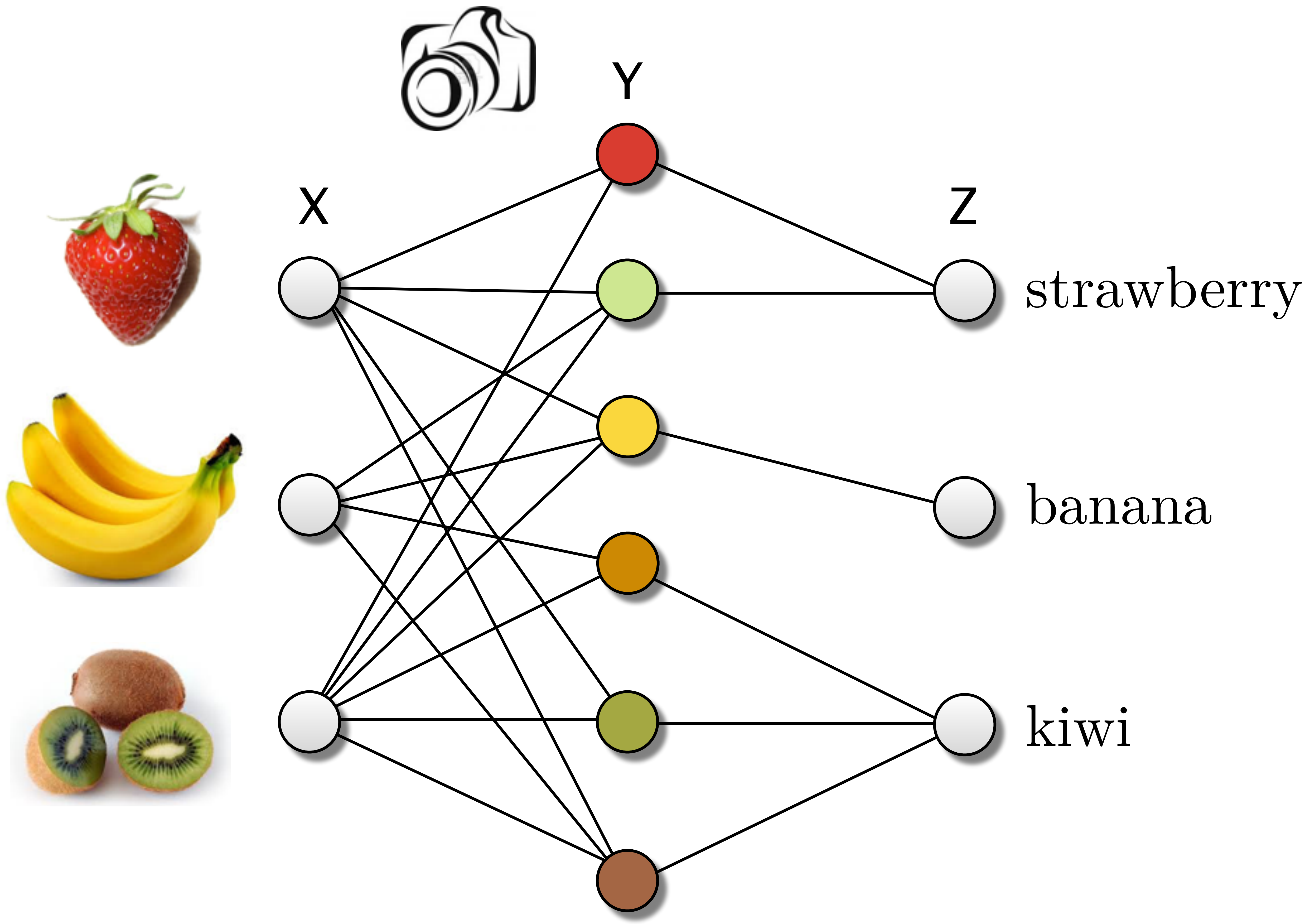
Machine Learning: Classification

$$X \xrightarrow{\Pr(Y|X)} Y$$



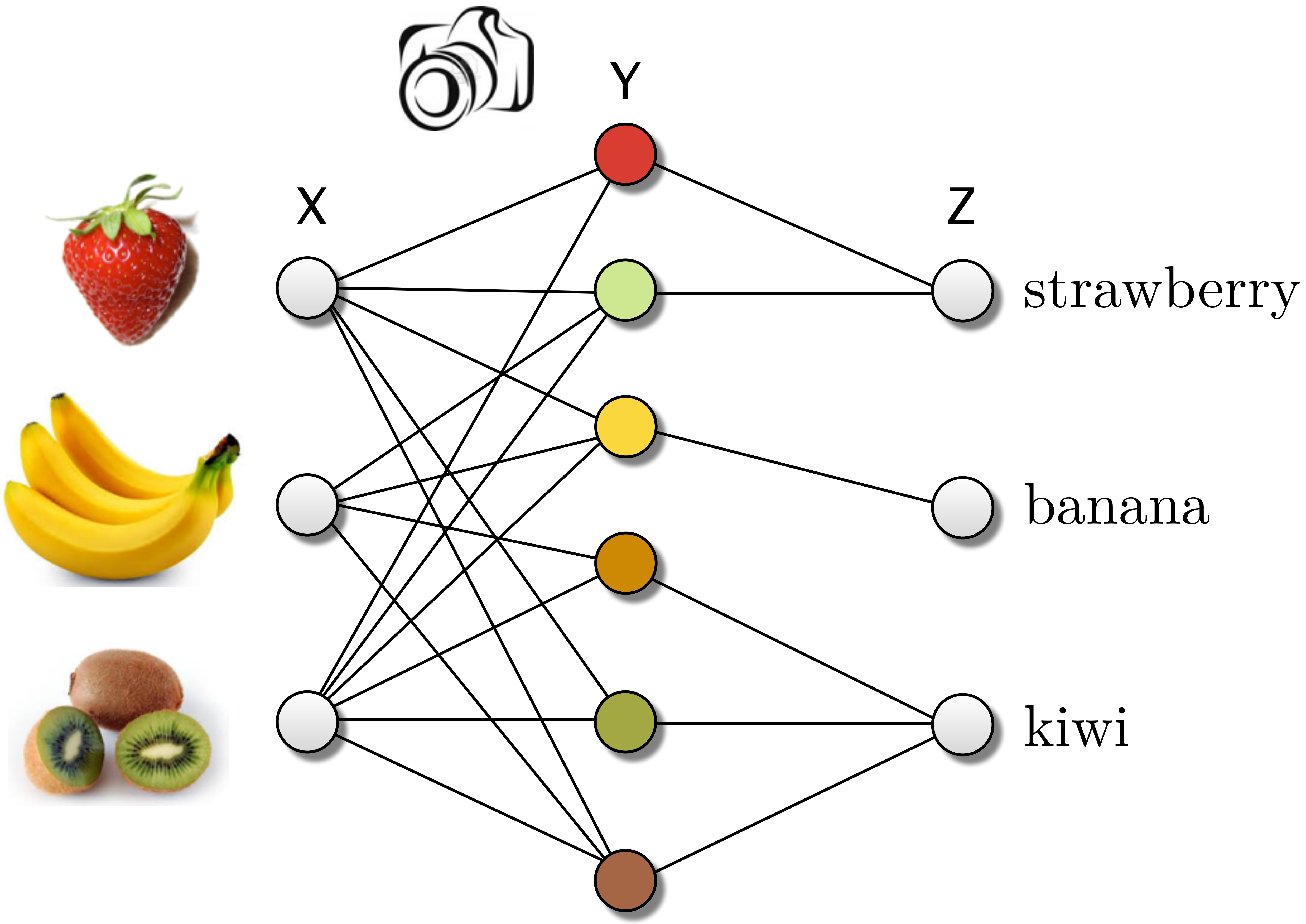
Machine Learning: Classification

$$\mathsf{X} \xrightarrow{\Pr(\mathsf{Y}|\mathsf{X})} \mathsf{Y} \xrightarrow{Q} \mathsf{Z}$$



Machine Learning: Classification

$$\mathsf{X} \xrightarrow{\Pr(\mathsf{Y}|\mathsf{X})} \mathsf{Y} \xrightarrow{Q} \mathsf{Z}$$



Minimum-entropy optimal classifier Q^*:

$$Q^* = \arg\min_{Q} H(\mathsf{Z}|\mathsf{X})$$

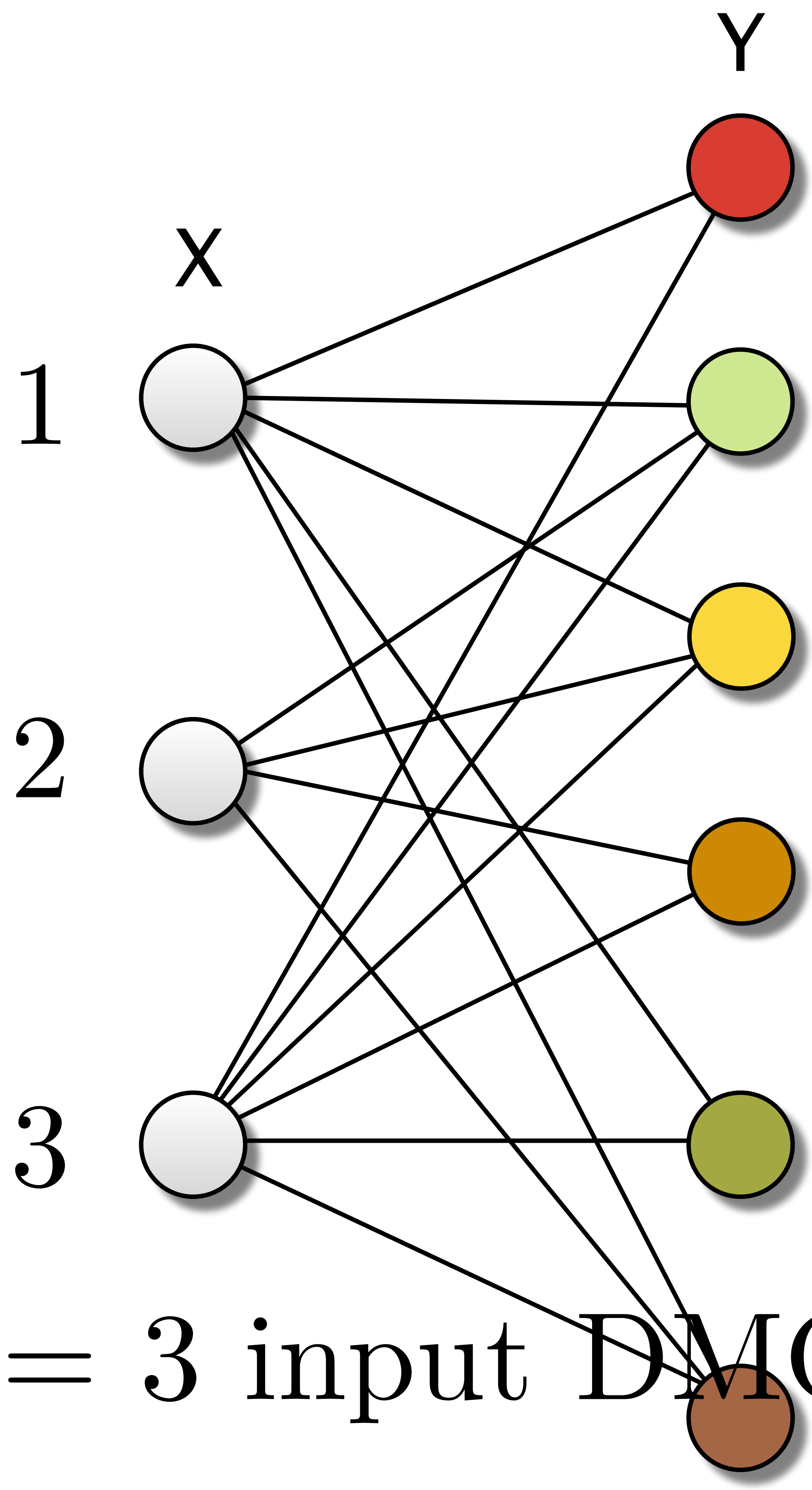
Exploit existing results in machine learning:

$$\underbrace{\max_{\text{LUT}} I(\mathsf{X}_3; \mathsf{Z})}_{\text{our problem}} = H(\mathsf{X}_3) - \underbrace{\min_{Q} H(\mathsf{Z}|\mathsf{X}_3)}_{\text{mach. learn.}}$$

Reverse Channel as a Vector

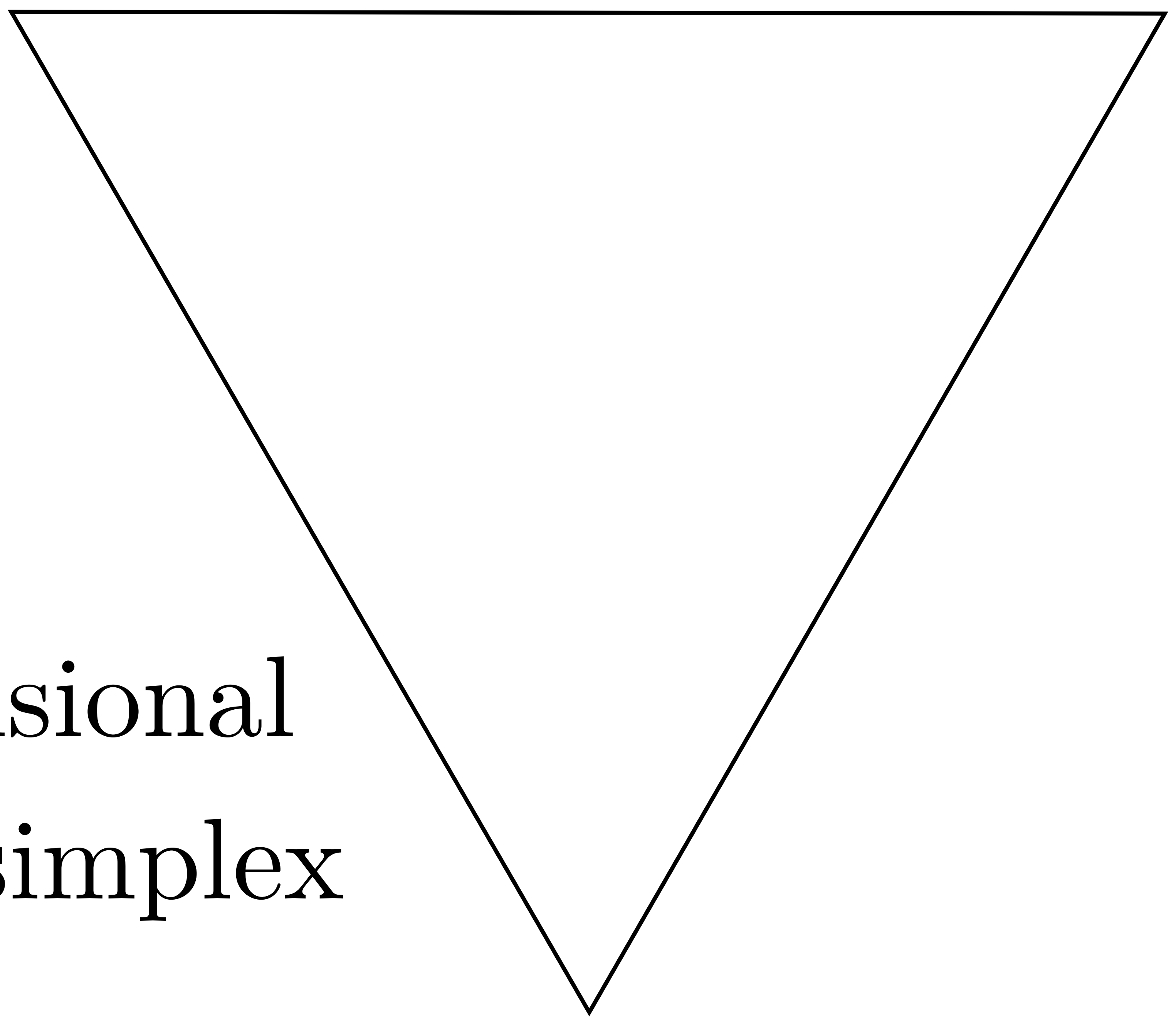
Consider the "reverse channel" $\Pr(\mathsf{X} \mid \mathsf{Y})$:

$$\mathbf{u}_y = \Big[\Pr(\mathsf{X}=1|\mathsf{Y}=y), \Pr(\mathsf{X}=2|\mathsf{Y}=y), \ldots, \Pr(\mathsf{X}=J|\mathsf{Y}=y)\Big]$$



$J = 3$ input DMC

$\mathbf{u}_y = [1\ 0\ 0]$ $\mathbf{u}_y = [0\ 1\ 0]$



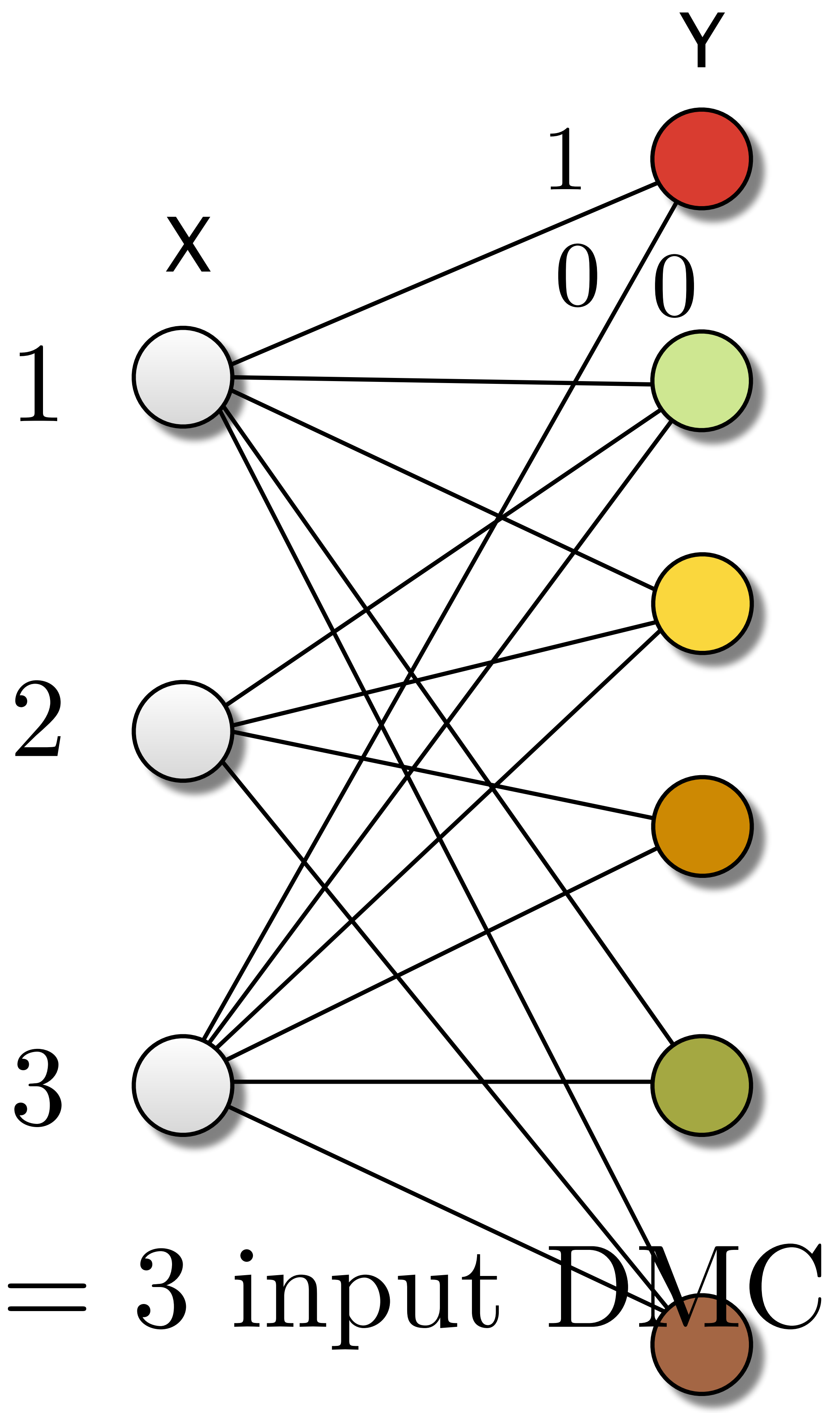
$J = 3$ dimensional probability simplex

$\mathbf{u}_y = [0\ 0\ 1]$

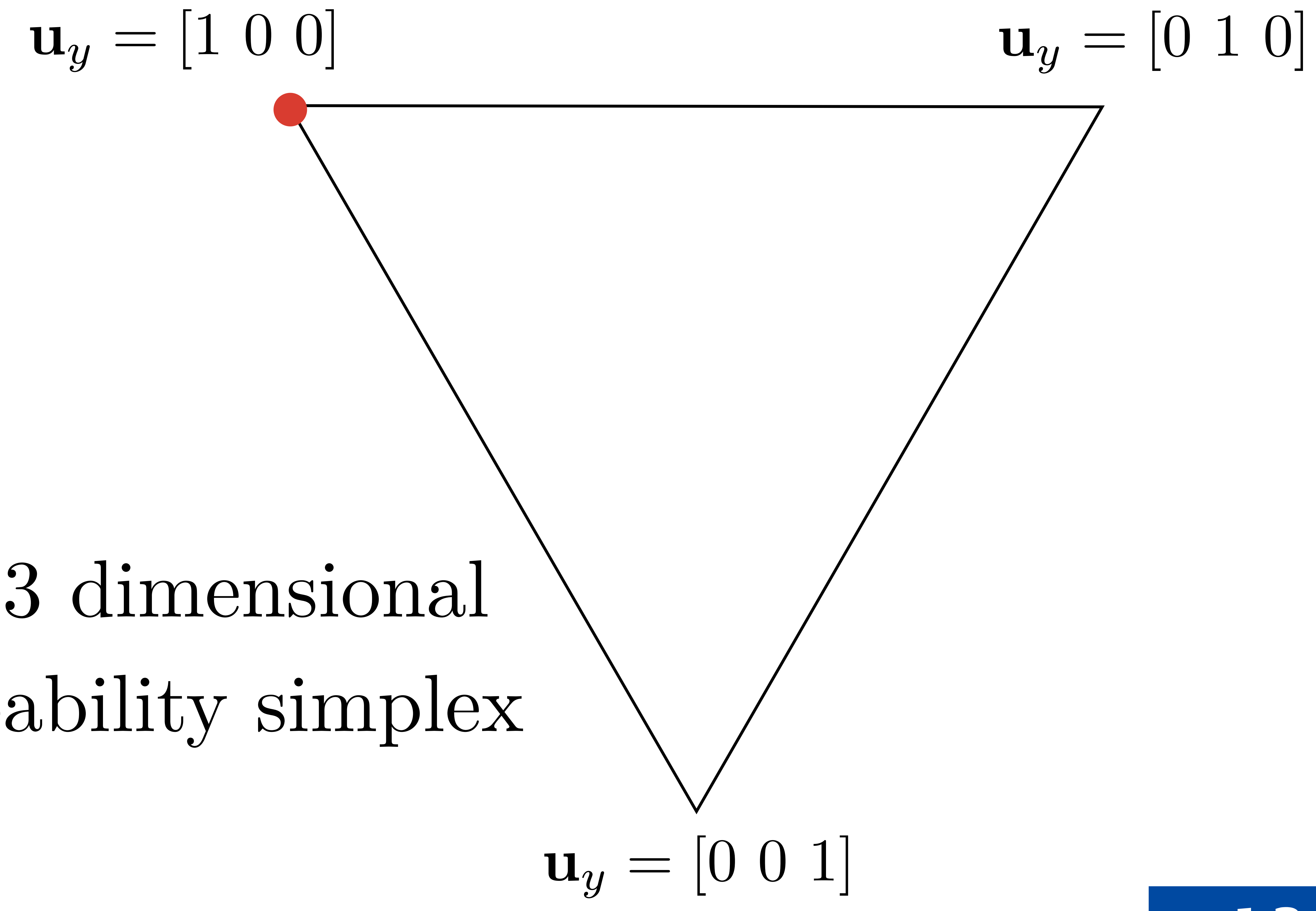
Reverse Channel as a Vector

Consider the "reverse channel" Pr(X | Y):

$$\mathbf{u}_y = \left[\Pr(\mathsf{X}=1|\mathsf{Y}=y), \Pr(\mathsf{X}=2|\mathsf{Y}=y), \ldots, \Pr(\mathsf{X}=J|\mathsf{Y}=y)\right]$$



$J = 3$ input DMC

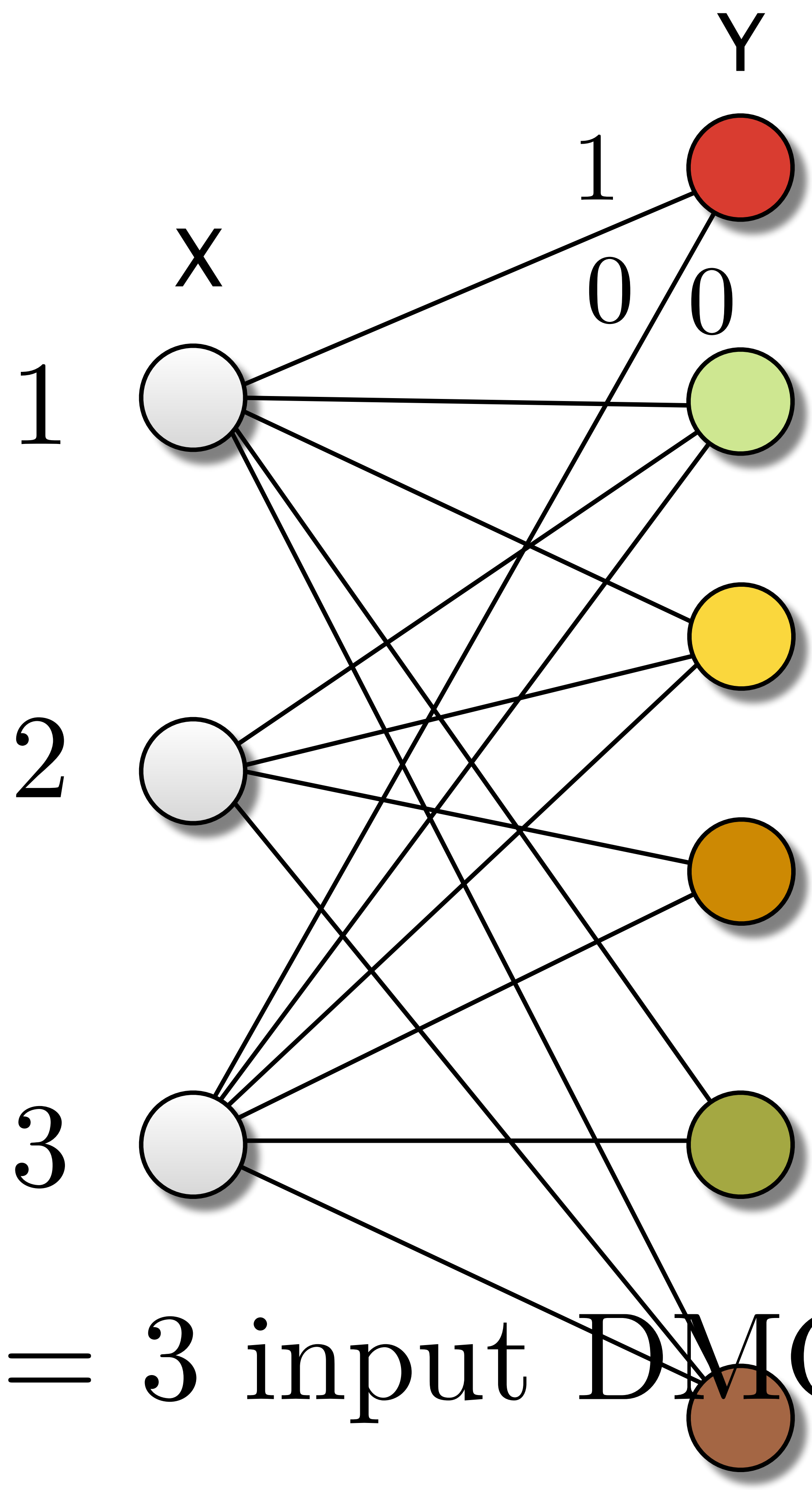


$J = 3$ dimensional
probability simplex

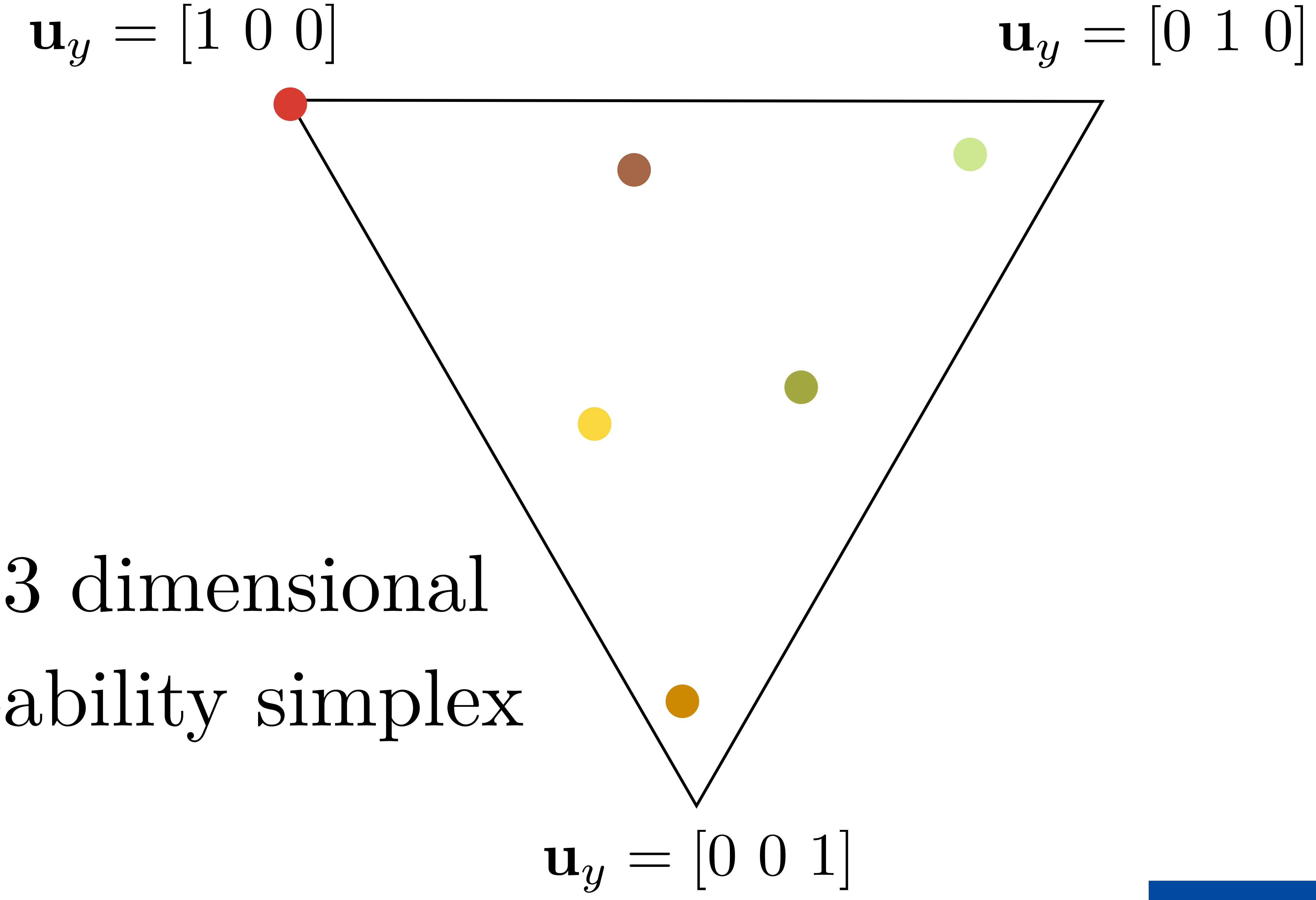
Reverse Channel as a Vector

Consider the "reverse channel" $\Pr(\mathsf{X} \mid \mathsf{Y})$:

$$\mathbf{u}_y = \Big[\Pr(\mathsf{X}=1|\mathsf{Y}=y), \Pr(\mathsf{X}=2|\mathsf{Y}=y), \ldots, \Pr(\mathsf{X}=J|\mathsf{Y}=y)\Big]$$

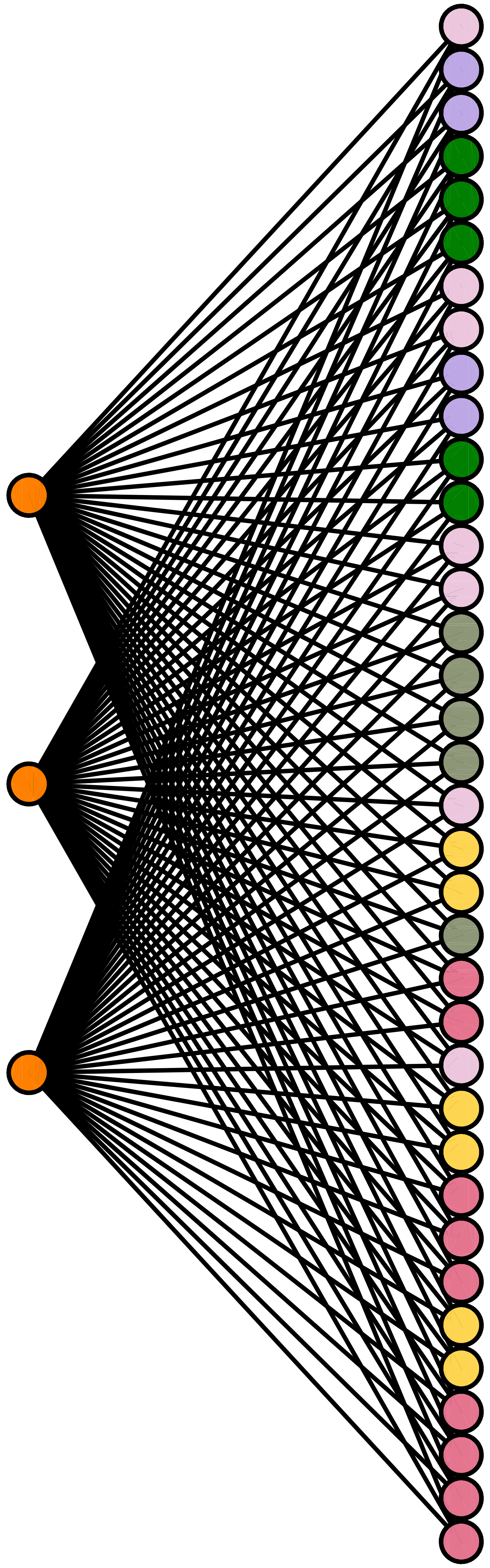


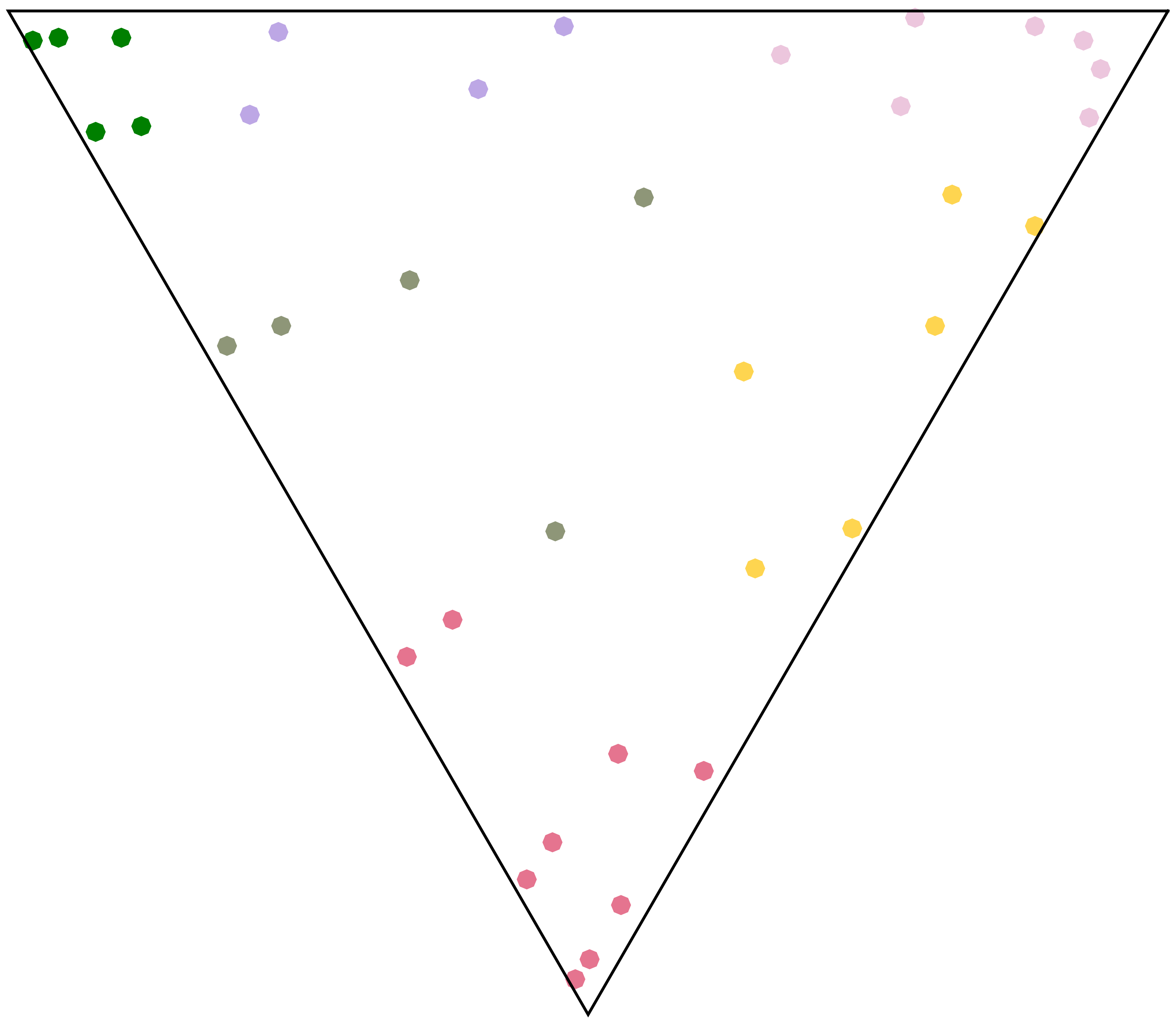
$J = 3$ input DMC



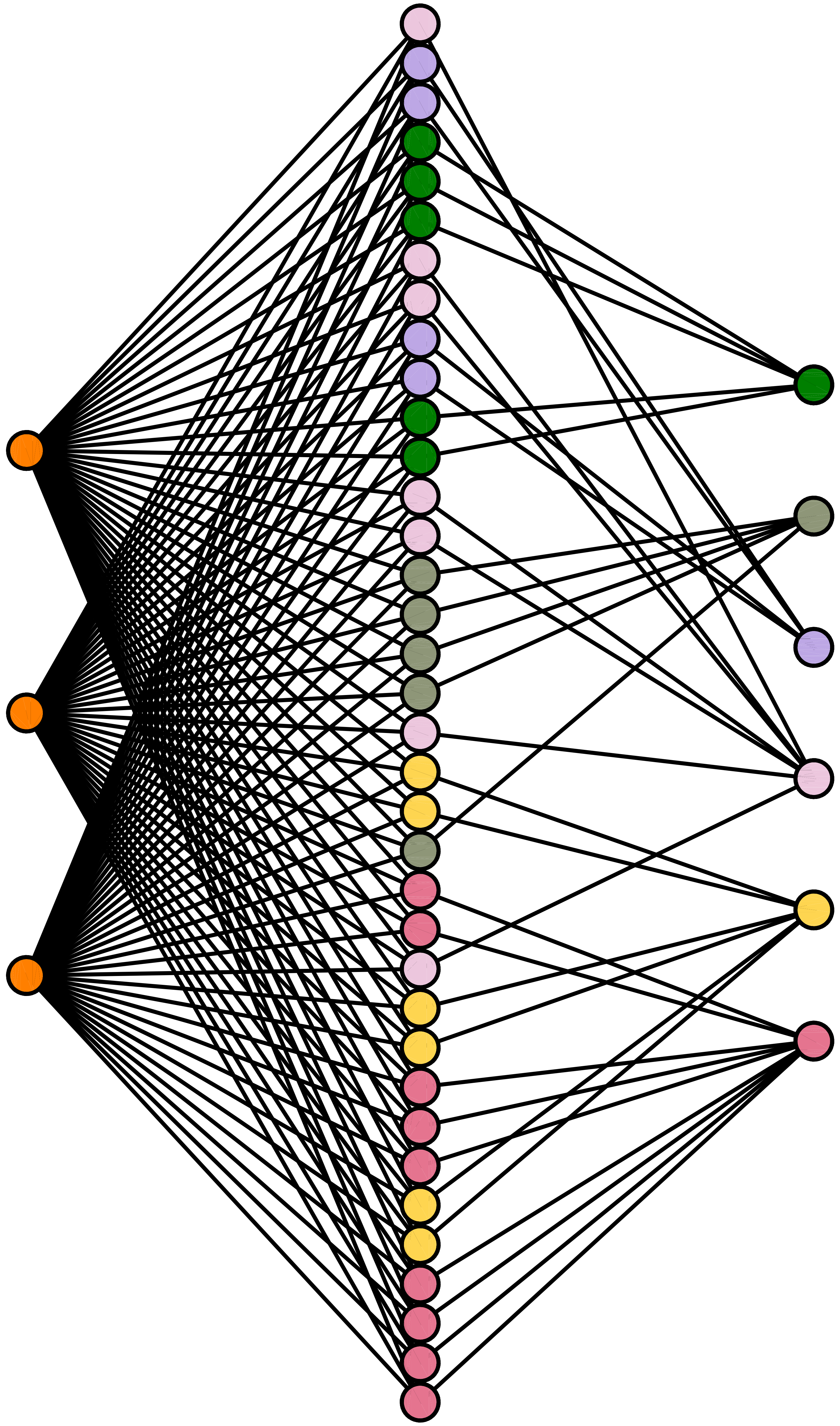
$J = 3$ dimensional probability simplex

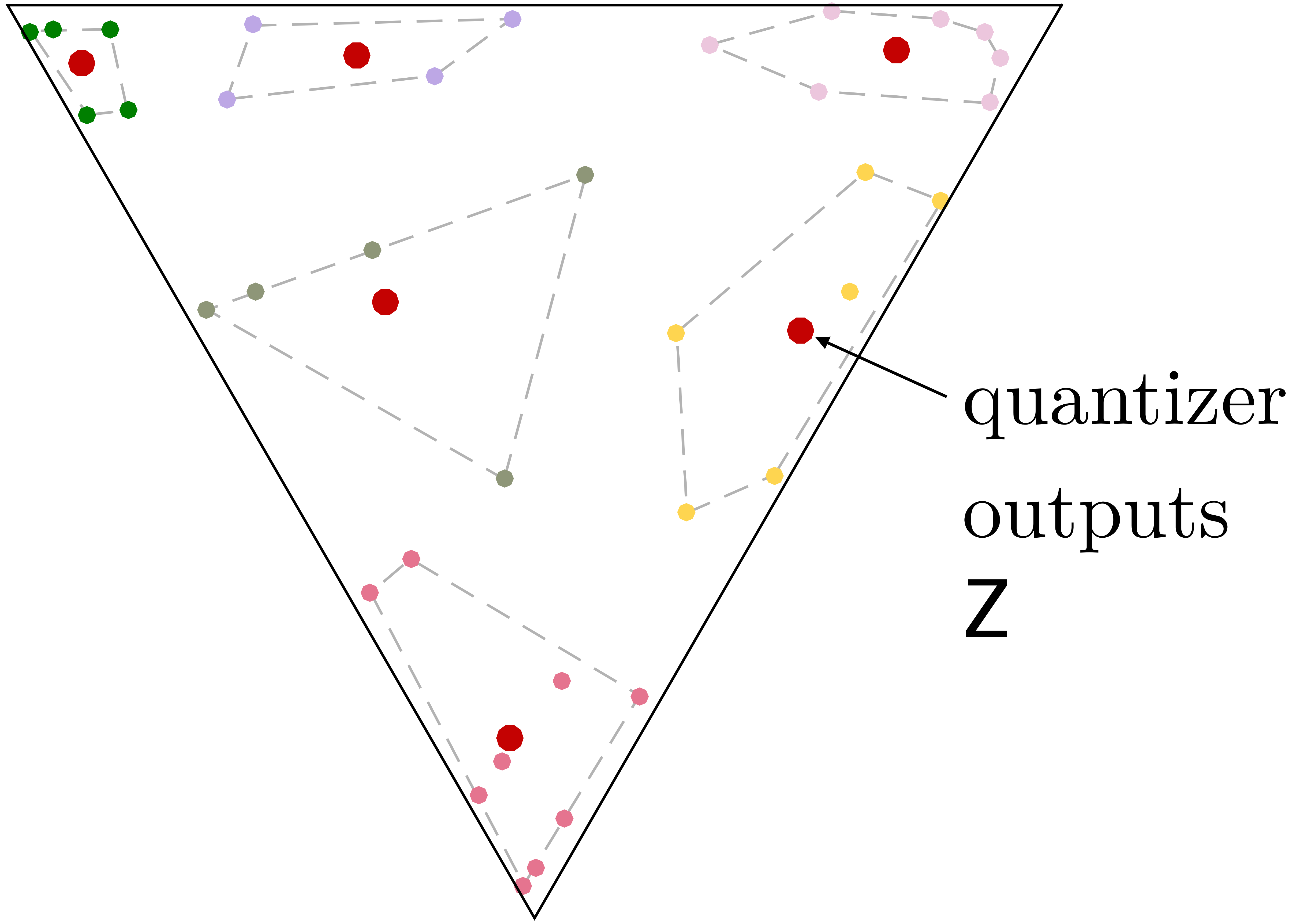
Quantization in Backwards Channel





Quantization in Backwards Channel





Maximization of Mutual Information is Minimization of KL Divergence

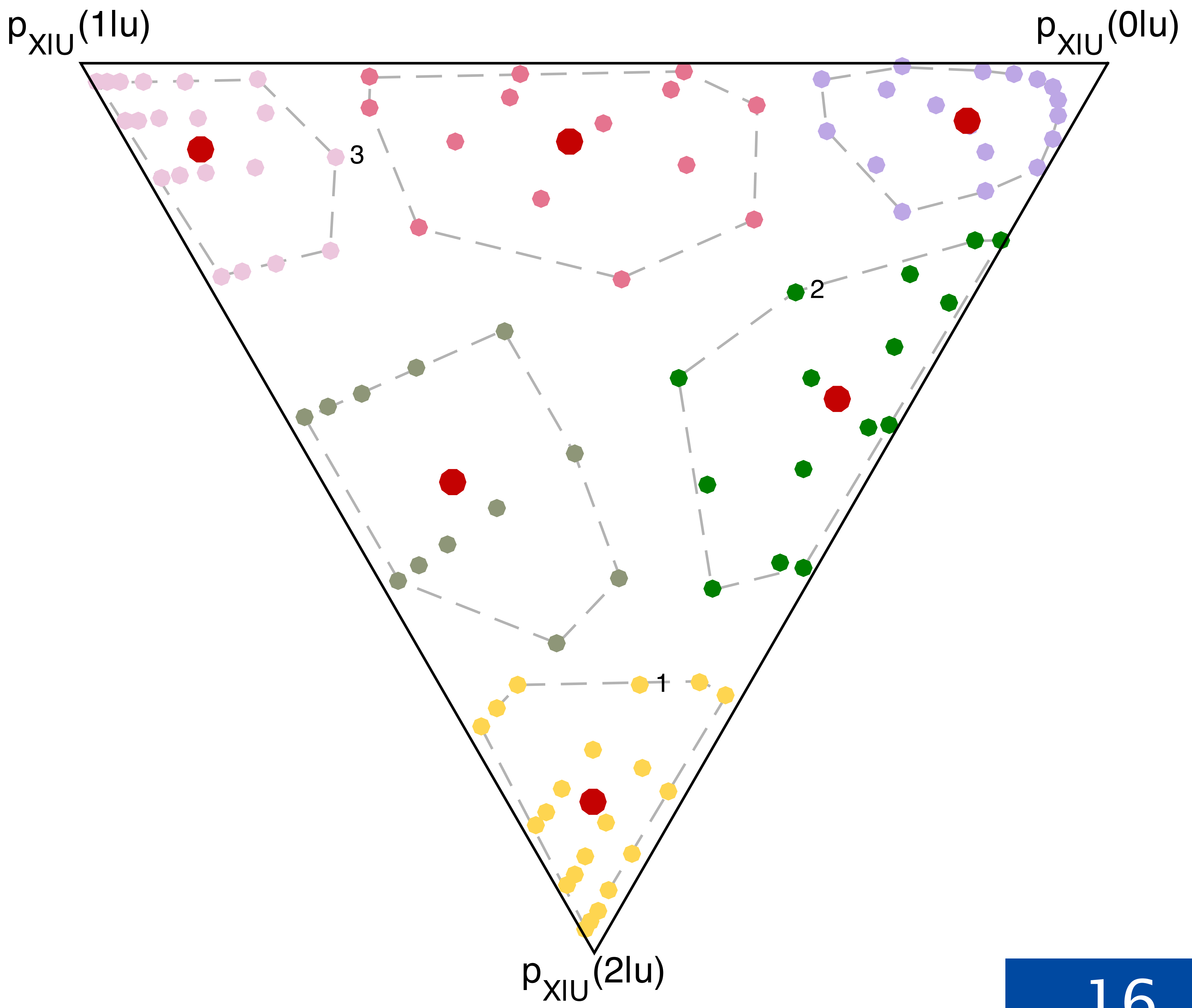
Consider random vector versions:

$$\mathbf{U} = \left[\, \Pr(\mathsf{X} = 1|\mathsf{Y}), \ldots, \Pr(\mathsf{X} = J|\mathsf{Y}) \right]$$
$$\mathbf{V} = \left[\, \Pr(\mathsf{X} = 1|\mathsf{Z}), \ldots, \Pr(\mathsf{X} = J|\mathsf{Z}) \right]$$

Then, it can be shown:

$$I(\mathsf{X};\mathsf{Y}) - I(\mathsf{X};\mathsf{Z}) = E\big(D(\mathbf{U}||\mathbf{V})\big)$$

$D(\cdot||\cdot)$ is the Kullback-Leiber divergence and E is expectation.



K-Means Algorithm

K-means clustering: vector quantization method, or clustering method.

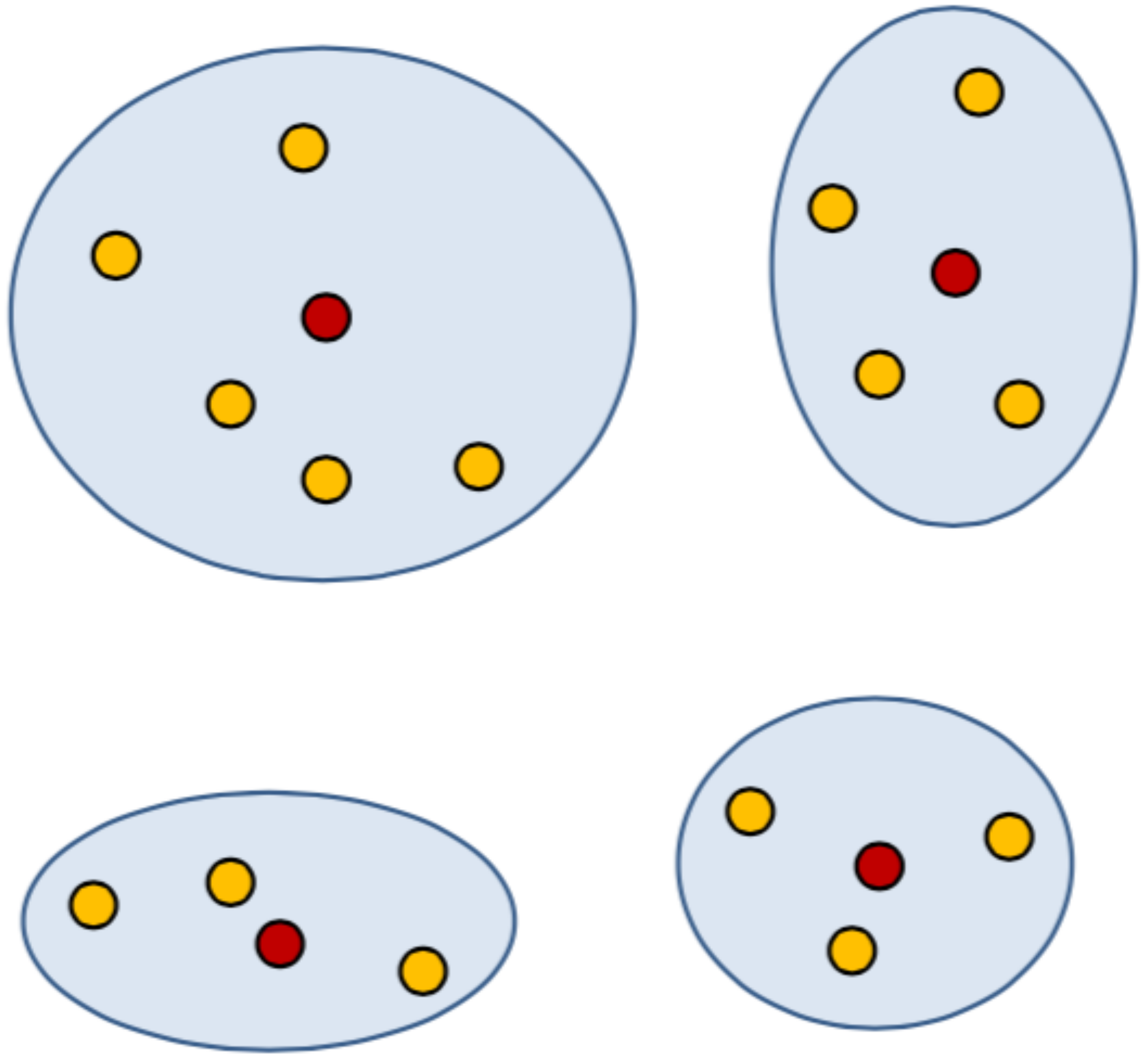
iterate

1. given n-dimensional data set, randomly choose K "means"
2. each of K clusters consists of data points closest to its mean in Euclidean distance
3. move the mean to the center of the cluster

Not optimal, but works well in practice.

Complexity is linear in M

Hugely successful in machine learning

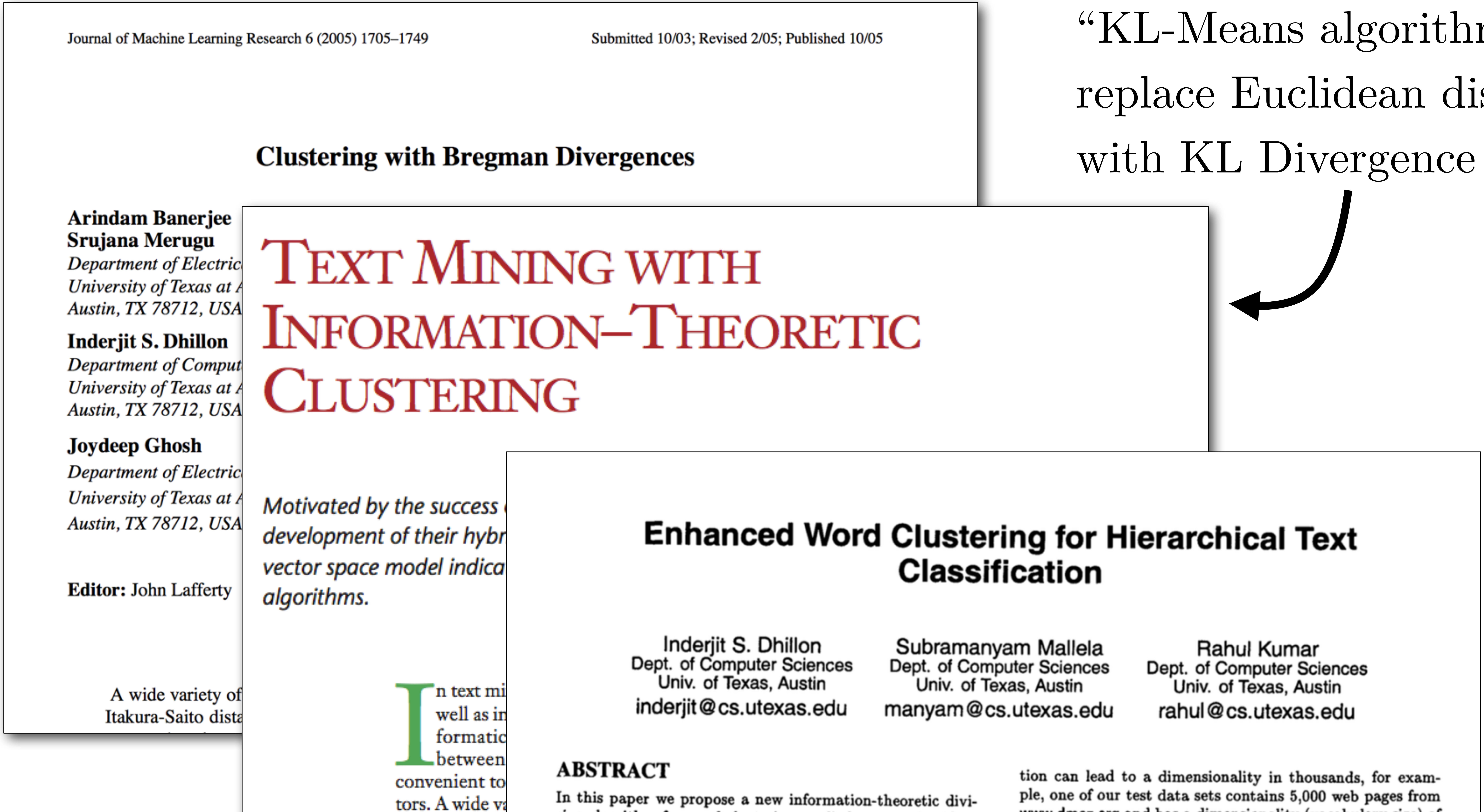


data $\boldsymbol{u}_m = (u_1 u_2 \dots u_J)_m$

means $\boldsymbol{v}_k = (v_1 v_2 \dots v_J)_k$

KL-Means Algorithm: KL Divergence

"KL-Means algorithm"
replace Euclidean distance
with KL Divergence



Journal of Machine Learning Research 6 (2005) 1705–1749 Submitted 10/03; Revised 2/05; Published 10/05

Clustering with Bregman Divergences

Arindam Banerjee
Srujana Merugu
Department of Electric
University of Texas at A
Austin, TX 78712, USA
Inderjit S. Dhillon
Department of Comput
University of Texas at A
Austin, TX 78712, USA
Joydeep Ghosh
Department of Electric
University of Texas at A
Austin, TX 78712, USA

Editor: John Lafferty

A wide variety of
Itakura-Saito dista

TEXT MINING WITH INFORMATION–THEORETIC CLUSTERING

Motivated by the success
development of their hybr
vector space model indica
algorithms.

In text mi
well as in
formatic
between
convenient to
tors. A wide va

Enhanced Word Clustering for Hierarchical Text Classification

Inderjit S. Dhillon
Dept. of Computer Sciences
Univ. of Texas, Austin
inderjit@cs.utexas.edu

Subramanyam Mallela
Dept. of Computer Sciences
Univ. of Texas, Austin
manyam@cs.utexas.edu

Rahul Kumar
Dept. of Computer Sciences
Univ. of Texas, Austin
rahul@cs.utexas.edu

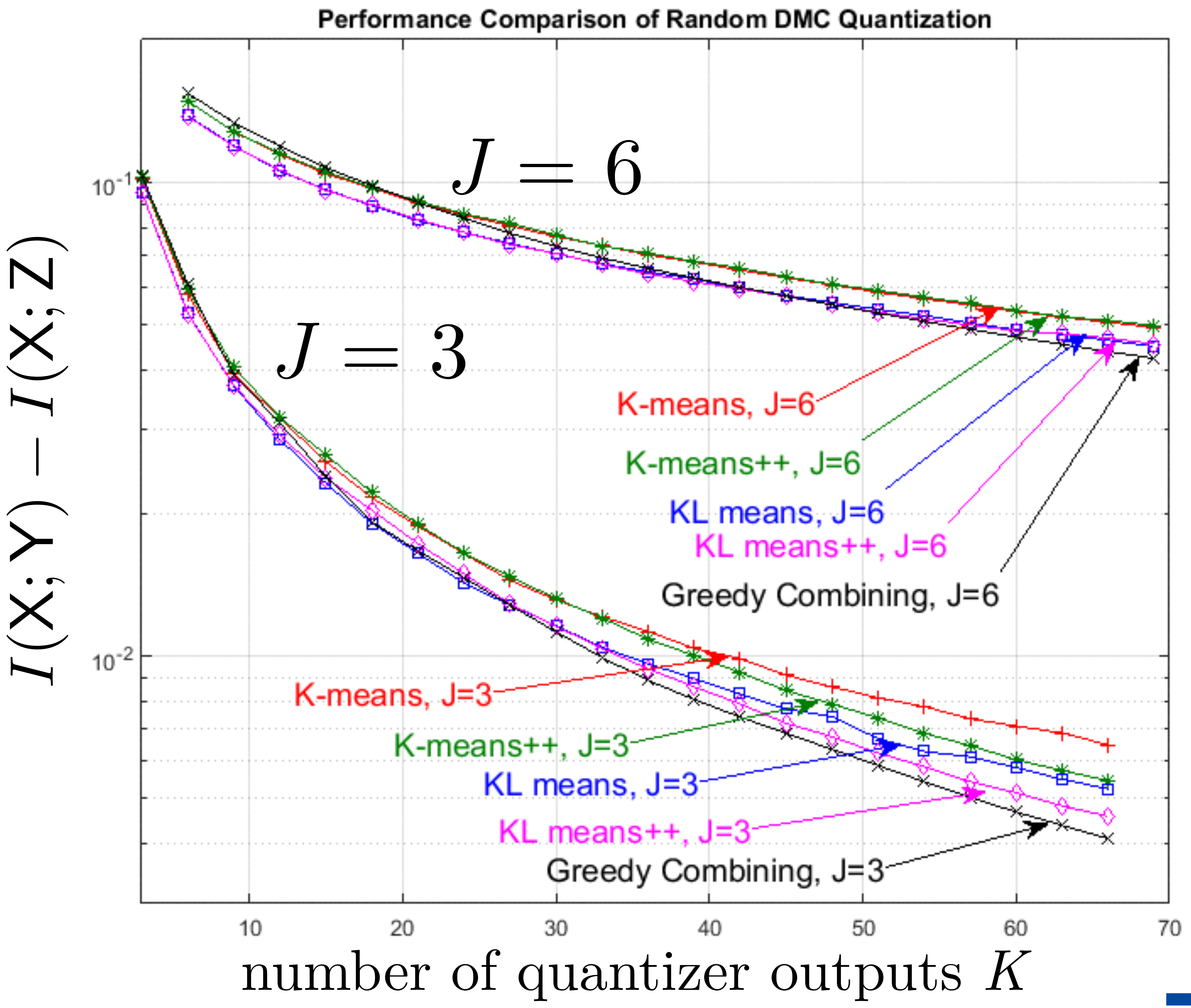
ABSTRACT

In this paper we propose a new information-theoretic divi-

tion can lead to a dimensionality in thousands, for exam-
ple, one of our test data sets contains 5,000 web pages from

Quantization of Non-Binary Channels

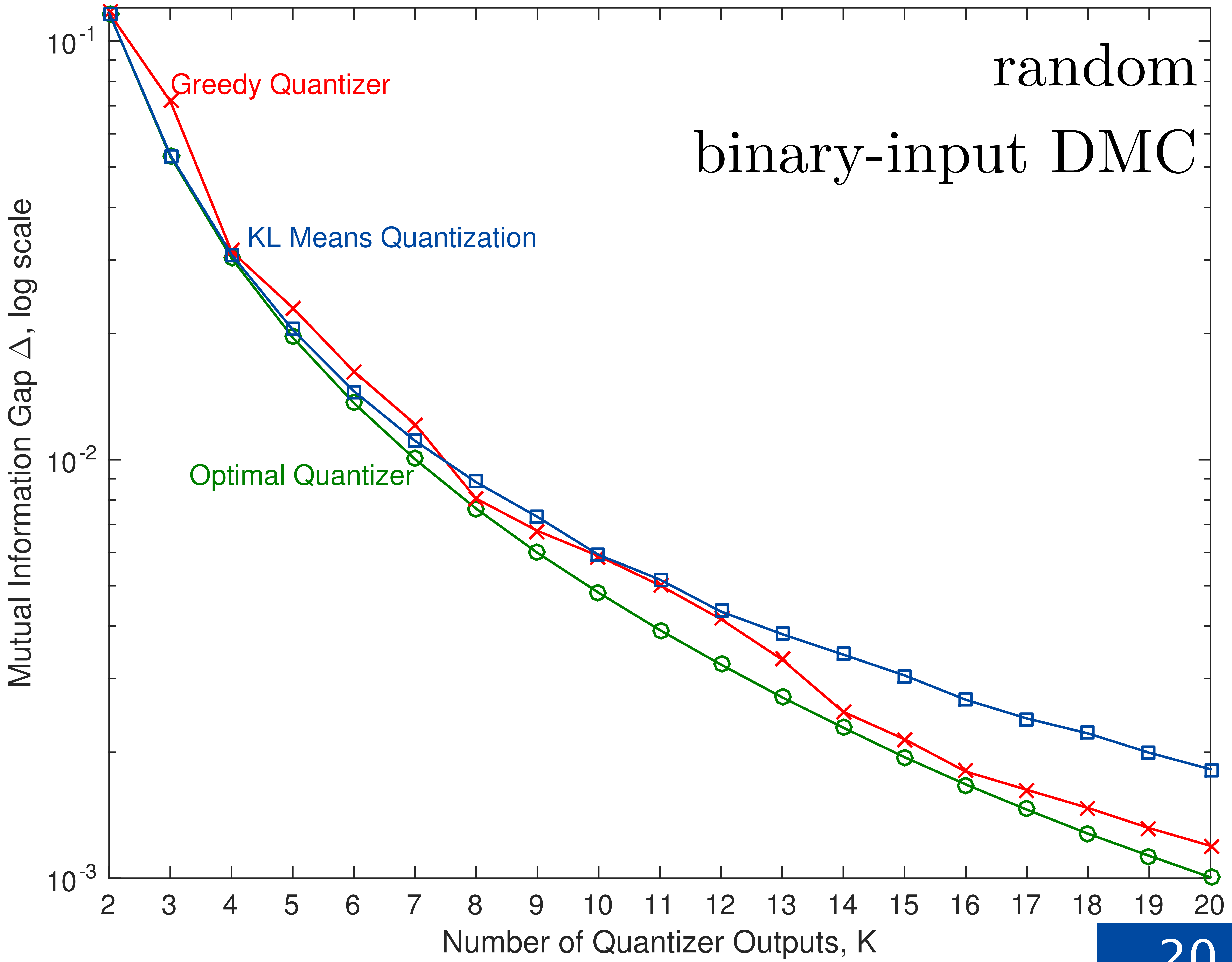
- Random DMC with 512 outputs
- Efficient optimal quantization is not known
- Greedy combining performs well, but has complexity M^3
- KL-Means has complexity M



"KL-Means" MI-max Quantizers

Replace the Euclidean distance metric with Kullback-Lieber divergence

Objective is now to maximize mutual information



Conclusion

For the Markov chain $\mathsf{X} \to \mathsf{Y} \to \mathsf{Z}$, finding the optimal quantizer $Q^* = \arg\max I(\mathsf{X};\mathsf{Z})$ has exponential complexity.

Motivated by Max-LUT method for factor graph hardware implementations

Described **KL Means Algorithm**, K-Means clustering with KL distance

- Efficient but suboptimal
- KL Means is equivalent to informant bottleneck method [SCC17]
- moving closer to optimal: e.g. KL-Means++ and IB-based soft clustering
- need optimal quantization as a benchmark
- application to Max-LUT method for non-binary factor graphs