Information-Theoretic Quantization and Its Connection to Classification

Brian M. Kurkoski
Japan Advanced Institute of Science and Technology

12 February 2019
12th International ITG Conference on Systems, Communication and Coding — SCC2019
Rostock, Germany

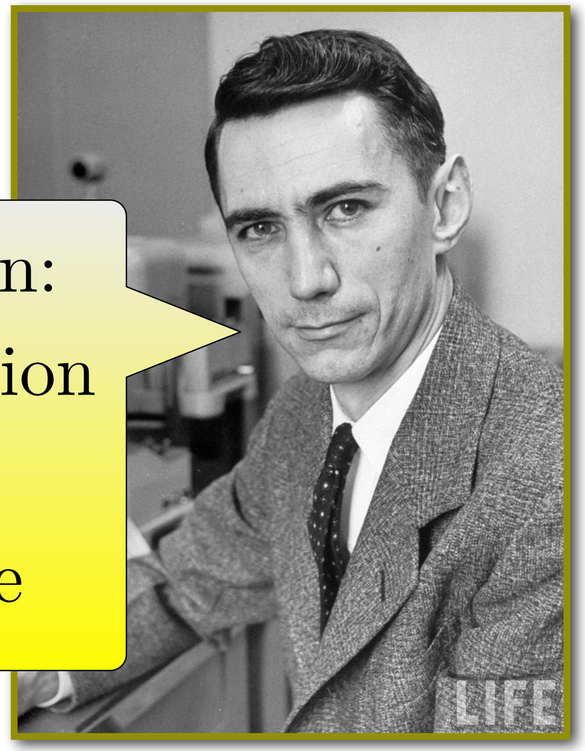
Just Enough Information Theory

What is the best code we can design?

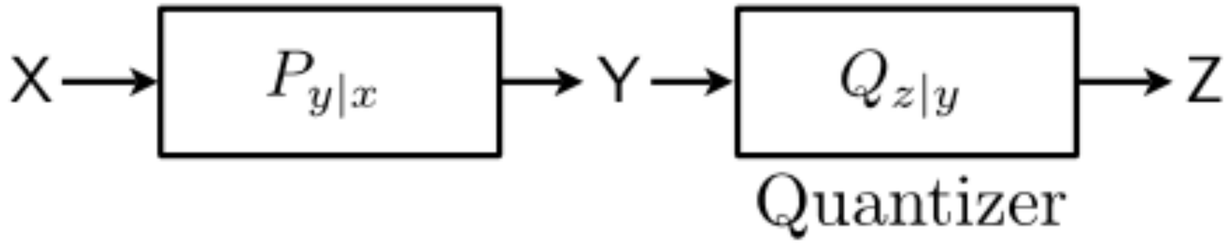
$$R < C = \max_{p_{\mathsf{X}}(x)} I(\mathsf{X};\mathsf{Y})$$

Code rate $<$ Channel Capacity

Claude Shannon: mutual information is the highest achievable rate



Highest Achievable Rate for a Communications over a Quantized Channel



Given a channel, find the *quantizer* Q which maximizes the achievable rate:

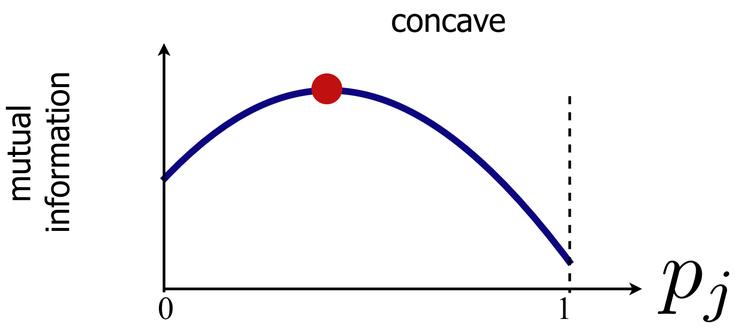
$$C = \max_{Q} I(\mathsf{X}; \mathsf{Z})$$

We will fix the input distribution $p_{\mathsf{X}}(x)$.

Jointly optimizing Q and $p_{\mathsf{X}}(x)$ is a much more difficult problem.

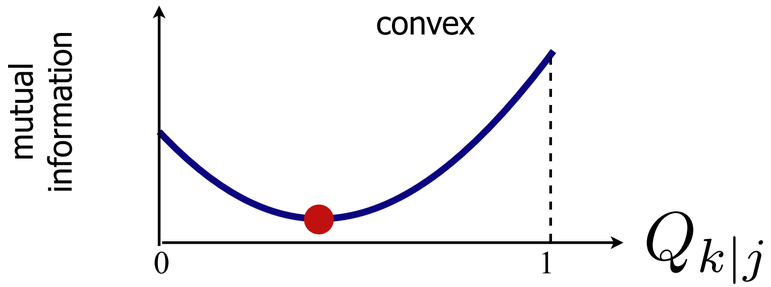
Comparison of Optimization Problems

Capacity computation



$$\max_{p_j} I(X;Z)$$

Rate-distortion computation

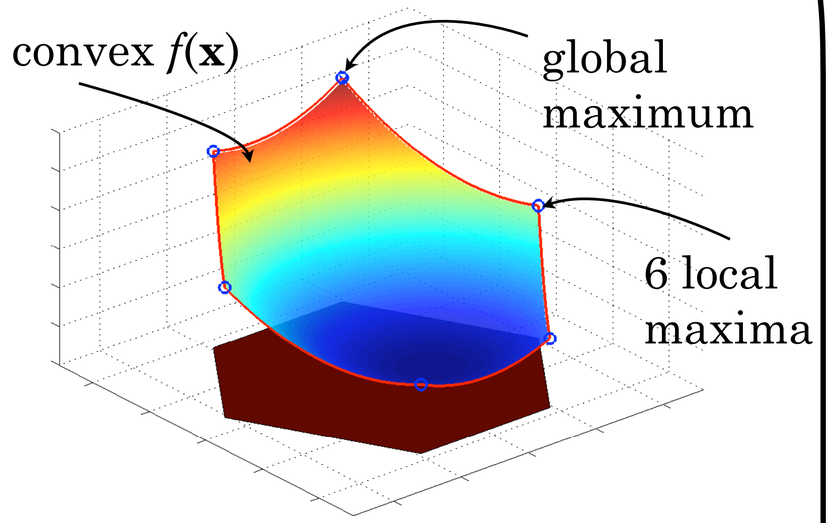


$$\min_{Q_{k|i}} I(X;Z)$$

Concave optimization—easy

Arimoto-Blahut, etc.

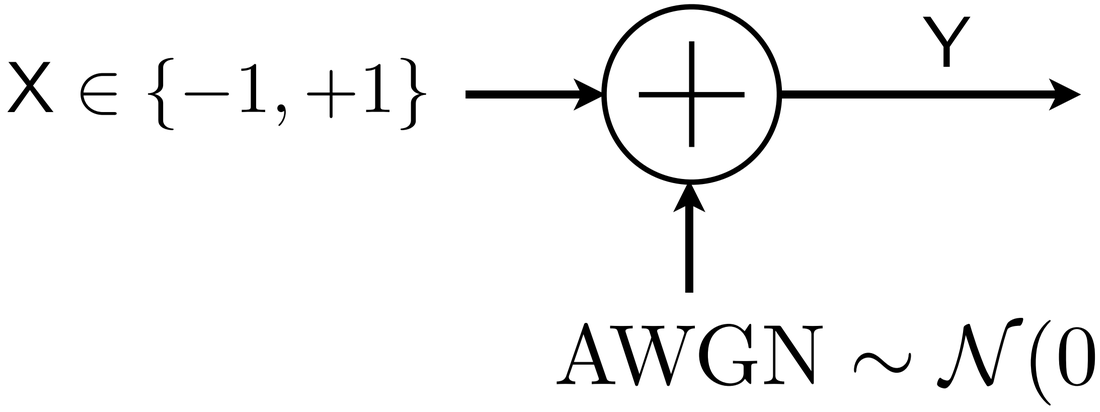
Channel Quantization

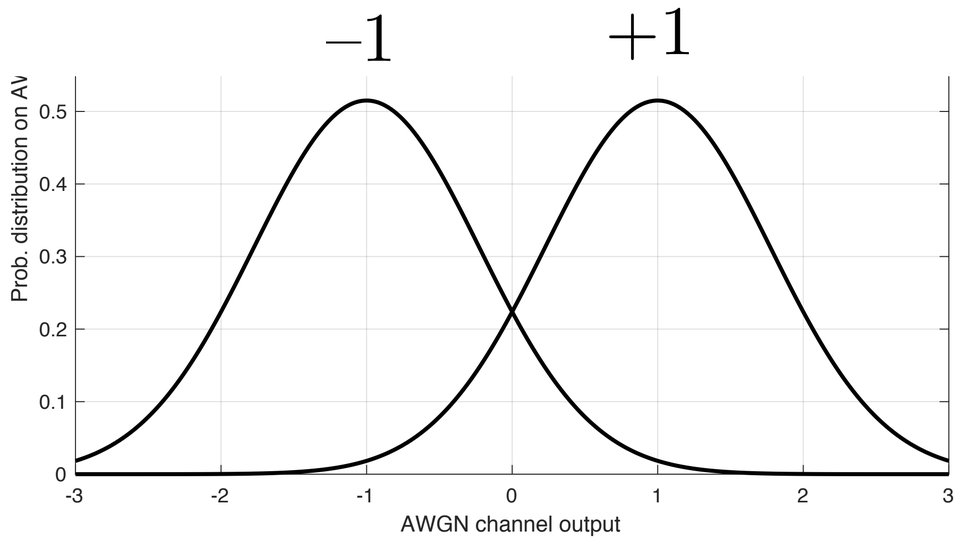


$$\max_{Q} I(\mathsf{X};\mathsf{Z})$$

Global optimization, hard!

Motivation 1: "Digitization"

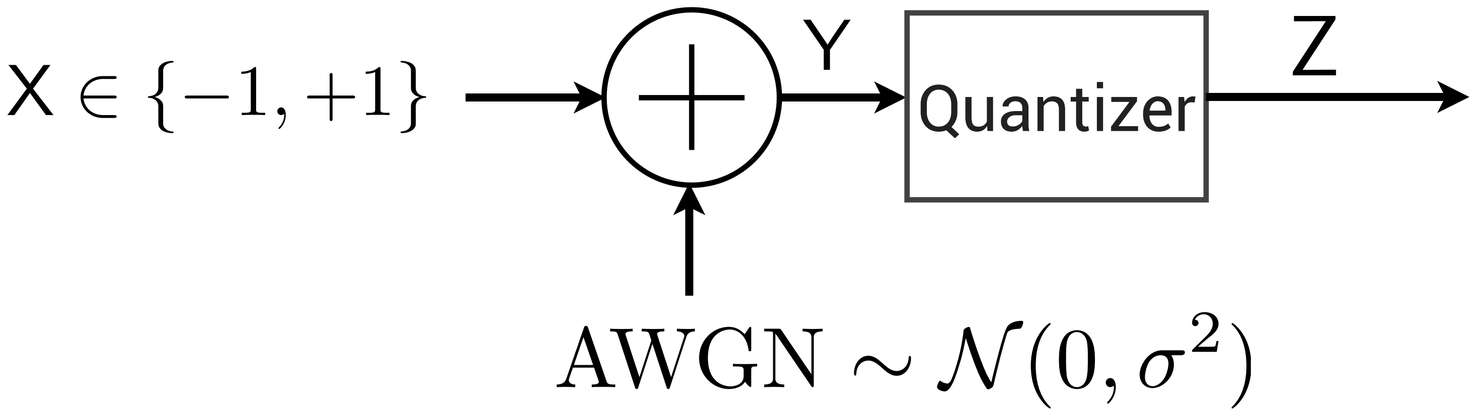


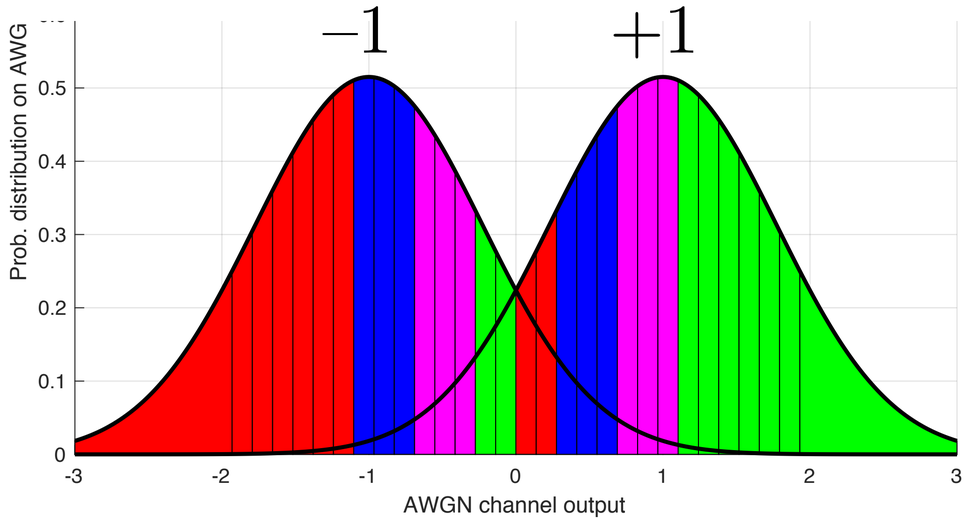


Given a continuous-output channel, we want to create a discrete version

- For example, digital circuits deal with discrete values.

Motivation 1: "Digitization"





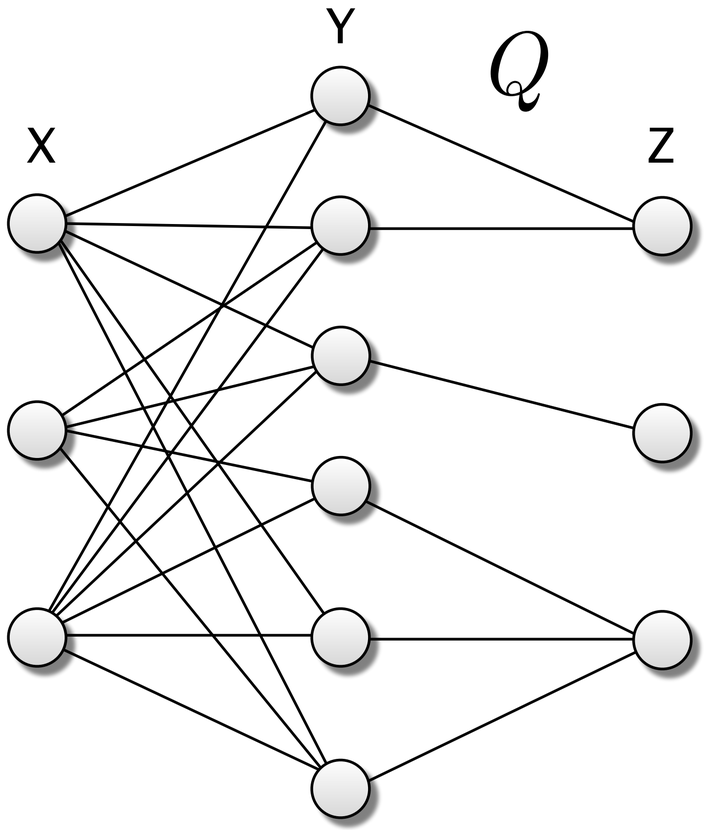
Given a continuous-output channel, we want to create a discrete version

- For example, digital circuits deal with discrete values.
- A quantizer Q maps real values Y to discrete values $\mathsf{Z} \in \{1, \ldots, M\}$
- How to choose the "quantization boundaries" to max $I(\mathsf{X}; \mathsf{Z})$?

X. Ma, X. Zhang, H. Yu, and A. Kavcic, "Optimal quantization for soft-decision decoding revisited," in Proc. Int. Symp. Inform. Theory Appl., Xian, China, Oct. 2002

“Quantization” of Discrete Memoryless Channel

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



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

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

with $|\mathsf{Z}| < |\mathsf{Y}|$.

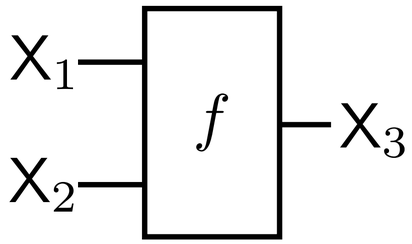
$|\mathsf{Z}| \geq |\mathsf{Y}|$ is trivial.

“Channel downgrading” in polar code design.

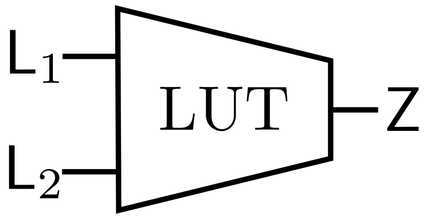
Motivation 2: Lookup Tables for Decoding

Factor Graph

Encoder-side function f

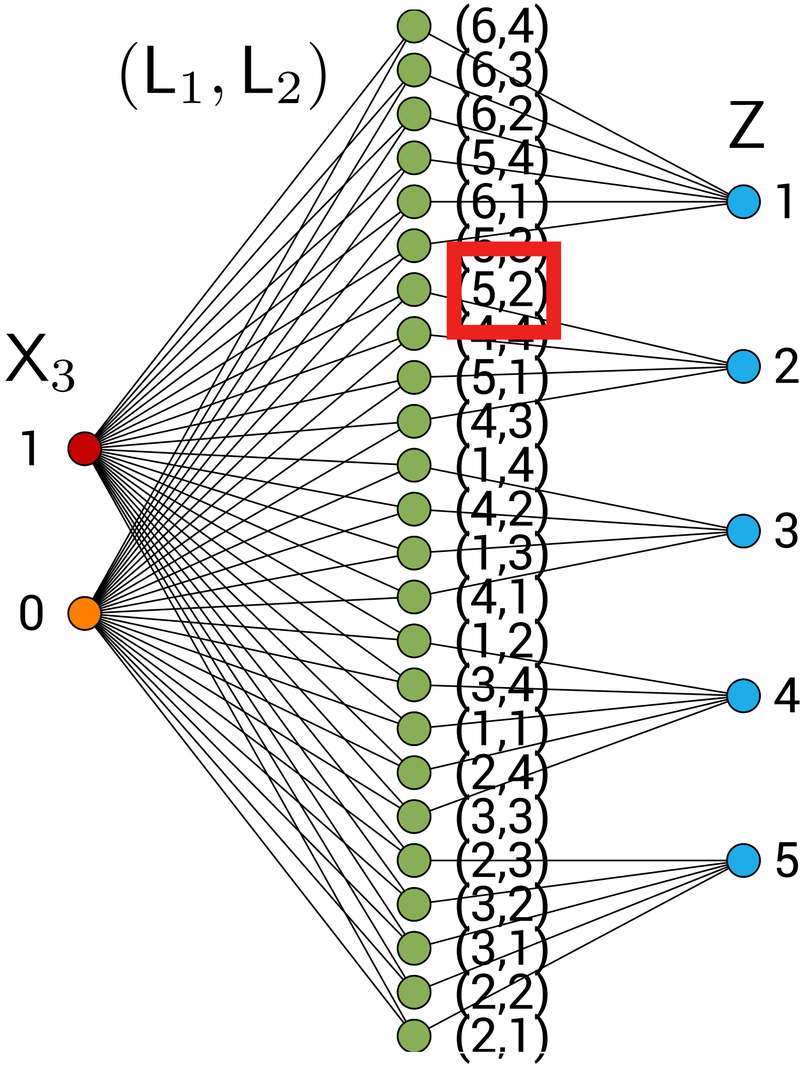


Decoder-side lookup-table



L_i is a noisy version of X_i

Z is a noisy version of X_3



Lookup table:

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

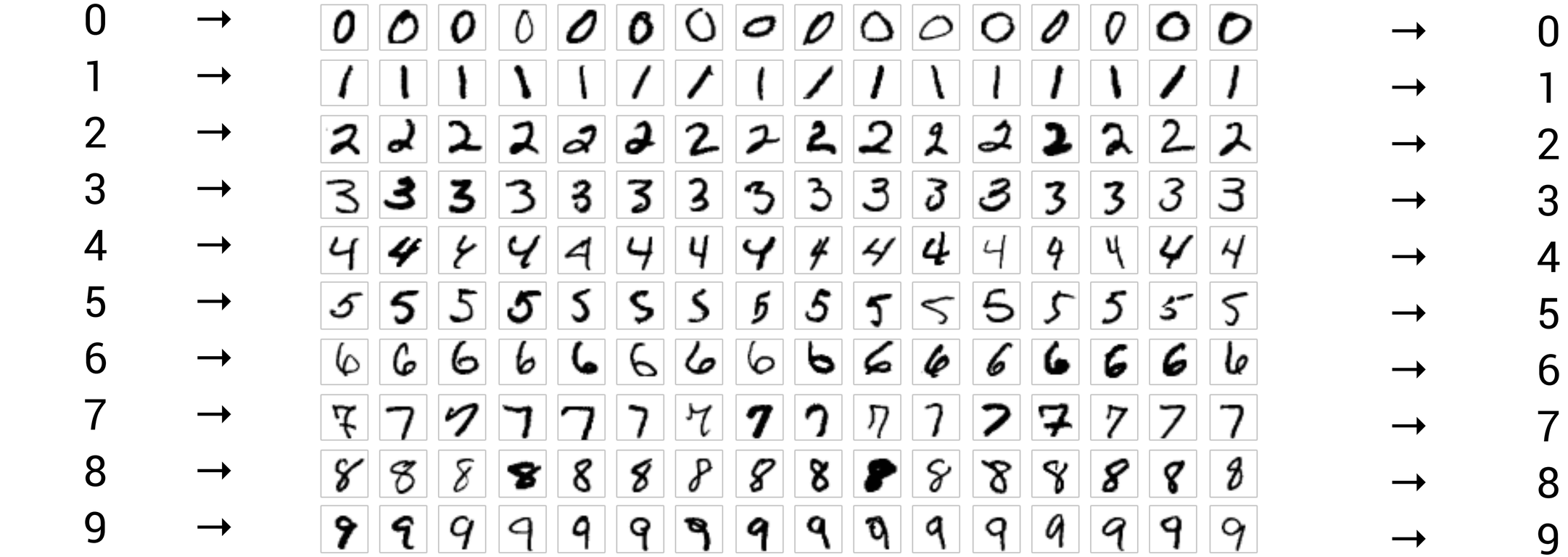
L_1 \ L_2	1	2	3	4
1	4	4	3	3
2	5	5	5	4
3	5	5	4	4
4	3	3	2	2
5	2	2	1	1
6	1	1	1	1

$$\max_{\mathrm{LUT}} I\big(X_3; \mathrm{LUT}(L_1, L_2)\big)$$

J. Lewandowsky, M. Stark, and G. Bauch, “Information bottleneck graphs for receiver design,” in Proceedings of IEEE International Symposium on Information Theory. IEEE, July 2016, pp. 2888–2892.

Handwriting Recognition is Typical Example of Classification

input X | noisy observation Y | estimate Z

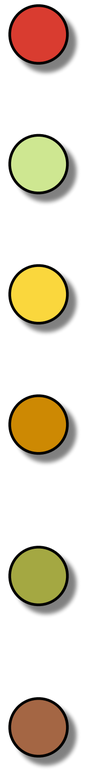


Deep neural networks solve this problem well, but are a bit complicated

Wikimedia/Josef Steppan using MNIST data set

Simple Example of Classification

one-pixel
camera



output: one color

Simple Example of Classification

Y

Classifier: $Q(y) = z$

X

Z

strawberry

durian

kiwi

Goal and Outline

Discuss tools for classification which can be applied to channel quantization

Can find effective or even optimal methods for channel quantization

Outline

- K-Means algorithm with Euclidean metrics
- K-Means with generalized metrics
  - “convexity theorem” by Burshtein et al
- Backwards channel perspective
- Three particular cases:
  - Quantizaiton of arbitrary DMC using K-Means
  - Optimal quantization of binary input DMC using dynamic programming
  - Quantization of binary-input, continuous-output channels

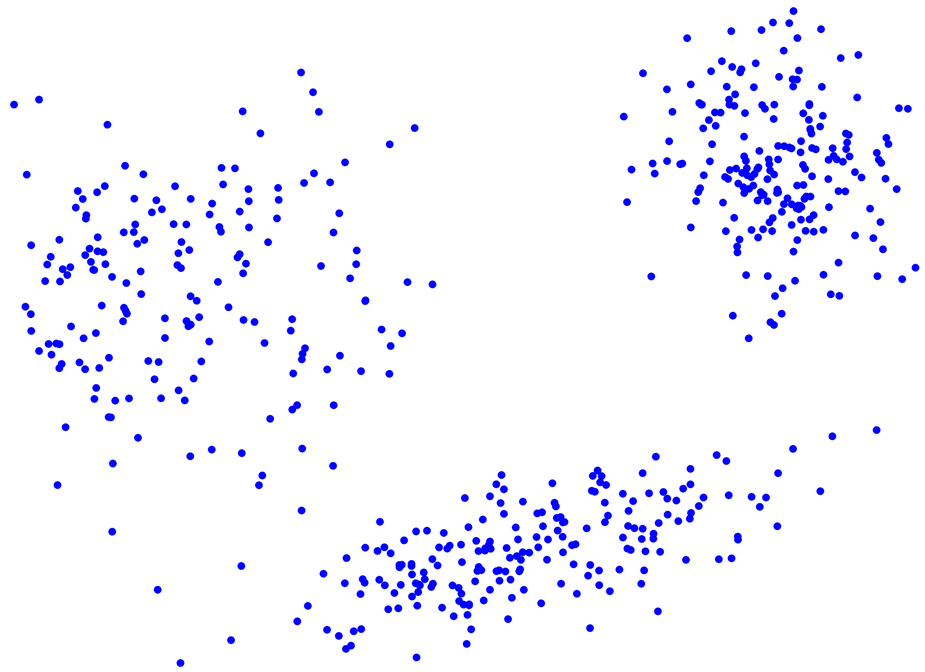
K-Means Algorithm

K-means algorithm partitions n observations into k clusters — each observation belongs to the cluster with the nearest mean. Can also be seen as vector quantization.

Attempts to minimize mean-squared of quantization

$$\min_{Q} E\left[(\mathsf{X} - Q(\mathsf{X}))^2\right]$$

Not optimal, but works well. Widely used in machine learning.



Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Data clustering: 50 years beyond K-means ☆

Anil K. Jain *

Department of Computer Science and Engineering, Michigan State University, East Lansing, Michigan 48824, USA
Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seoul, 136-713, Korea

ARTICLE INFO

ABSTRACT

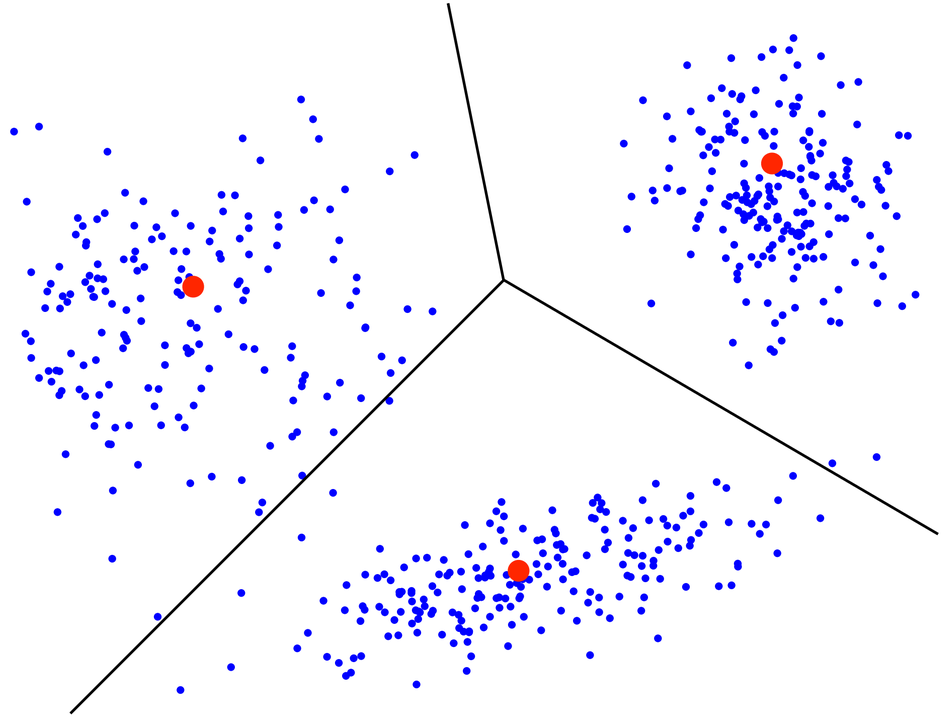
K-Means Algorithm

K-means algorithm partitions n observations into k clusters — each observation belongs to the cluster with the nearest mean. Can also be seen as vector quantization.

Attempts to minimize mean-squared of quantization

$$\min_{Q} E\left[(\mathsf{X} - Q(\mathsf{X}))^2\right]$$

Not optimal, but works well. Widely used in machine learning.



Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Data clustering: 50 years beyond K-means ☆

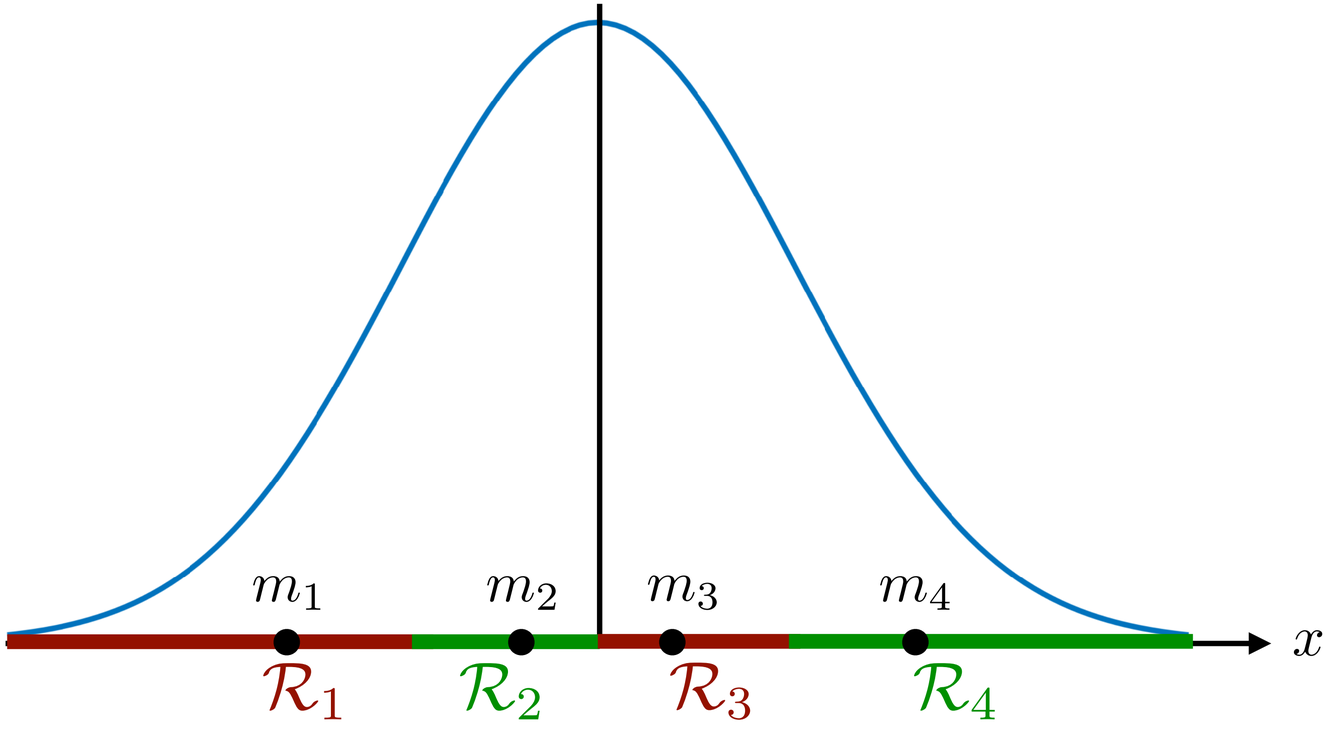
Anil K. Jain *

Department of Computer Science and Engineering, Michigan State University, East Lansing, Michigan 48824, USA
Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seoul, 136-713, Korea

ARTICLE INFO

ABSTRACT

Two Observations

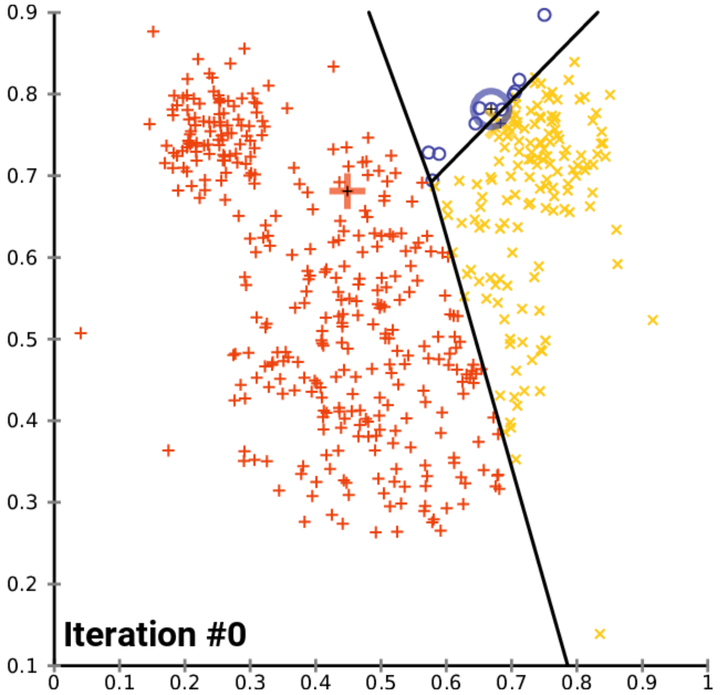


- Given reconstruction points $\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_K$, the reconstruction region should minimize the average error. Or, any point $\mathbf{y}$ should be **assigned** to the closest point $\mathbf{m}_i$.
- Within some region $\mathcal{R}_i$, its reconstruction point $\mathbf{m}_i$ should be **updated** to minimize the average error for that region.

K-Means Algorithm with Euclidean Distance Metric

iterate

1. given n-dimensional data set, randomly choose K means (centroids)
2. **Assignment step** K clusters consists of data points closest to its mean in Euclidean distance
3. **Update step** move the mean to the center of the cluster



wikipedia

Nearest in Euclidean-distance sense. What about other metrics?

K-Means with Generalized Metrics

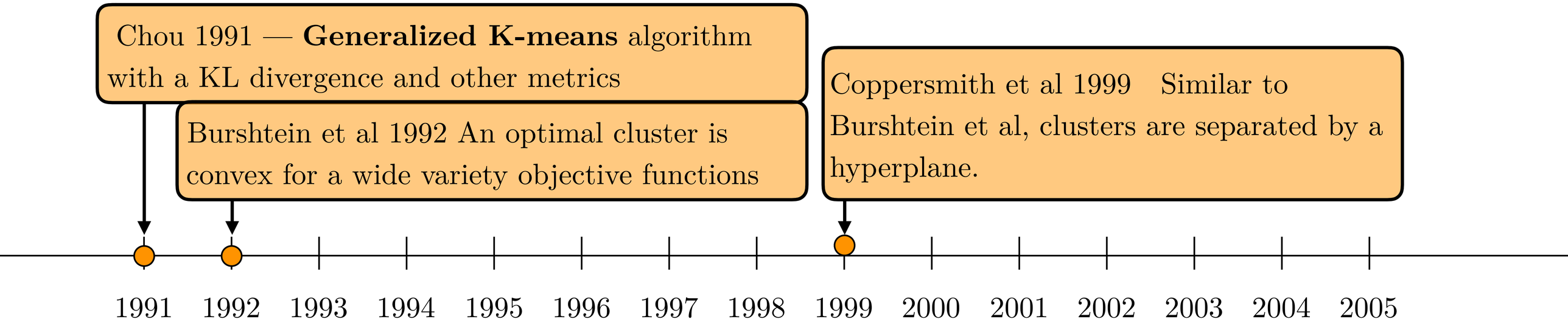
Generalize objective function:

$$\min_{Q} E\left[(\mathsf{X} - Q(\mathsf{X}))^2\right]$$

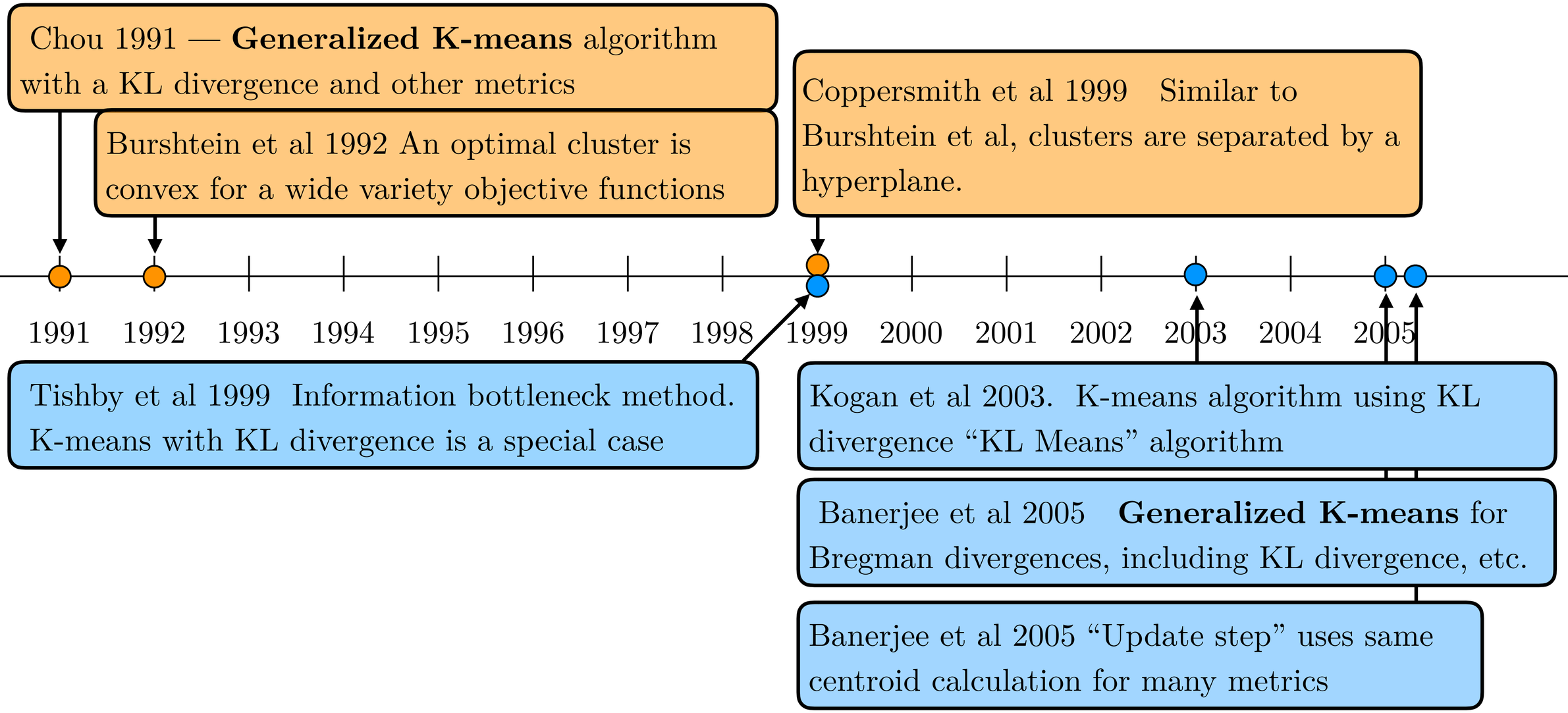
Euclidean distance

KL divergence, Gini index, Hamming error, Itakura-Saito, logistic loss, etc.

Change the Assignment/Update "local" function in the to match the objective function.



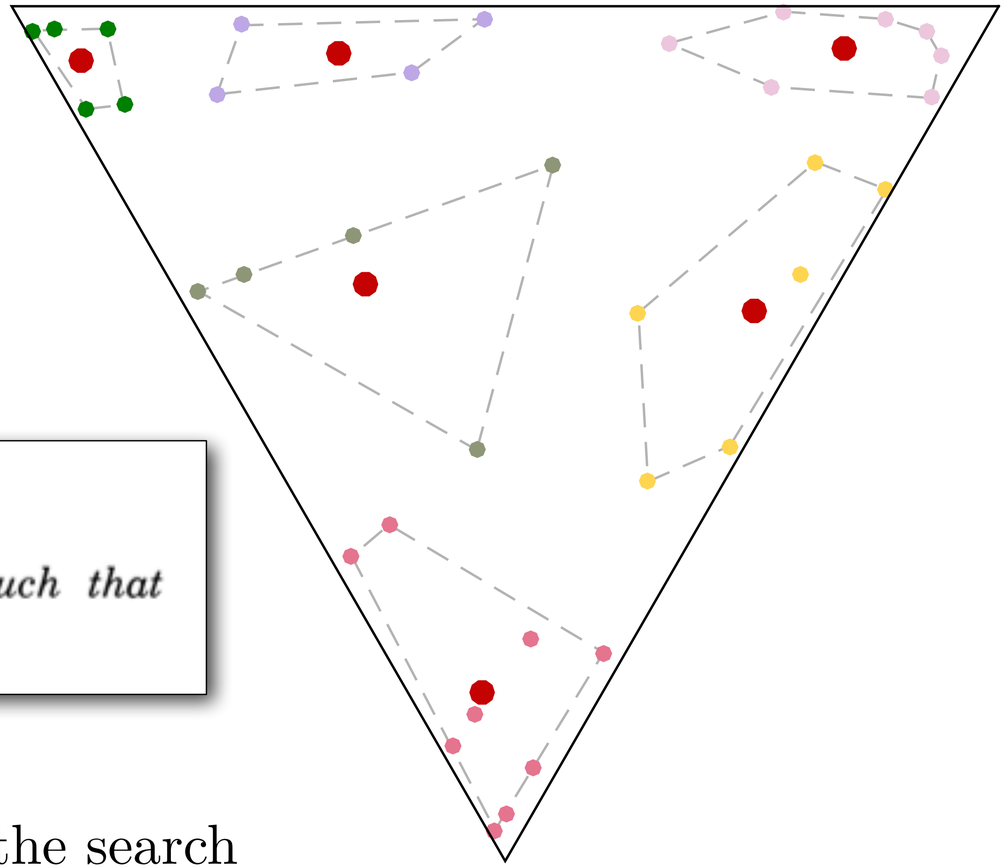
K-Means with Generalized Metrics



Two independent streams: Tishby et al, Banerjee et al did not cite the Chou stream.

Optimality of Assignment Step

Consider a broad class of convex objective functions. There exists an optimal assignment such that each cluster is a convex set in the “backward channel”



The Annals of Statistics
1992, Vol. 20, No. 3, 1637–1646

MINIMUM IMPURITY PARTITIONS

BY DAVID BURSHTEIN, VINCENT DELLA PIETRA, DIMITRI KANEVSKY AND ARTHUR NÁDAS

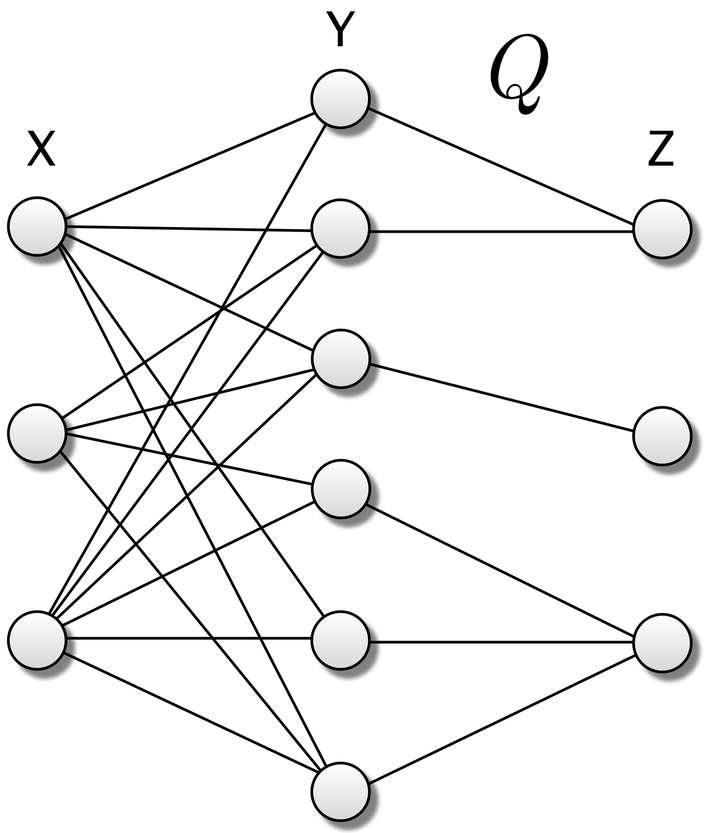
2. Results.

THEOREM 1. *For any* C: $\mathcal{X} \to \mathcal{C}$ *there exists a* $\tilde{C}$: $\mathcal{Y} \to \mathcal{C}$ *such that* $\Psi(\tilde{C}(Y)) \leq \Psi(C)$ *and such that* $\tilde{C}^{-1}(c)$ *is convex for all* $c \in \mathcal{C}$.

Significance: for a wide variety of objective functions, the search can be restricted to convex clusters.

Original Problem: Maximization of Mutual Information

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



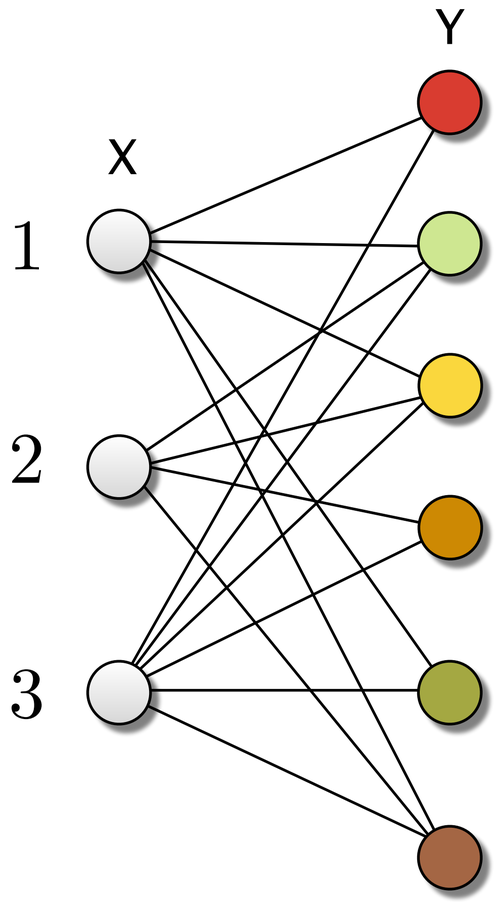
Given a discrete memoryless channel and input distribution $p_{\mathsf{XY}}(x, y)$, find the quantizer Q which maximizes mutual information:

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

with $|\mathsf{Z}| < |\mathsf{Y}|$.

$|\mathsf{Z}| \geq |\mathsf{Y}|$ is trivial.

Problem Setup and Backwards Channel



Assume $p_{\mathsf{XY}}(x, y)$ is known; X is discrete. Examples show Y is discrete, but results can be extend to continuous case. Running example:

$$\mathsf{X} \in \{1, 2, 3\}$$

$$\mathsf{Y} \in \{\text{red}, \text{lime}, \text{yellow}, \text{orange}, \text{green}, \text{brown}\}$$

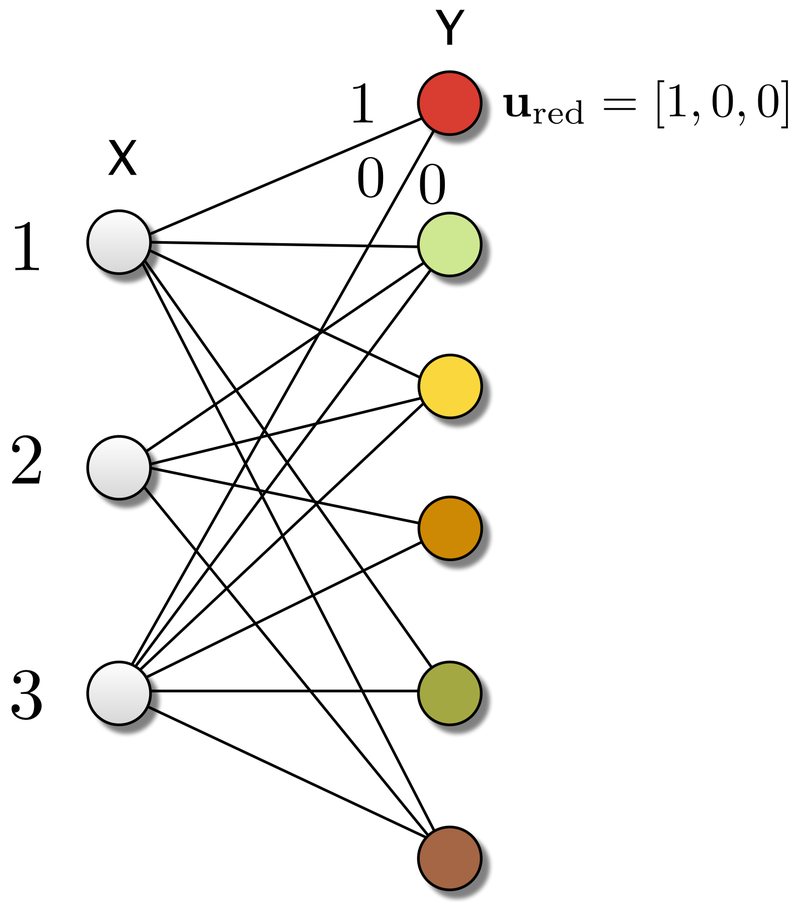
Work with the backward channel:

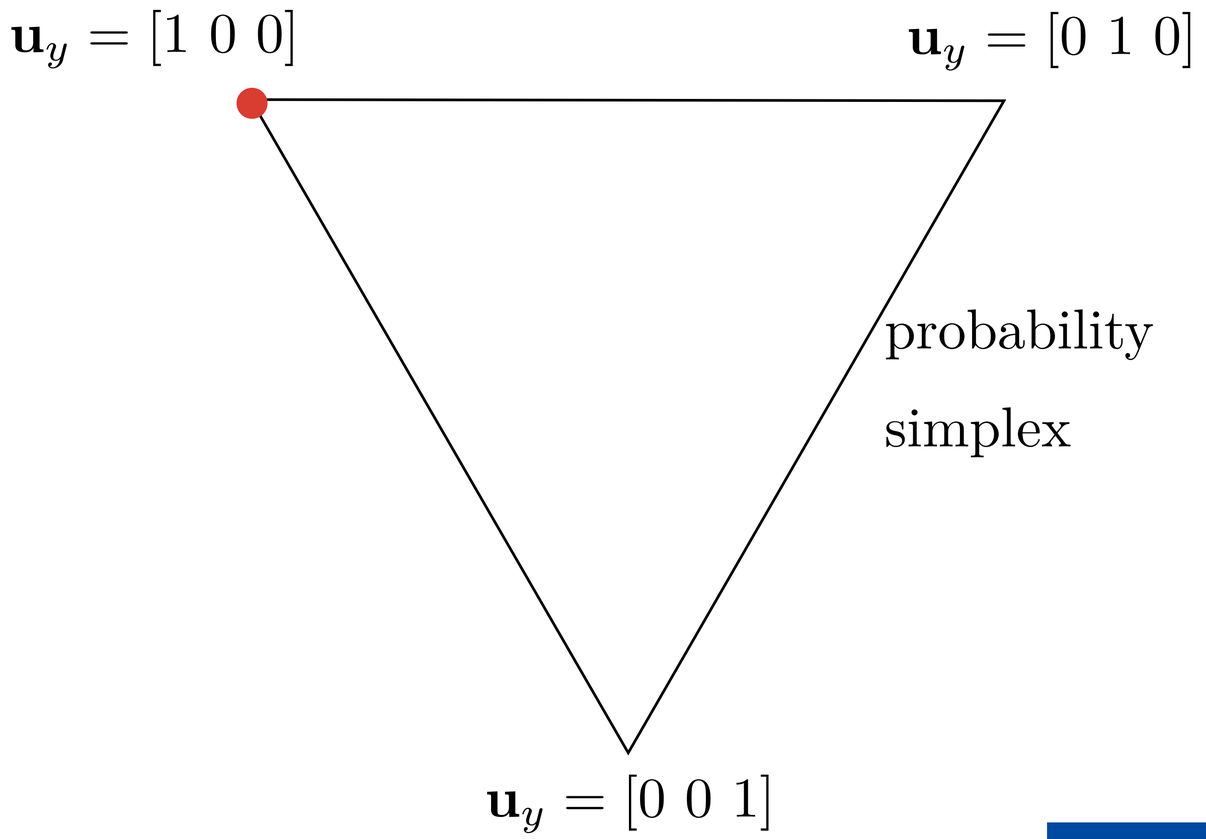
$$\mathbf{u}_y = \Big[\Pr(\mathsf{X} = 1|\mathsf{Y} = y), \ldots, \Pr(\mathsf{X} = J|\mathsf{Y} = y)\Big]$$

Justify this later.

Backward Channel $\Pr(\mathsf{X} \mid \mathsf{Y})$ as a Vector

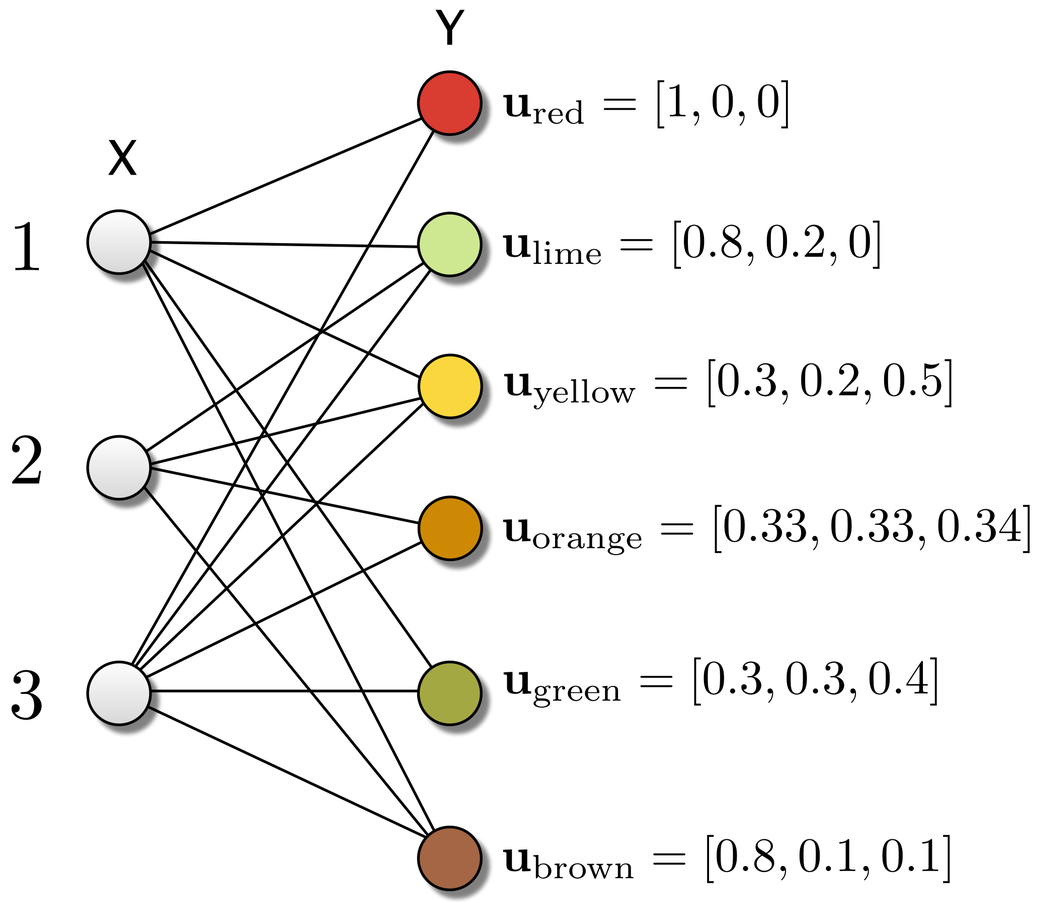
$$\mathbf{u}_y = \Big[\Pr(\mathsf{X}=1|\mathsf{Y}=y), \Pr(\mathsf{X}=2|\mathsf{Y}=y), \Pr(\mathsf{X}=3|\mathsf{Y}=y)\Big]$$

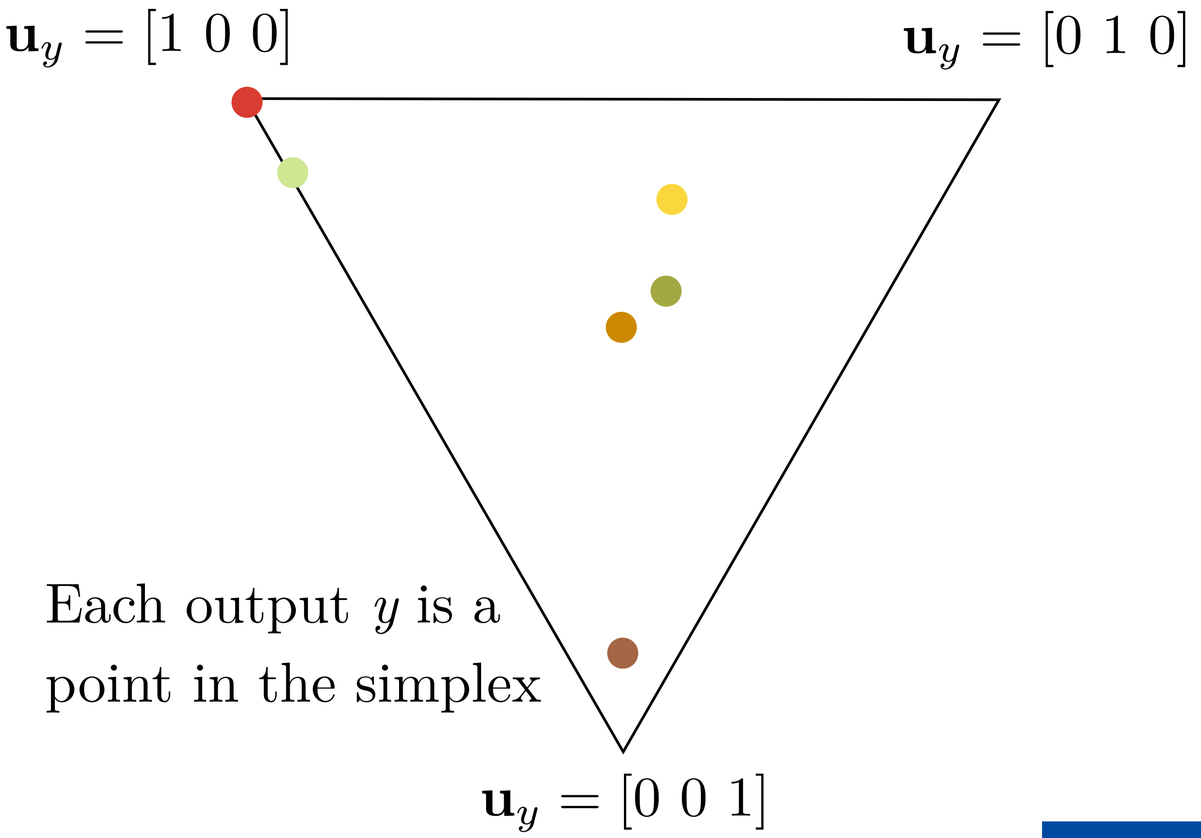




Backward Channel $\Pr(\mathsf{X} \mid \mathsf{Y})$ as a Vector

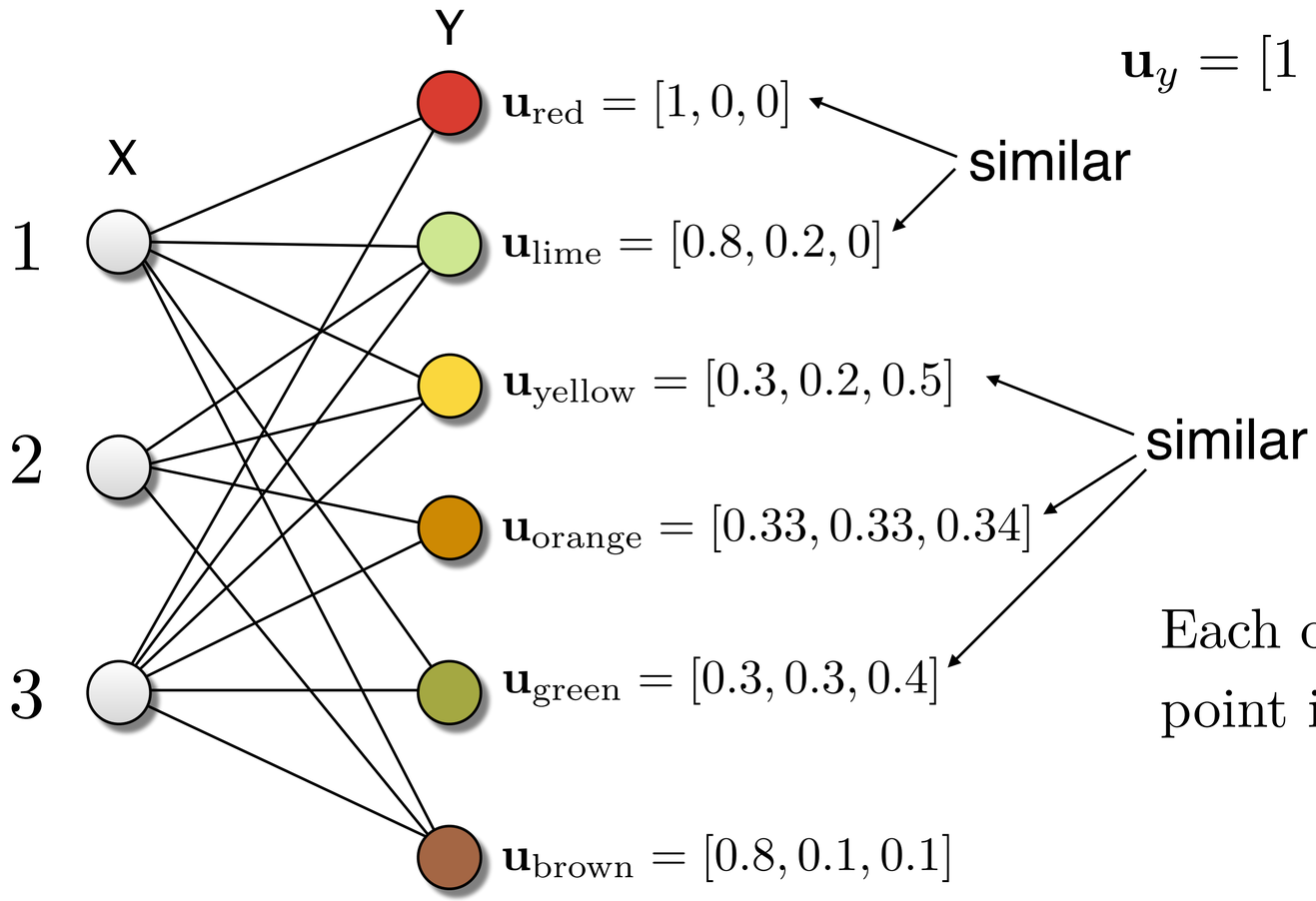
$$\mathbf{u}_y = \Big[\Pr(\mathsf{X}=1|\mathsf{Y}=y), \Pr(\mathsf{X}=2|\mathsf{Y}=y), \Pr(\mathsf{X}=3|\mathsf{Y}=y)\Big]$$

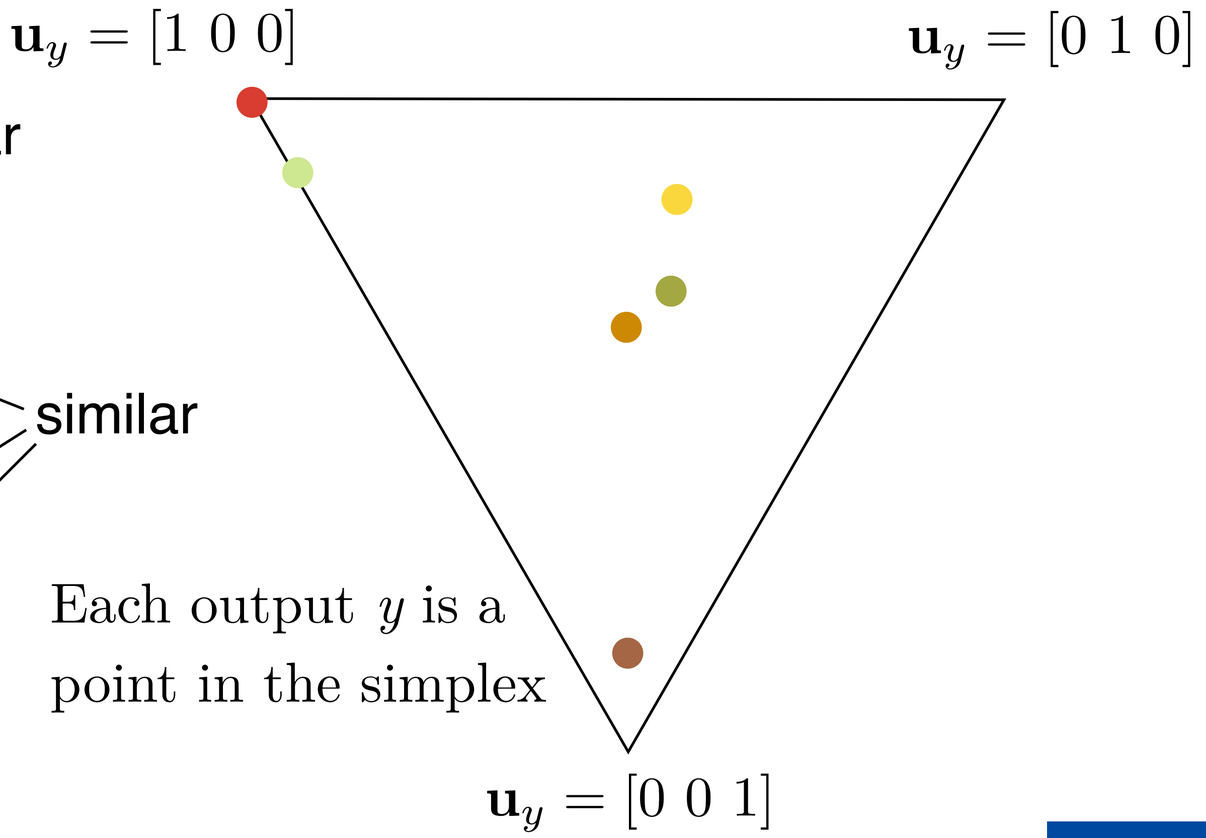




Backward Channel $\Pr(\mathrm{X} \mid \mathrm{Y})$ as a Vector

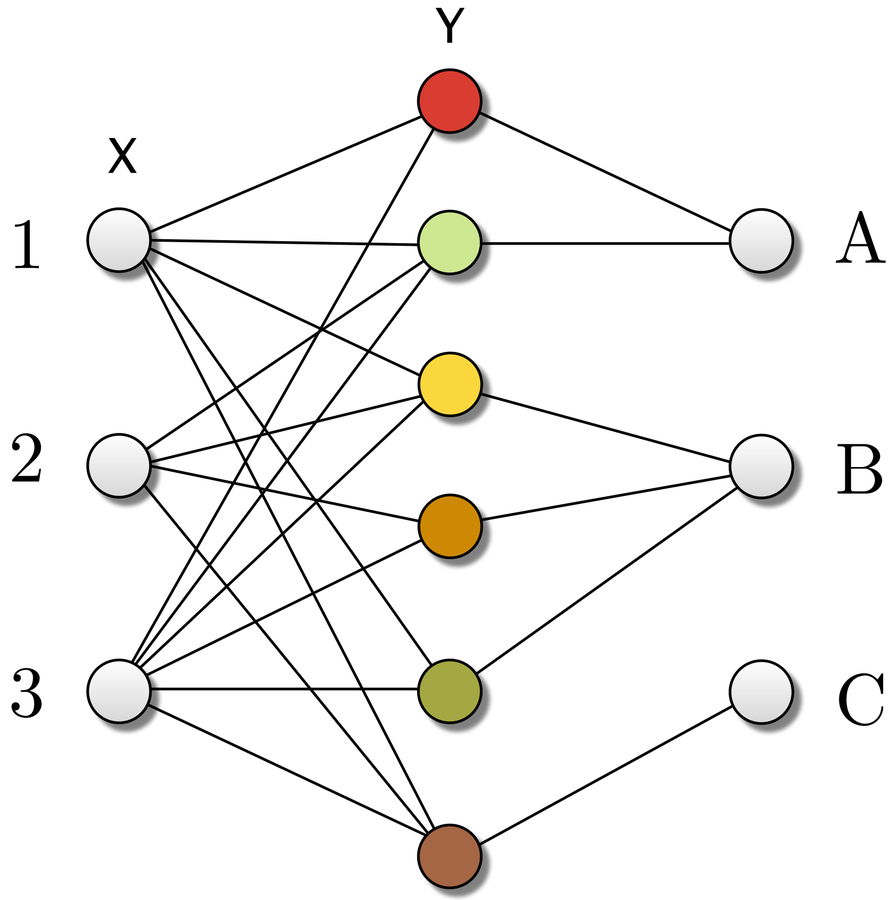
$$\mathbf{u}_y = \Big[\Pr(\mathrm{X}=1|\mathrm{Y}=y), \Pr(\mathrm{X}=2|\mathrm{Y}=y), \Pr(\mathrm{X}=3|\mathrm{Y}=y)\Big]$$

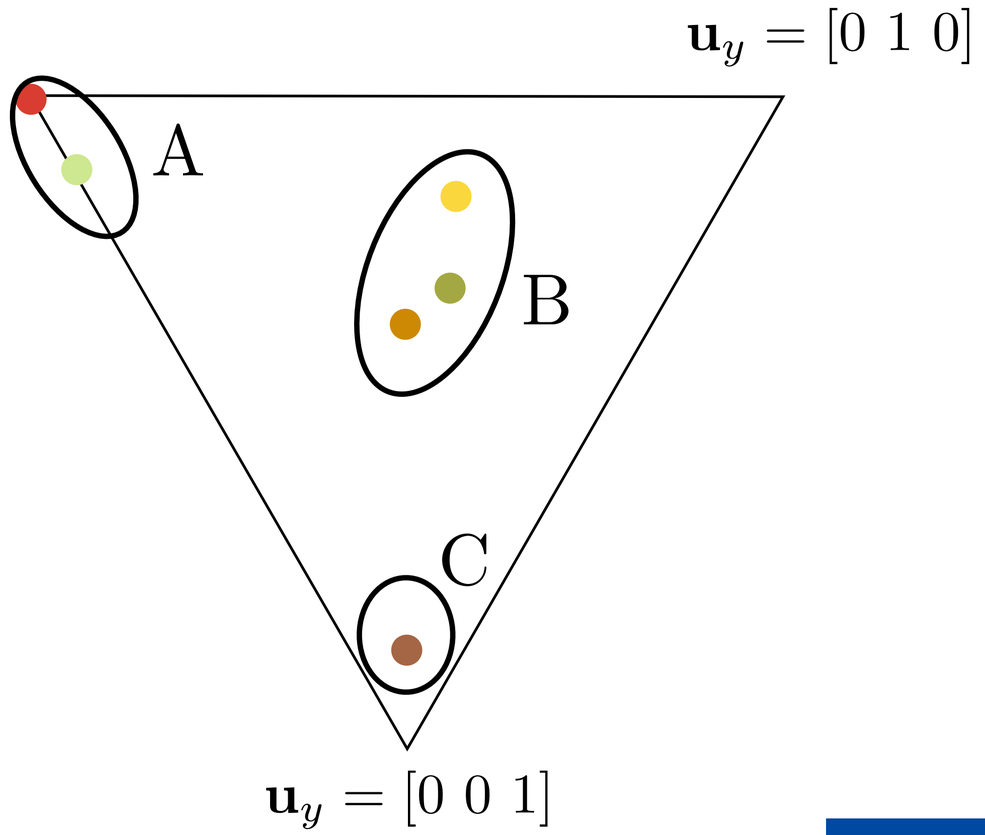




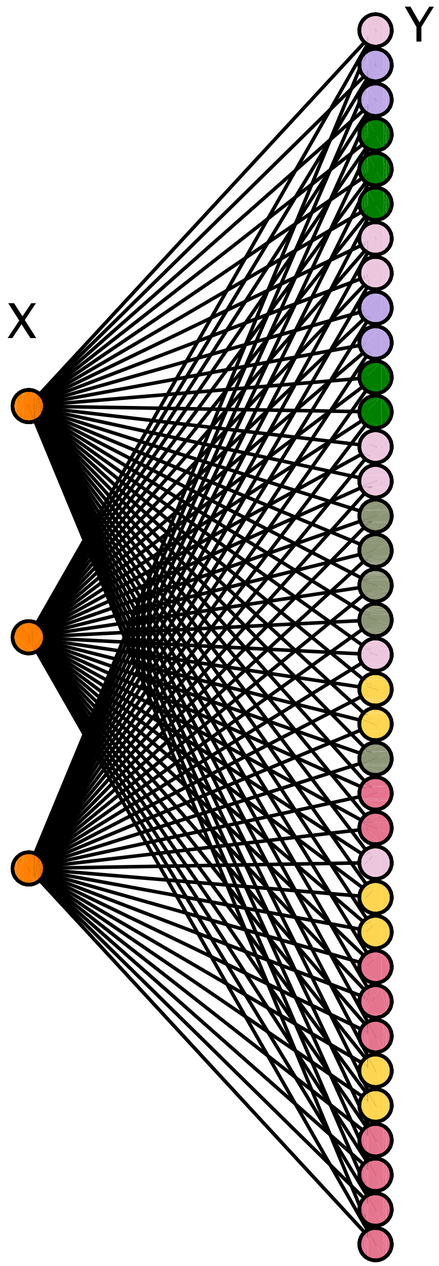
Backward Channel $\Pr(\mathrm{X} \mid \mathrm{Y})$ as a Vector

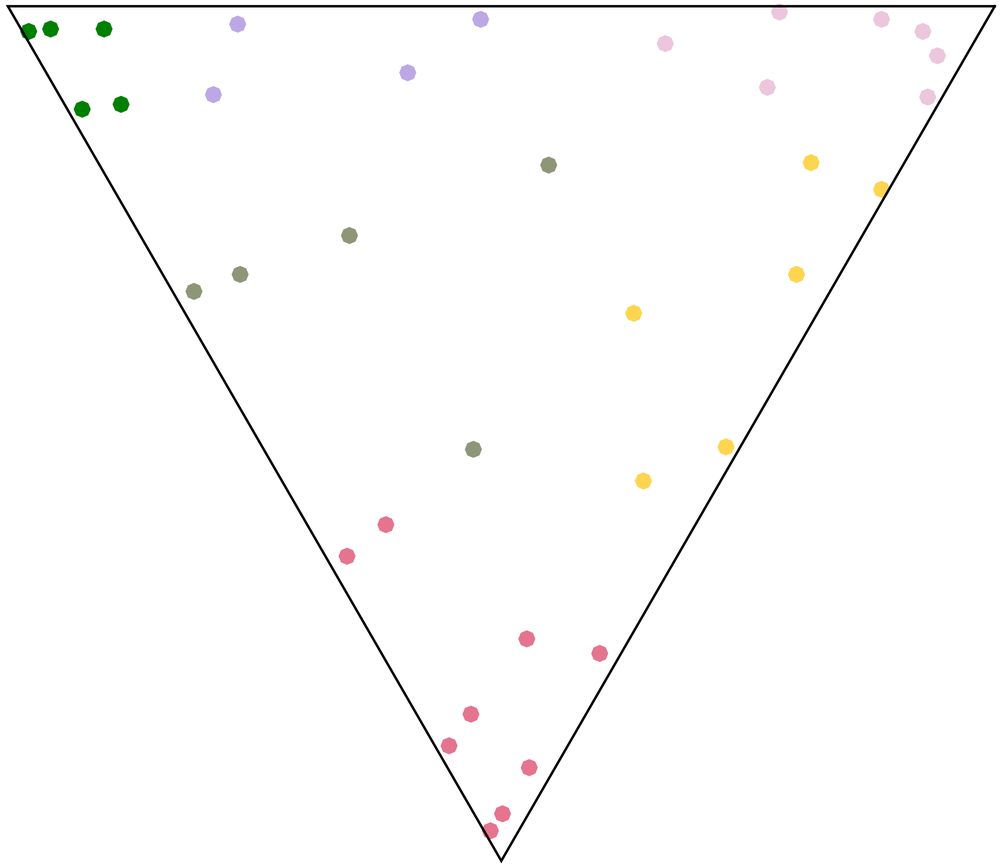
$$\mathbf{u}_y = \Big[\Pr(\mathrm{X}=1|\mathrm{Y}=y), \Pr(\mathrm{X}=2|\mathrm{Y}=y), \Pr(\mathrm{X}=3|\mathrm{Y}=y)\Big]$$



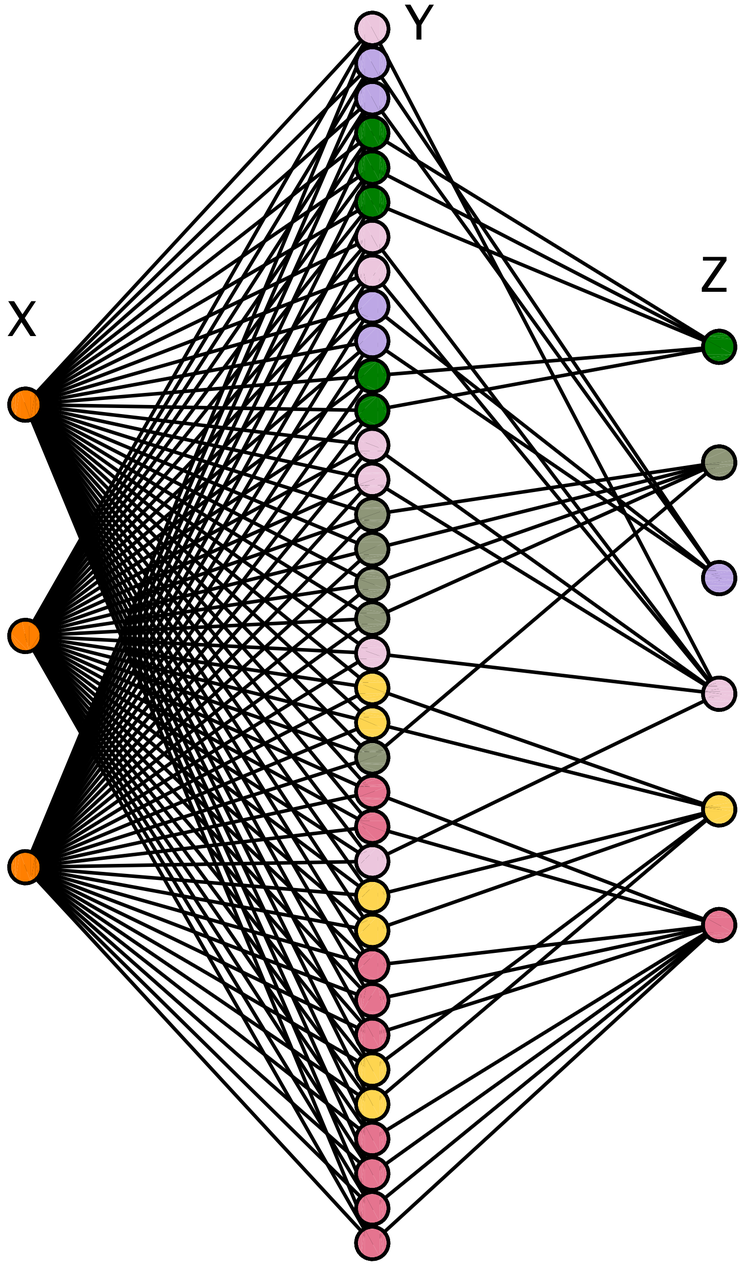


Visualizing the “Means”

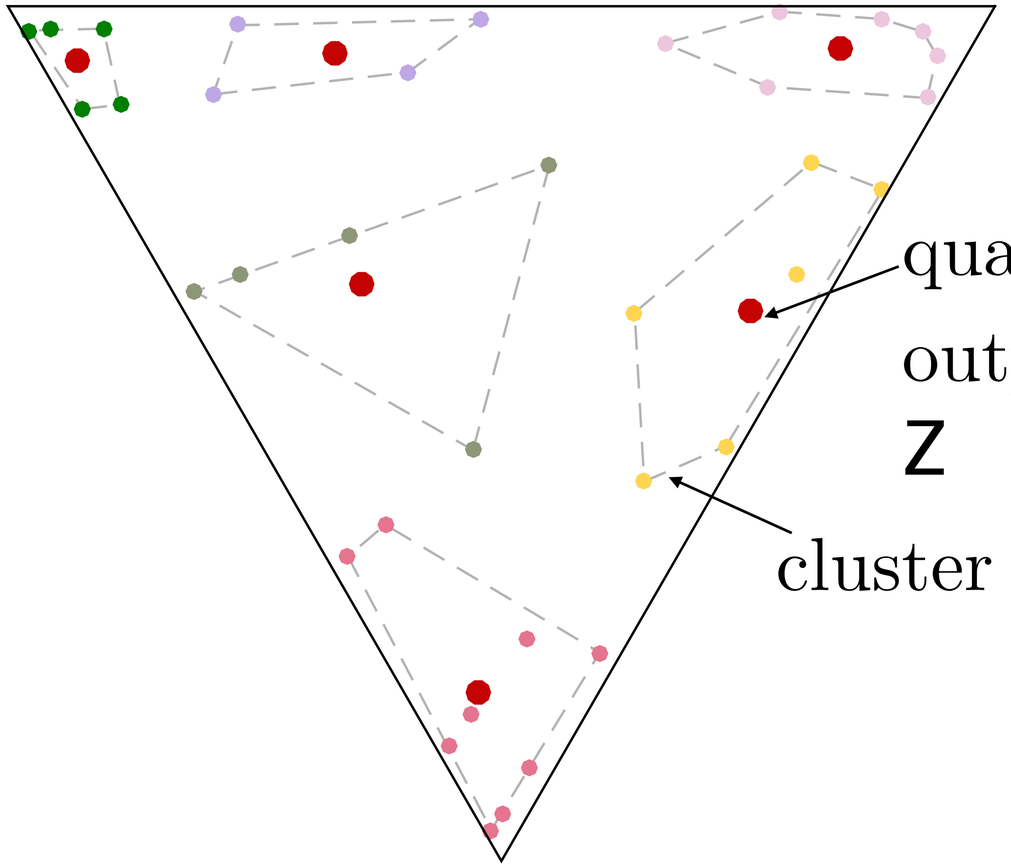




Visualizing the “Means”



$p_{X|Z}(x|z)$ as well as $p_{X|Y}(x|y)$



quantizer outputs Z

cluster

From Mutual Information to KL Divergence

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

$$\text{Random vector versions: } \mathbf{U} = \big[\Pr(\mathsf{X}=1|\mathsf{Y}), \Pr(\mathsf{X}=2|\mathsf{Y}), \ldots, \Pr(\mathsf{X}=J|\mathsf{Y})\big]$$
$$\mathbf{V} = \big[\Pr(\mathsf{X}=1|\mathsf{Z}), \Pr(\mathsf{X}=2|\mathsf{Y}), \ldots, \Pr(\mathsf{X}=J|\mathsf{Z})\big]$$

Then, the following holds:

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

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

$$Q^* = \arg\max_Q I(\mathsf{X};\mathsf{Z}) = \arg\min_Q E\big(D(\mathbf{U}||\mathbf{V})\big)$$

Thus, maximization of mutual information is minimization of KL divergence.

Three Particular Cases

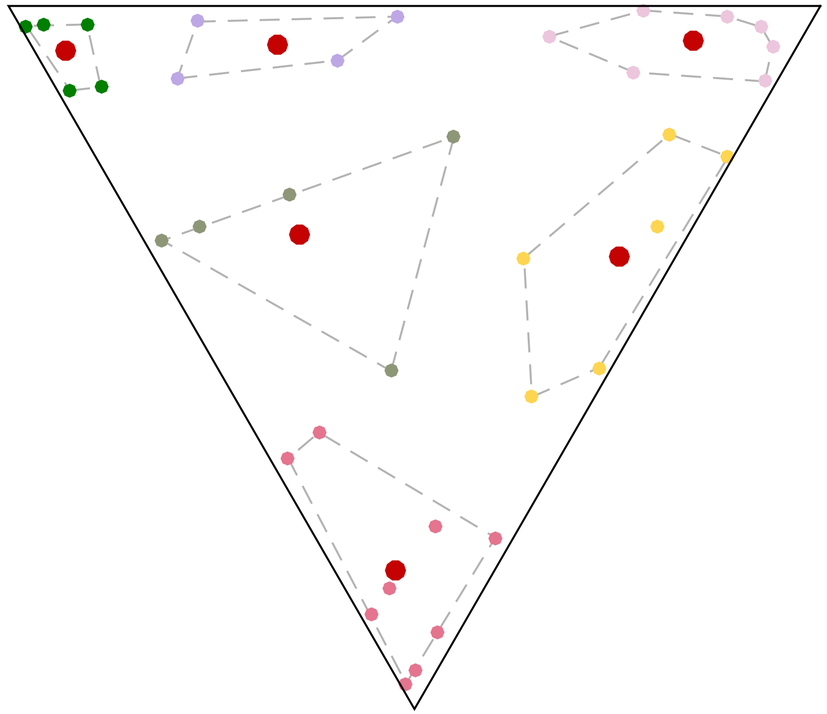
1. General discrete quantization:
    1. K-Means with KL Divergence metrics “KL-Means”
    2. Information bottleneck method
2. Binary-input, discrete output quantization: Dynamic programming quantization
3. Binary-input, continuous-output quantization with arbitrary noise

K-Means With KL Divergence Metric

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

Min KL divergence = max. mutual information

$$Q^* = \arg\max_Q I(\mathsf{X};\mathsf{Z}) = \arg\min_Q E\big(D(\mathbf{U}||\mathbf{V})\big)$$



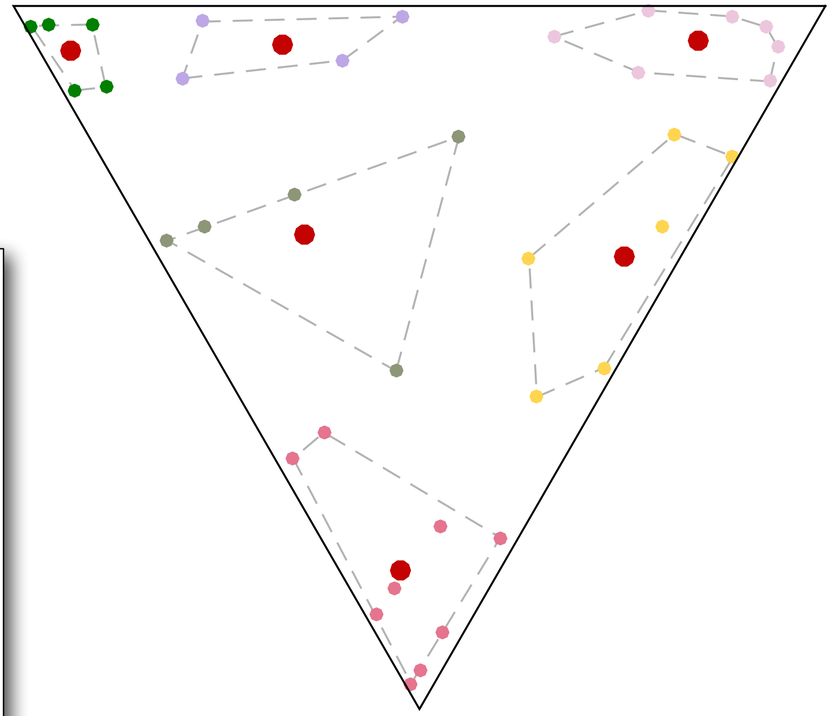
A. Zhang and B. Kurkoski, "Low-Complexity Quantization of Non-Binary Input DMCs" ISITA 2016.

K-Means With KL Divergence Metric

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

Min KL divergence = max. mutual information

$$Q^* = \arg\max_Q I(\mathsf{X};\mathsf{Z}) = \arg\min_Q E\big(D(\mathbf{U}||\mathbf{V})\big)$$



TEXT MINING WITH INFORMATION–THEORETIC CLUSTERING

Motivated by the success of hybrid information-retrieval algorithms, the authors report on the development of their hybrid clustering scheme. Scheme experiments on data in a reduced vector space model indicate a higher performance level over several existing clustering algorithms.

In text mining and information retrieval, as well as in other applications such as bioin- partitions for the second step of the procedure. The second step is a modification of a new divi-

A. inary Input DMCs" ISITA 2016.

K-Means With KL Divergence Metric

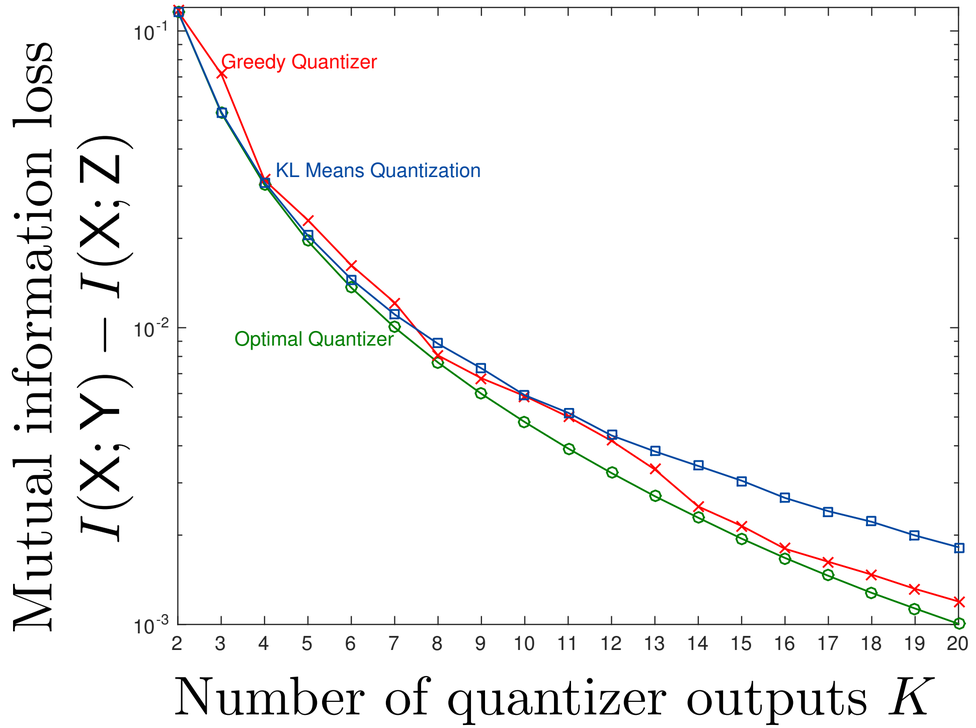
"KL-Means algorithm" replace Euclidean distance with KL distance

Min KL divergence = max. mutual information

$$Q^* = \arg\max_Q I(\mathsf{X};\mathsf{Z}) = \arg\min_Q E\big(D(\mathbf{U}||\mathbf{V})\big)$$

Numerical results show tradeoff:

- increasing number of quantizer outputs
- decreases the loss of mutual information



A. Zhang and B. Kurkoski, "Low-Complexity Quantization of Non-Binary Input DMCs" ISITA 2016.

Information Bottleneck Method

For Markov chain with joint distribution:

$$\mathsf{X} \to \mathsf{Y} \overset{Q}{\to} \mathsf{Z}$$

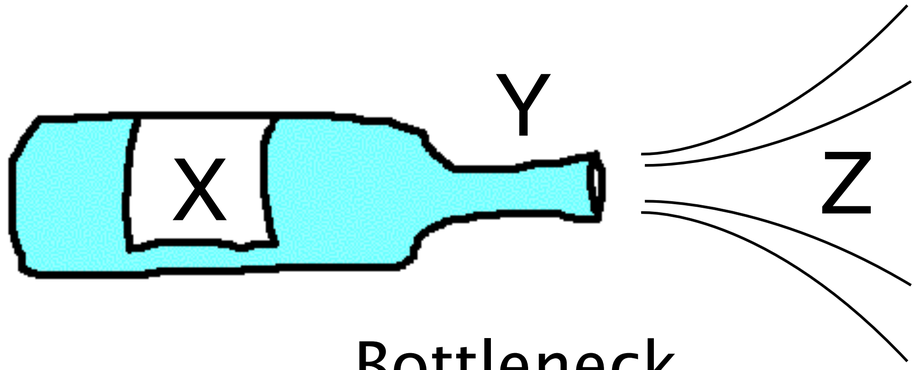
Tishby, Pereira, Bialek (1999) gave the information bottleneck method which finds:

$$\min_{p_{\mathsf{Z}|\mathsf{Y}}(z|y)} I(\mathsf{Y};\mathsf{Z}) - \beta I(\mathsf{X};\mathsf{Z})$$

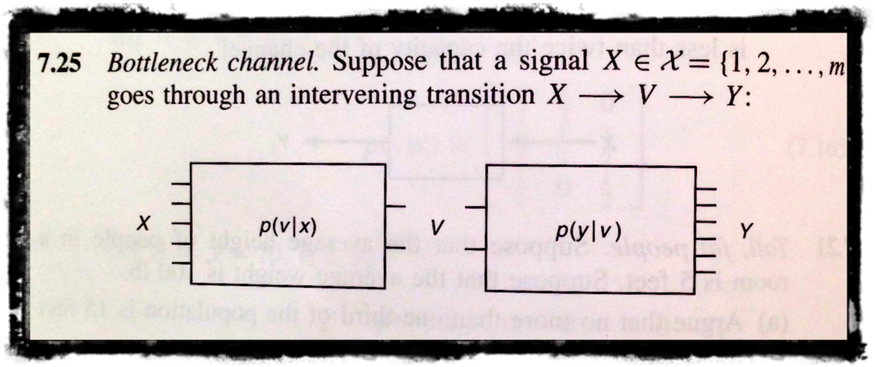
When:

- β finite: probabilistic clustering $p_{\mathsf{Z}|\mathsf{Y}}(z|y)$
- $\beta \to \infty$: hard clustring, $p_{\mathsf{Z}|\mathsf{Y}}(z|y) = 0$ or 1

Q is the probability distribution $p_{\mathsf{Z}|\mathsf{Y}}(z|y)$



Bottleneck



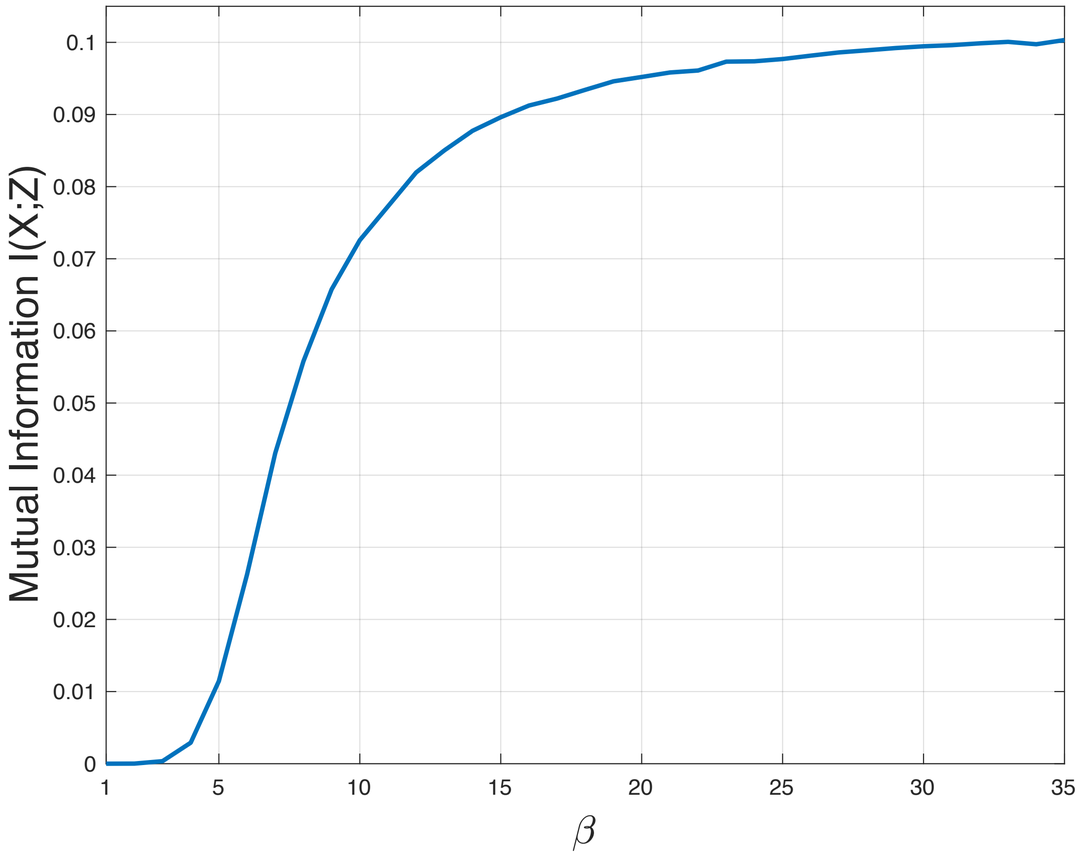
Cover and Thomas, *Elements of Information Theory*, 2006. Problem 7.52 on page 234.

Image credit: Boris Epshtein & Lena Gorelick

Information Bottleneck Method

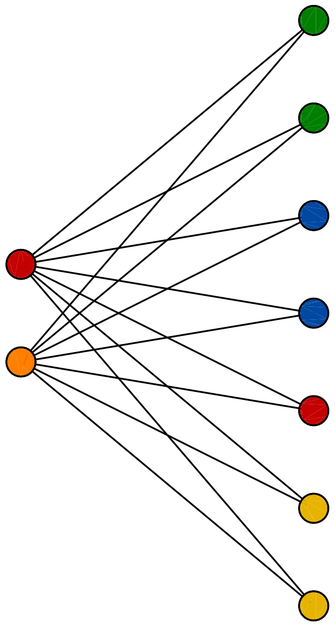
$\beta \geq 0$ sweeps a kind of rate-distortion curve.

When $\beta \to \infty$ we get hard clustering. In fact, the informaiton bottleneck method becomes K-means algorithm with KL divergence metric. "KL means algorithm"



B. M. Kurkoski, "On the relationship between the KL means algorithm and the information bottleneck method," in 11th International ITG Conference on Systems, Communications and Coding (SCC2017), (Hamburg, Germany), pp. 1-5, February 2017.

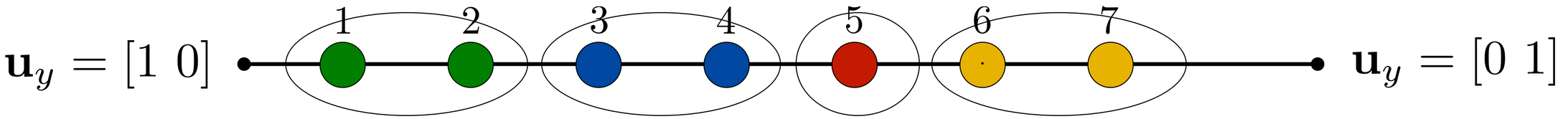
Quantizer Design for Binary Input DMC



- Consider a binary-input channel
- For an optimal quantizer, there must exist an convex cluster (Burshtein et al)
- Convex in the space:

$$u_y = \left[\Pr[\mathsf{X} = 0 | \mathsf{Y} = y], \Pr[\mathsf{X} = 1 | \mathsf{Y} = y] \right]$$

- **Dynamic programming** approach is suggested: search over all boundaries
- Algorithm is optimal, complexity is $|\mathcal{Y}|^3$

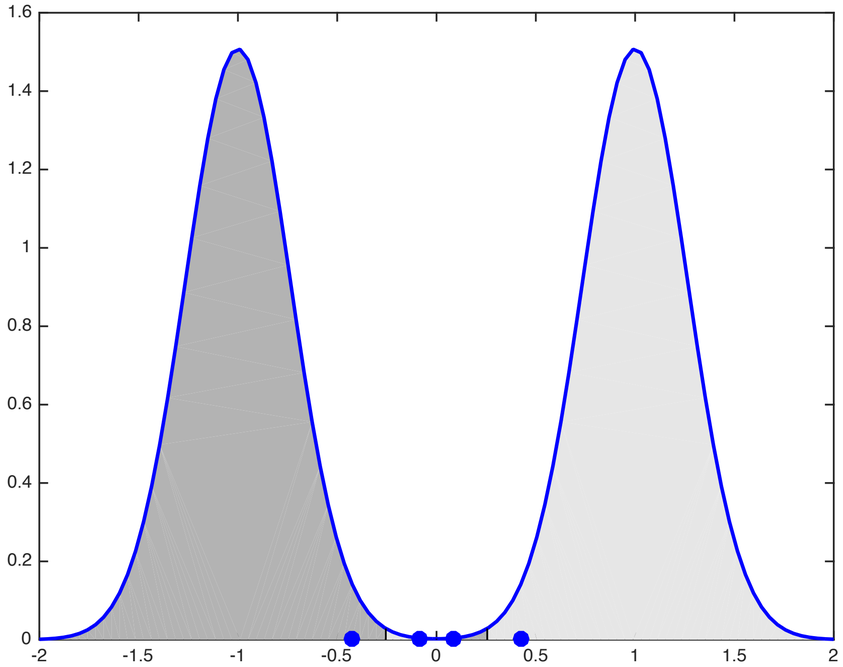


B. Kurkoski and H. Yagi, "Quantization of Binary-Input Discrete Memoryless Channels," IEEE Trans on Information Theory, May 2014.

Structure of Optimal Solution

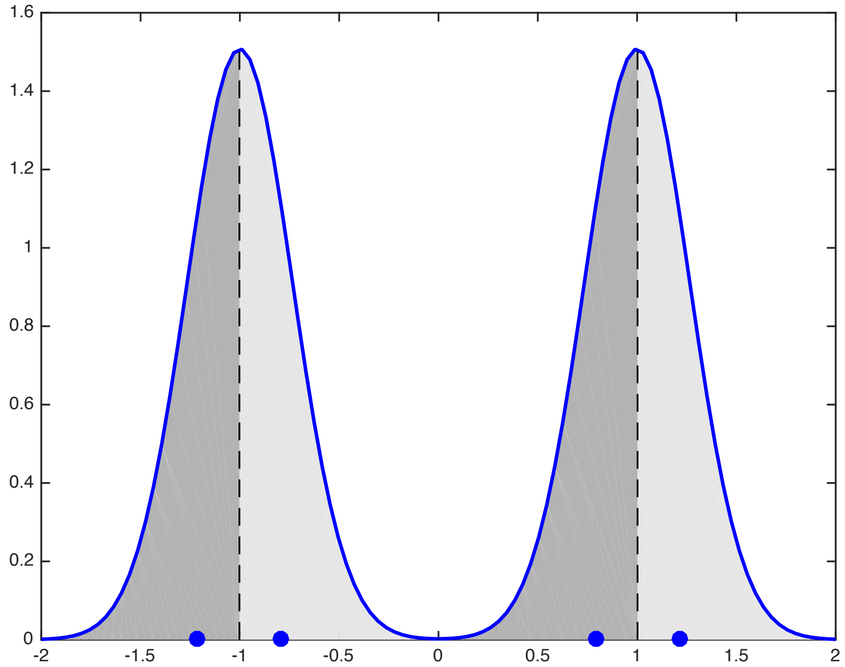
$$\text{variance} = 0.07$$

KL-Means



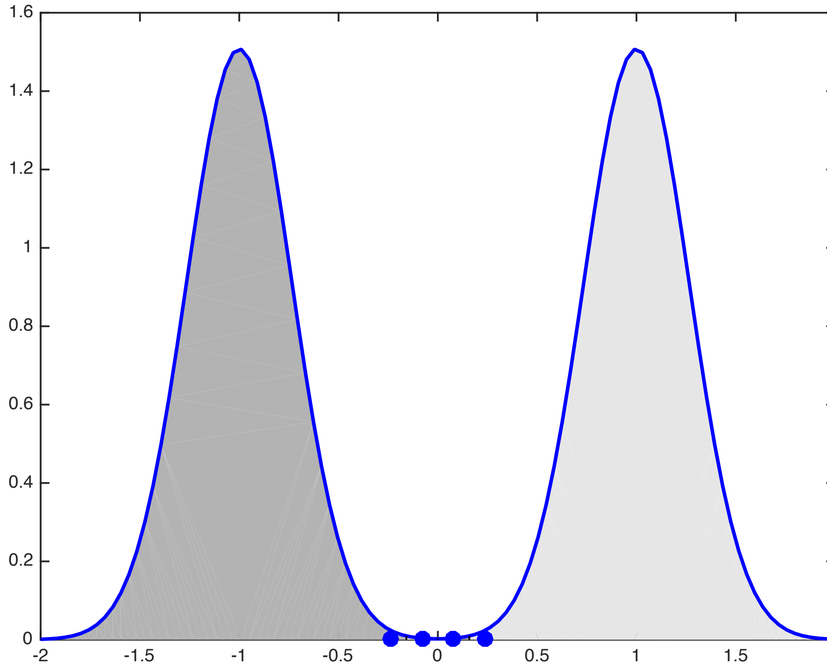
$$\mathrm{I(X;Z)} = 0.9995$$

K-Means



$$\mathrm{I(X;Z)} = 0.9989$$

Uniform

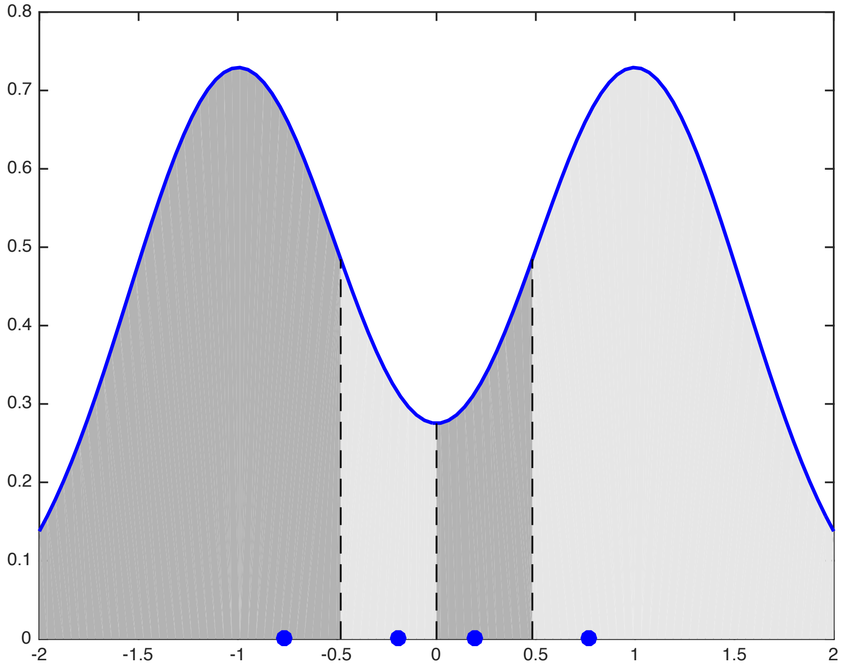


$$\mathrm{I(X;Z)} = 0.9995$$

Structure of Optimal Solution

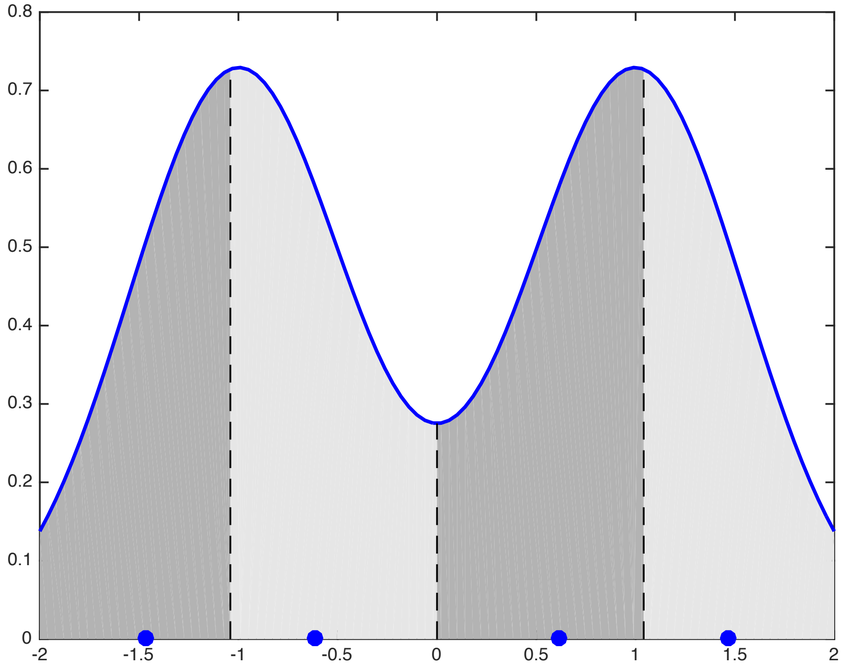
$$\text{variance} = 0.3$$

KL-Means



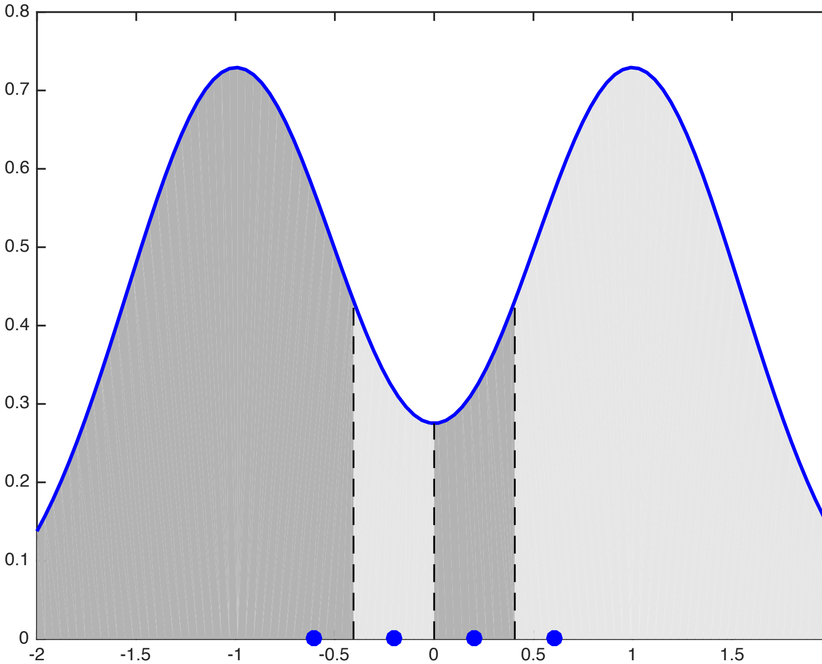
$$\mathrm{I(X;Z)} = 0.8529$$

K-Means



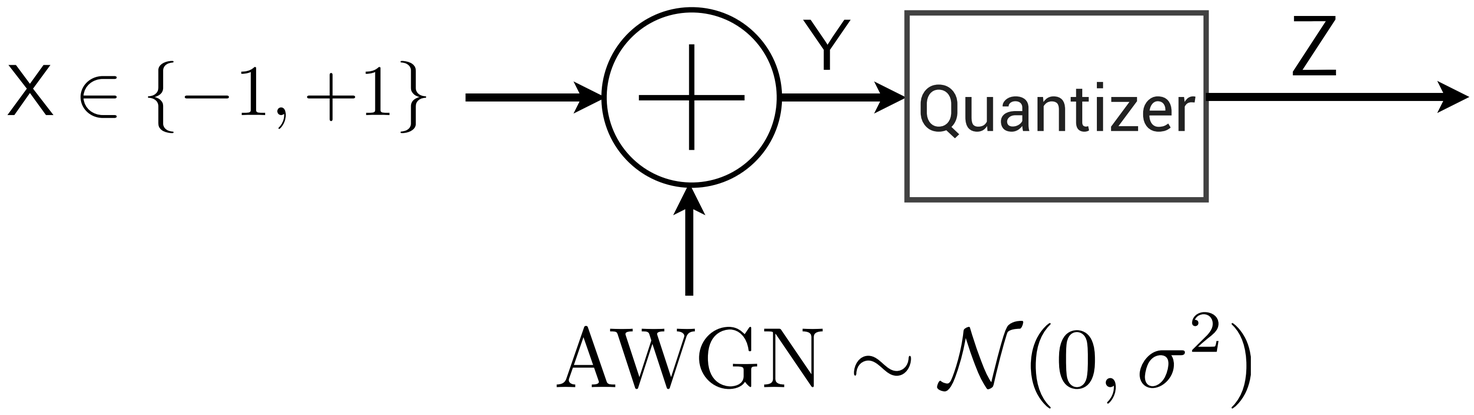
$$\mathrm{I(X;Z)} = 0.8171$$

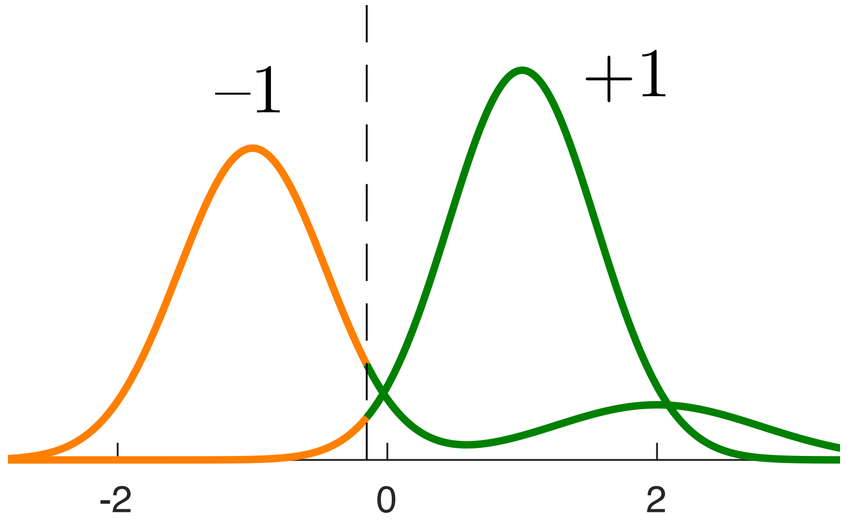
Uniform



$$\mathrm{I(X;Z)} = 0.8541$$

Continuous-Output Channel Quantization





Consider binary-input, continuous-output $y \in \mathbb{R}$ channel:

- For BI-AWGN solution is obvious, i.e. threshold at 0
- Arbitrary, data-dependent noise $p_{\mathsf{Y}|\mathsf{X}}(y|+1)$ different from $p_{\mathsf{Y}|\mathsf{X}}(y|-1)$
- Quantization to **one bit**
- Using Burshtein et al Theorem, will not restrict ourselves to a single threshold

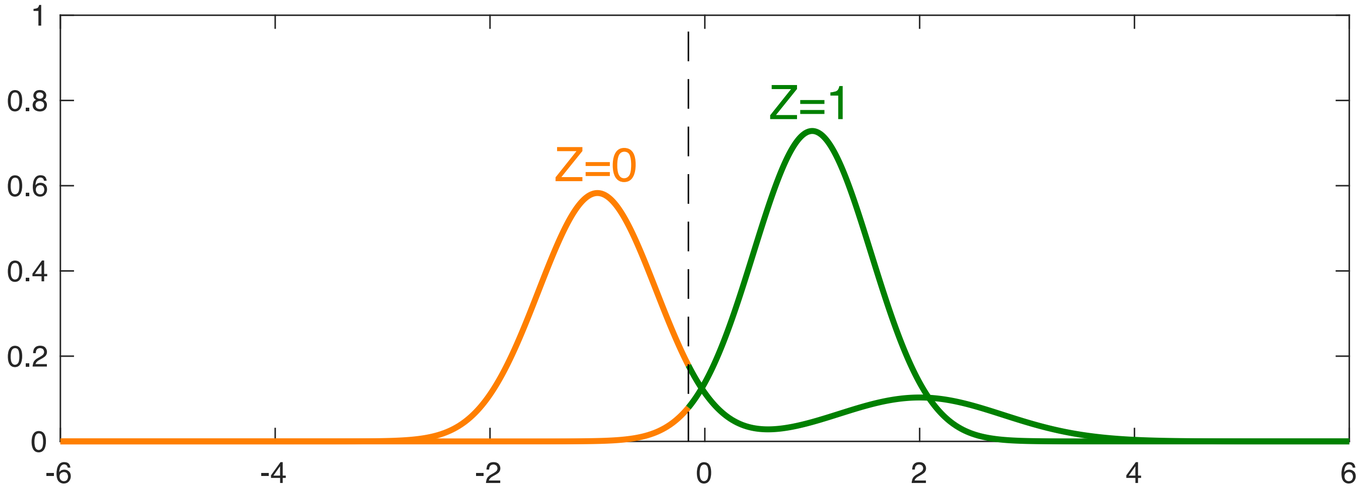
Quantization By Threshold Search

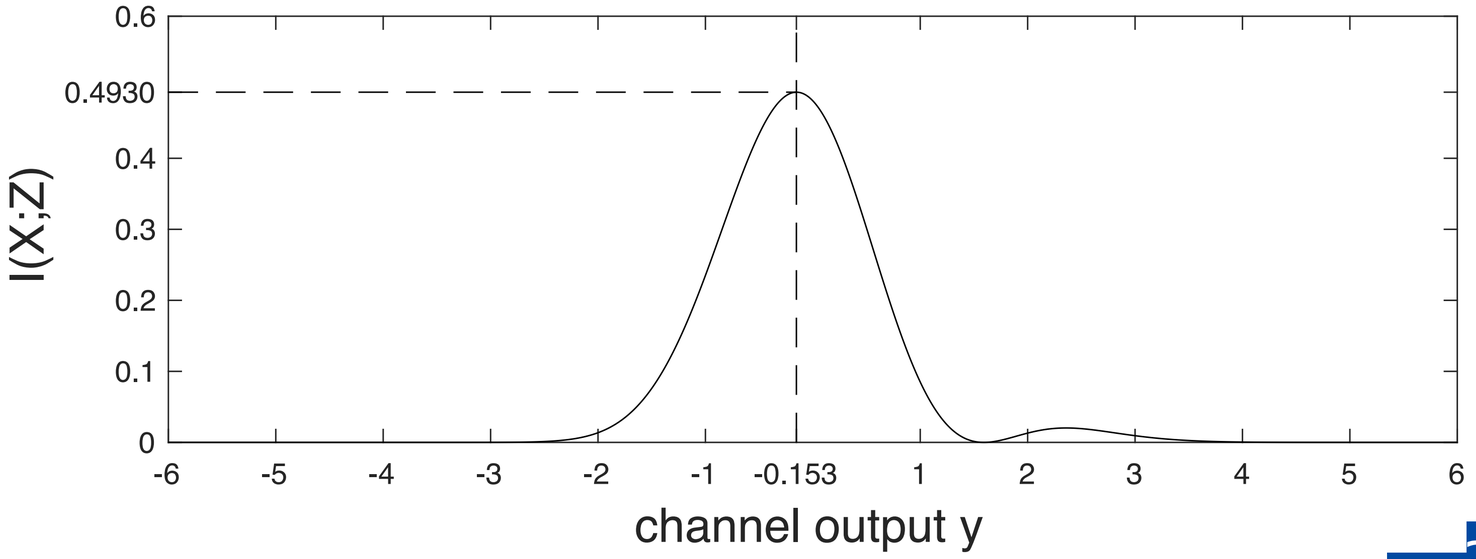
Quantizer:

$$z = \begin{cases} 0 & y < a \\ 1 & y \geq a \end{cases}$$

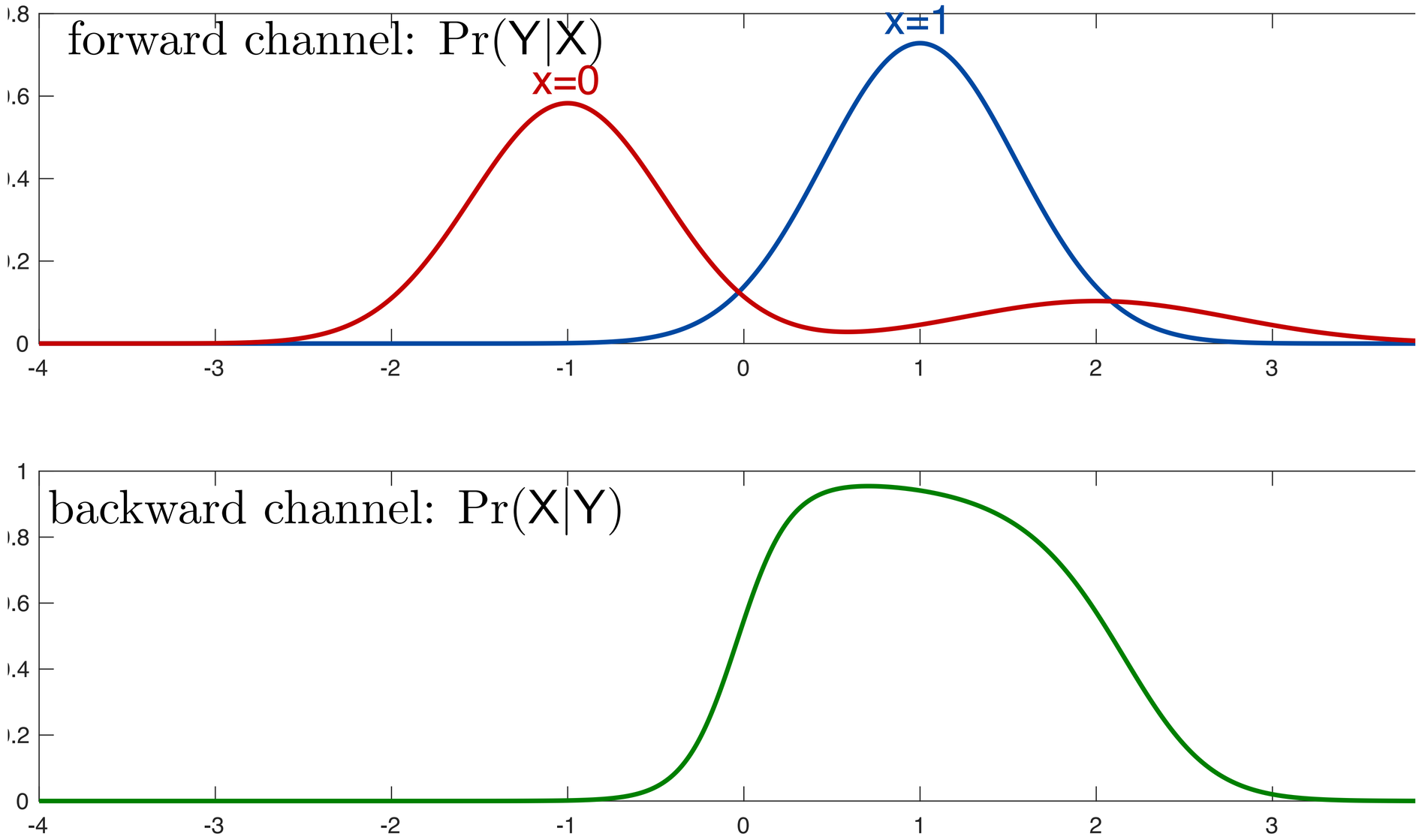
Example: data-dependent Gaussian noise mixtures

max I(X;Z) = 0.493 at
a^* = -0.153





Backward Channel $\Pr(X = x|Y = y)$

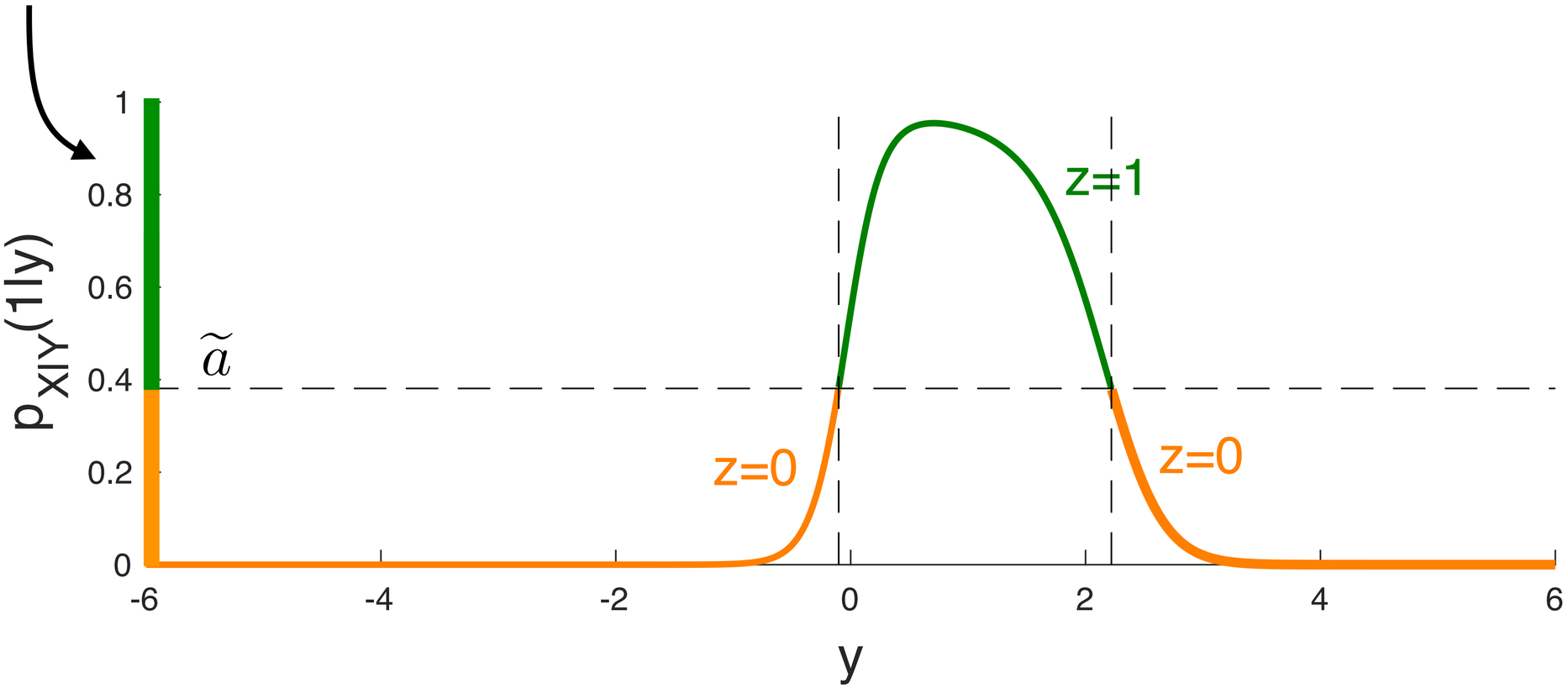


Convex Backward Channel Quantizer

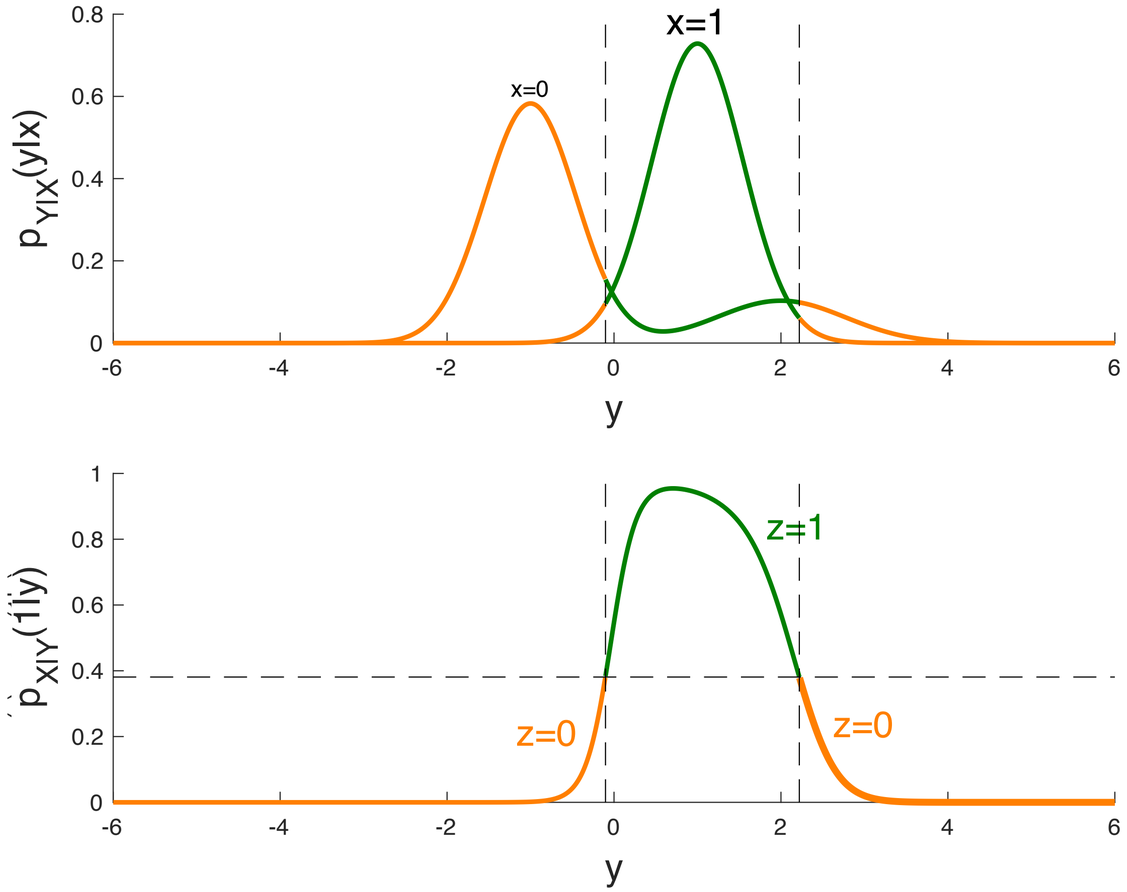
Threshold $\widetilde{a}$. Backward channel quantizer $\widetilde{Q}$:

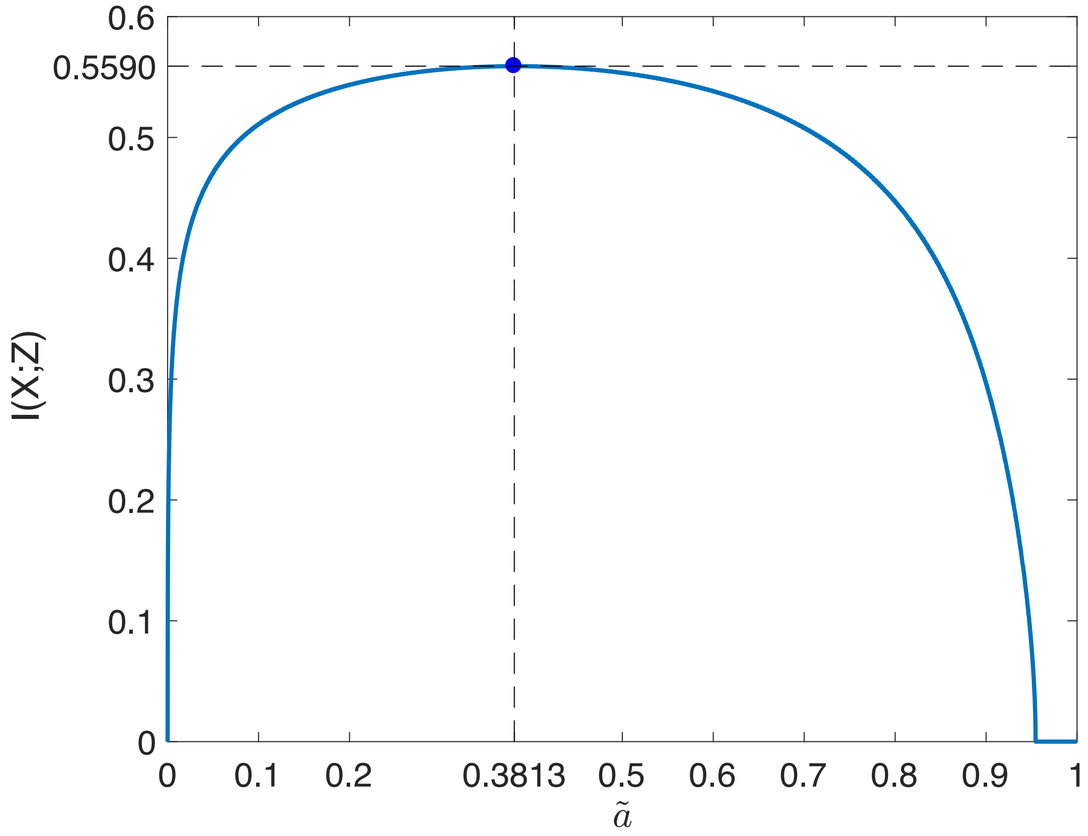
$$z = \begin{cases} 0 & \Pr(X=1|y) < \widetilde{a} \\ 1 & \Pr(X=1|y) \geq \widetilde{a} \end{cases}$$

z preimage is convex



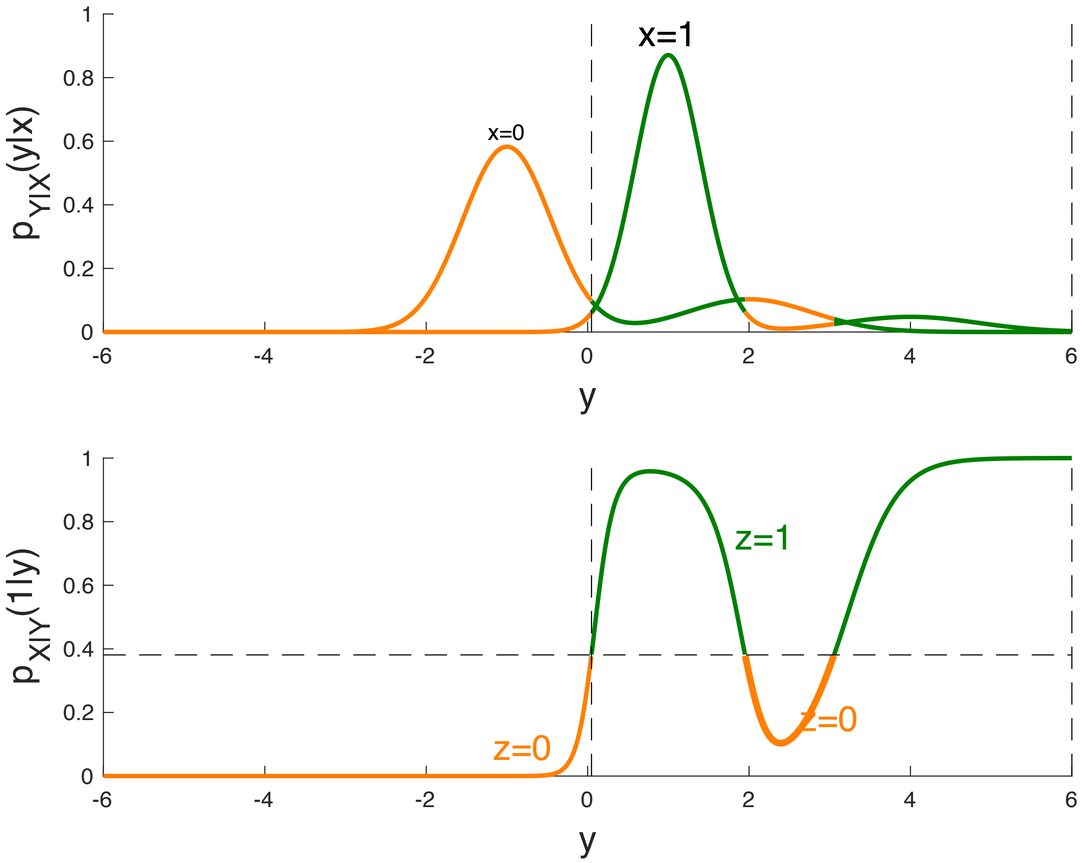
Optimal Quantizer in Backward Channel

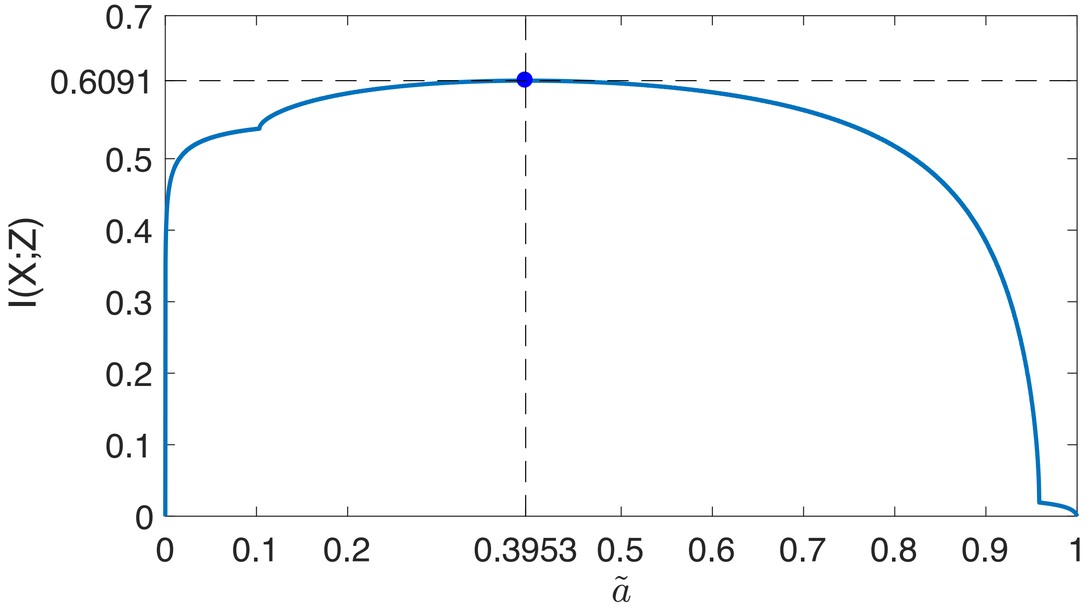




Happens to be concave. But not in general

Even Backwards Channel is Hard





Mutual information is not convex in $\widetilde{a}$

Optimal Convex Forward Quantizer

Lemma 2 If the channel log-likelihood ratio satisfies:

$$\log \frac{\Pr(y|X=1)}{\Pr(y|X=0)} \leq \log \frac{\Pr(y'|X=1)}{\Pr(y'|X=0)}$$

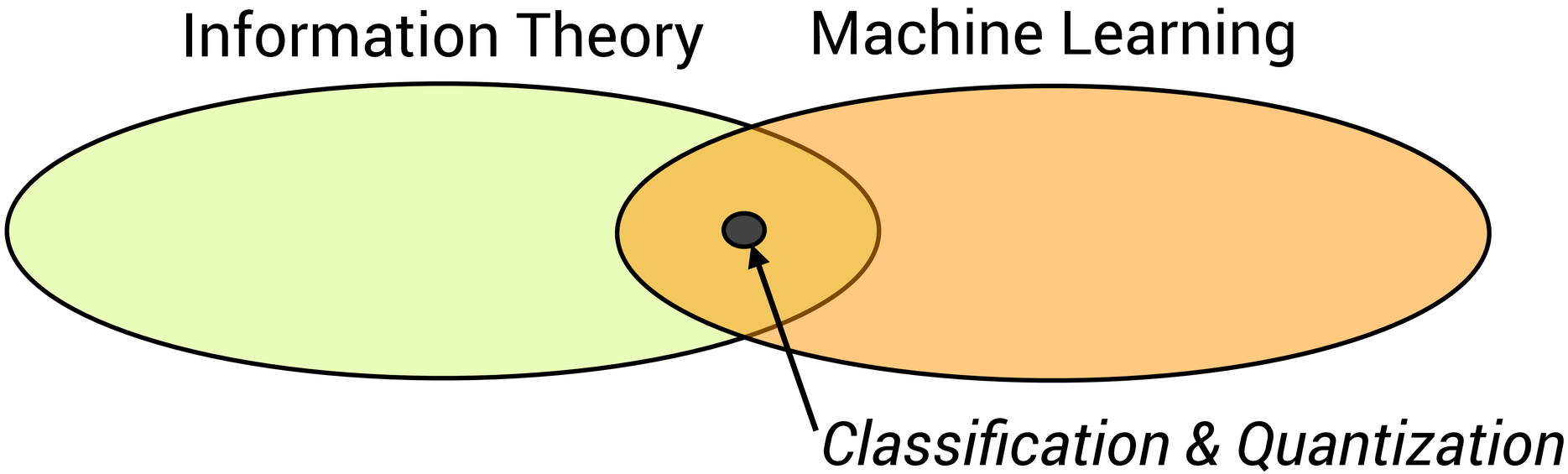
for all $y < y'$, then there exists an optimal forward channel quantizer Q^* which is a convex quantizer.

Consequences:

- For many well-behaved channels, the optimal forward channel quantizer is convex.
- The BI-AWGN channel satisfies this condition

B. M. Kurkoski and H. Yagi, "Single-bit quantization of binary-input, continuous-output channels," in Proceedings of IEEE International Symposium on Information Theory, (Aachen, Germany), pp. 2088-2092, June 2017

Channel Quantization and Classification



- Before 2002, little interest in channel quantization to maximize mutual information.
- But in the 1990s the pattern recognition community had developed necessary tools
- K-means with KL divergence/information bottleneck is effective but suboptimal way to quantize channels
- For binary-input, discrete-output channels, optimality is possible
- For binary-input, continuous output channels, still work to be done