From information theory to unsupervised learning

Susanne Still
University of Hawaii at Manoa

Reminder

- We derived the "information bottleneck" optimization problem from thermodynamic argument about physical limits of data representation efficiency.
- From the optimization problem, we derived an equation that the optimal data representation must obey.

IB algorithm

- Solve self-consistently:

$$p(y|x) = \frac{p(y)}{Z(x,\alpha)} e^{-\alpha D_{KL}[p(z|x)\|p(z|y)]}$$

$$Z(x,\alpha) = \sum_y p(y) e^{-\alpha D_{KL}[p(z|x)\|p(z|y)]}$$

$$p(y) = \sum_x p(y|x) p(x)$$

$$p(z|y) = \frac{1}{p(y)} \sum_x p(y|x) p(z,x)$$

Applications

- Clustering (unsupervised learning): Lossy compression summarizes data by keeping only relevant information and throwing away irrelevant information.
- Useful for domains without a known metric on the data space.
- Idea: instead of "guessing" at a metric, specify the relevant quantity z, and then $D_{KL}[p(z|x)\|p(z|y)]$ emerges as a measure of deviation.
- e.g. Document classification

With a known metric

- View algorithm as dynamical system
- Put knowledge about data space (metric) into the initialization of the "cluster centers"
- For the squared distance measure (euclidian distance), the dynamics of the IB algorithm lead to the "K-means" solution...
- This derivation allows us to interpret the "annealing" rate in the "soft K-means algorithm as the "magnification" we use on the problem.

Unsupervised Learning and Cluster Analysis

- Problem overview
- Roadmap to finding a solution
- Example: K-means
- Choices
- Overcoming arbitrariness
- Information theory to the help!

Different kinds of learning

- Supervised -- there is a teacher signal
- Unsupervised -- no teacher signal

- Example: cancer tissue (from UC Irvine ML repository)

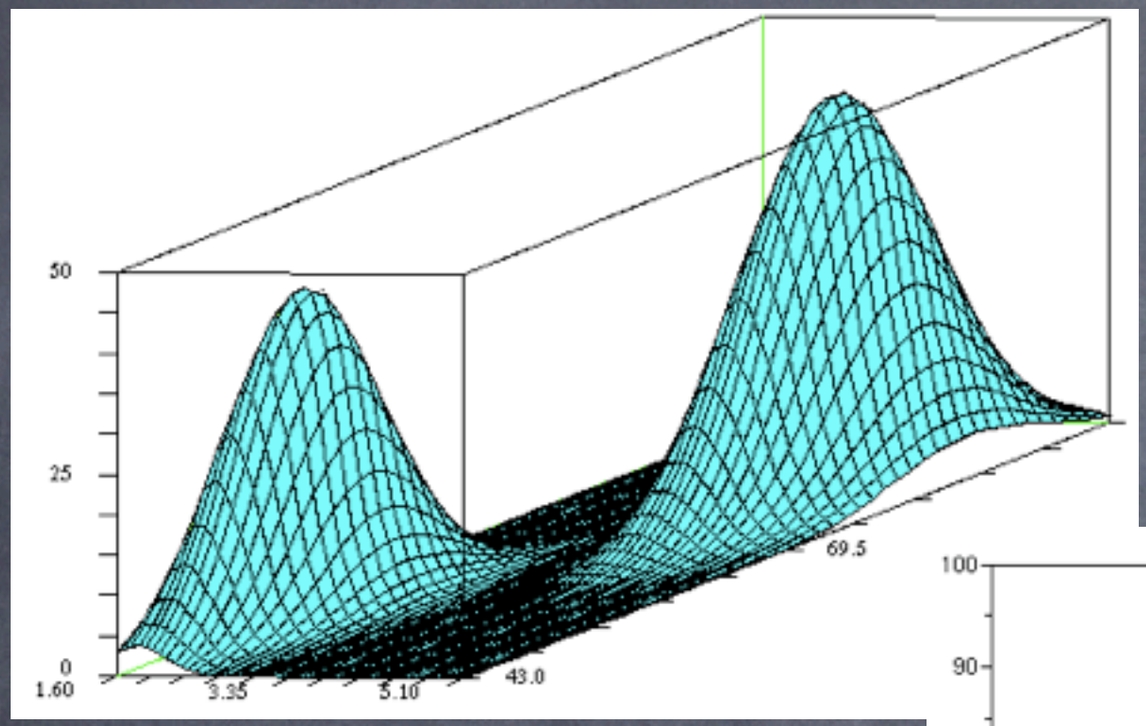
diameter	perimeter	texture	...	outcome	time	training example
13.71	20.83	90.2	...	Recurrence	77	1
13	21.82	87.5	...	Normal	119	2
12.46	24.04	83.97	...	Normal	76	3

- Binary classification: Predicting Yes/No label; here is the cancer recurrent or not.
- Regression: Predicting a continuous variable; here: the time to recurrence.

Different kinds of learning

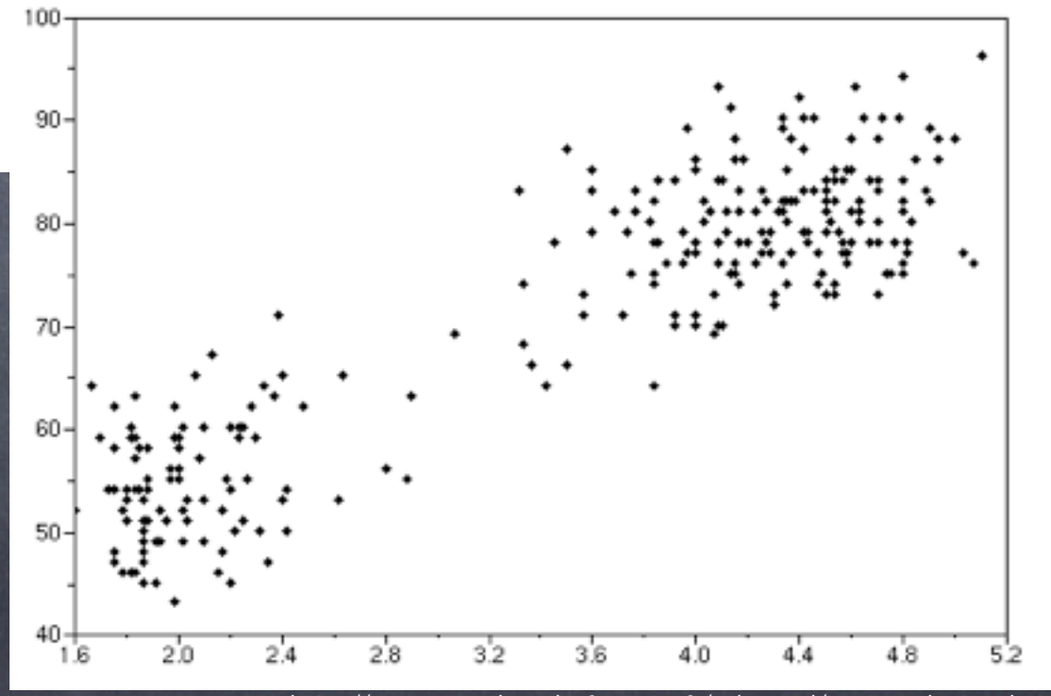
- Based on the role of the learner:
  active learning vs. passive learning
- Most animal learning is active!
- But most results in learning theory address passive learning. Complications for active learning. E. g. data are not i.i.d.
- Based on the problem definition:
  reinforcement learning: reward signal is present (a behavioral goal is specified)

Density Estimation



Process, described by probability distribution, generates data

Inference task:
Infer the underlying distribution from the data

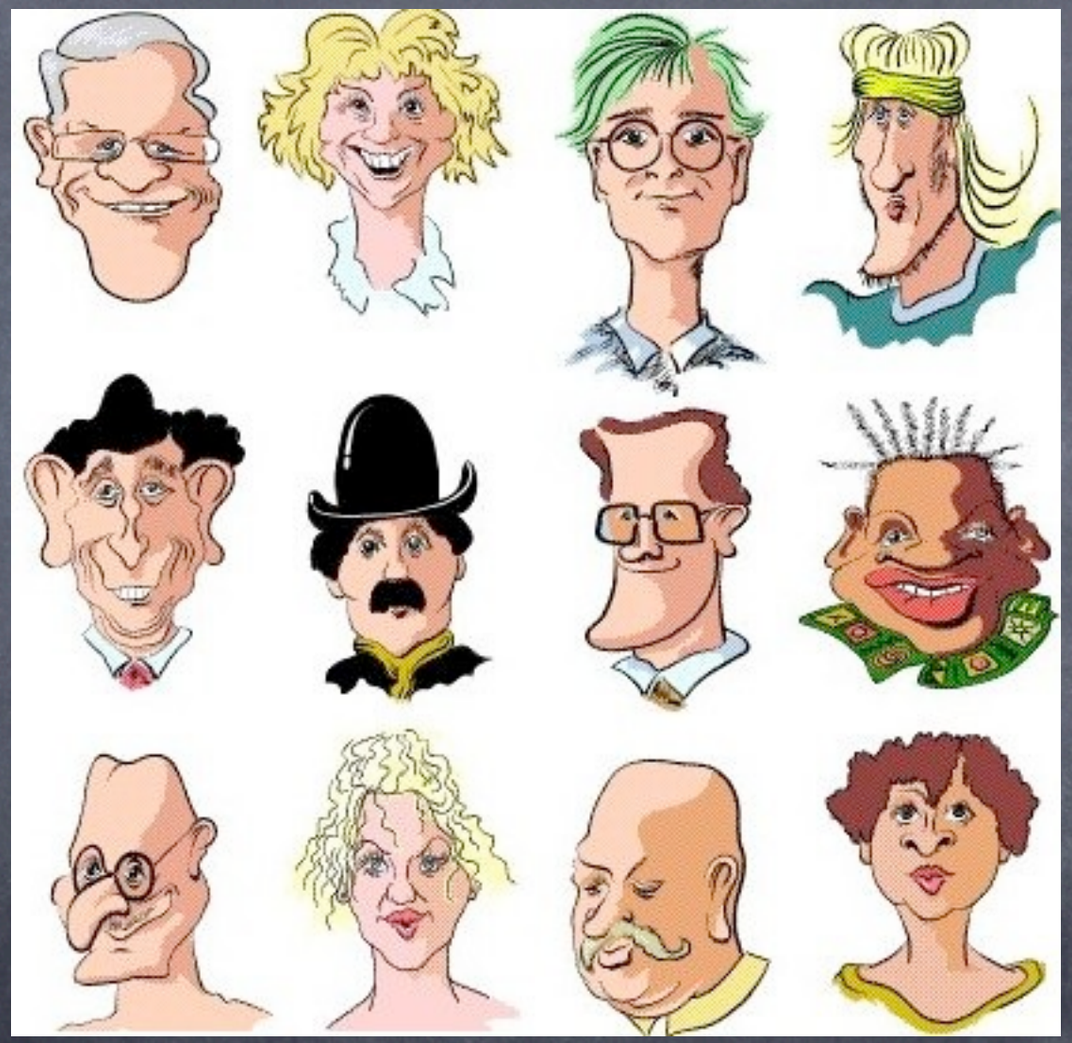


http://www-math.univ-fcomte.fr/mixmod/screenshots.php

Solving a simpler problem

- Supervised Learning - Find decision boundary.
- Unsupervised learning/clustering:
  Find high density regions of support.
- Find a partition of the input.
- "Group similar objects together, different objects into different groups"

Raw data



How do we describe the data?

Measured Features

case	sex	glasses	moustache	smile	hat
1	m	y	n	y	n
2	f	n	n	y	n
3	m	y	n	n	n
4	m	n	n	n	n
5	m	n	n	y?	n
6	m	n	y	n	y
7	m	y	n	y	n
8	m	n	n	y	n
9	m	y	y	y	n
10	f	n	n	n	n
11	m	n	y	n	n
12	f	n	n	n	n

http://149.170.199.144/multivar/ca.htm

What does it mean for two faces to be similar?

Roadmap to Clustering

- Similarity measure
- Clustering criterion (objective function)

→ Algorithm (finds optimal K partition)

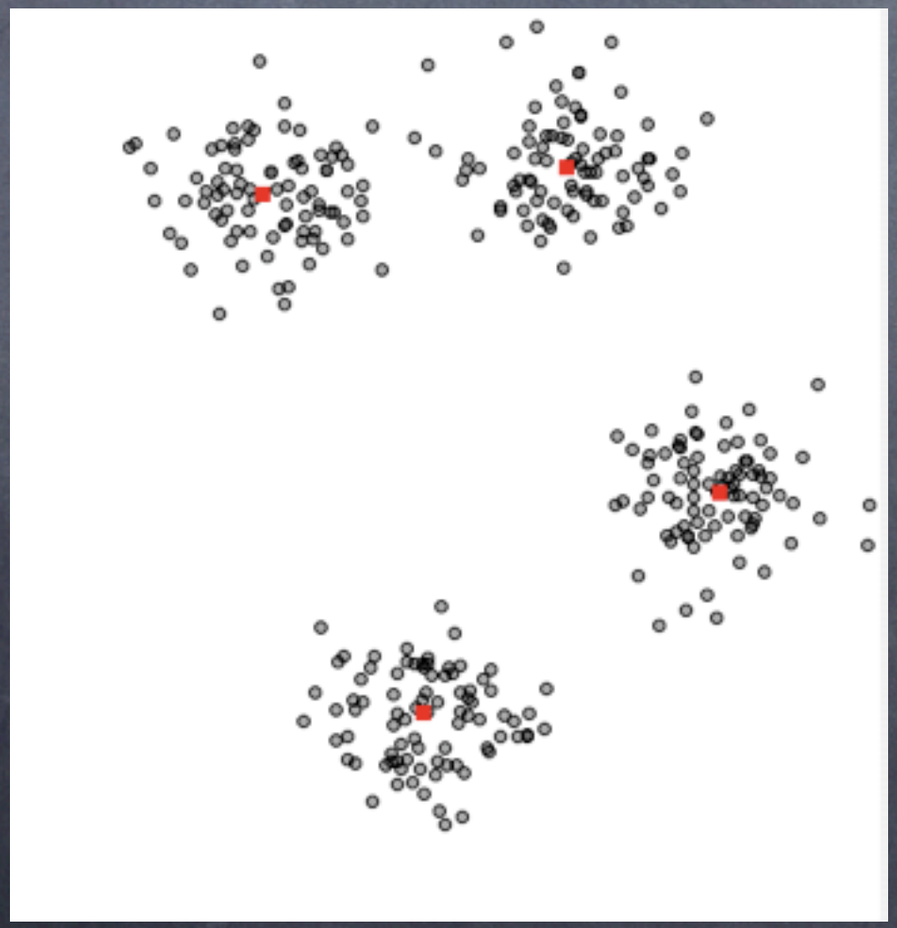
- Number of clusters K

Need to specify these.
Problem: The clustering solution depends on the choices we make.
Theoretical guidance?
Underlying principles?

Example: K-means clustering

(MacQueen, 1967)

The data, visualized in 2-dim input space:

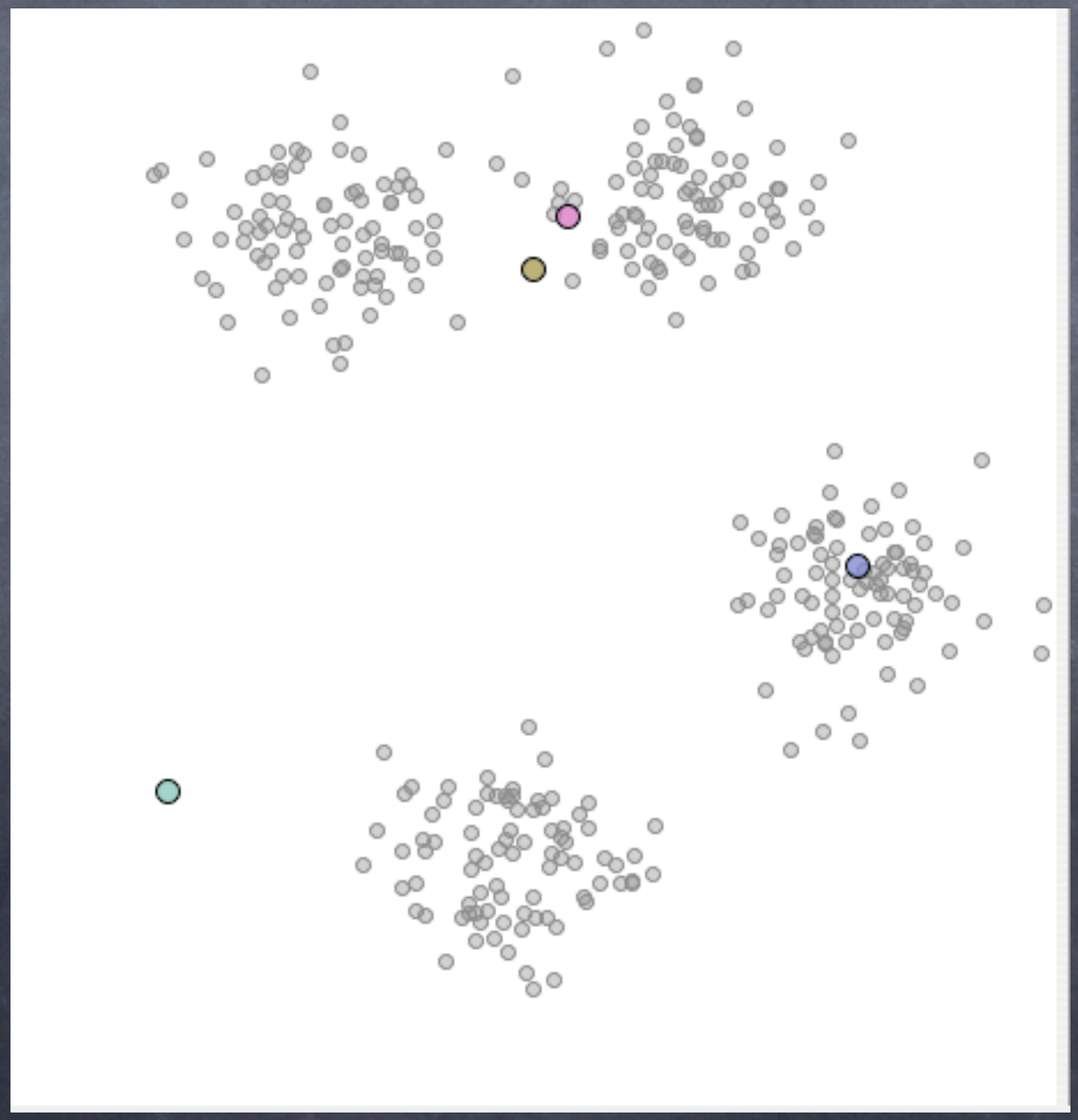


The red squares indicate the "true" cluster centers.

Algorithm doesn't know those.

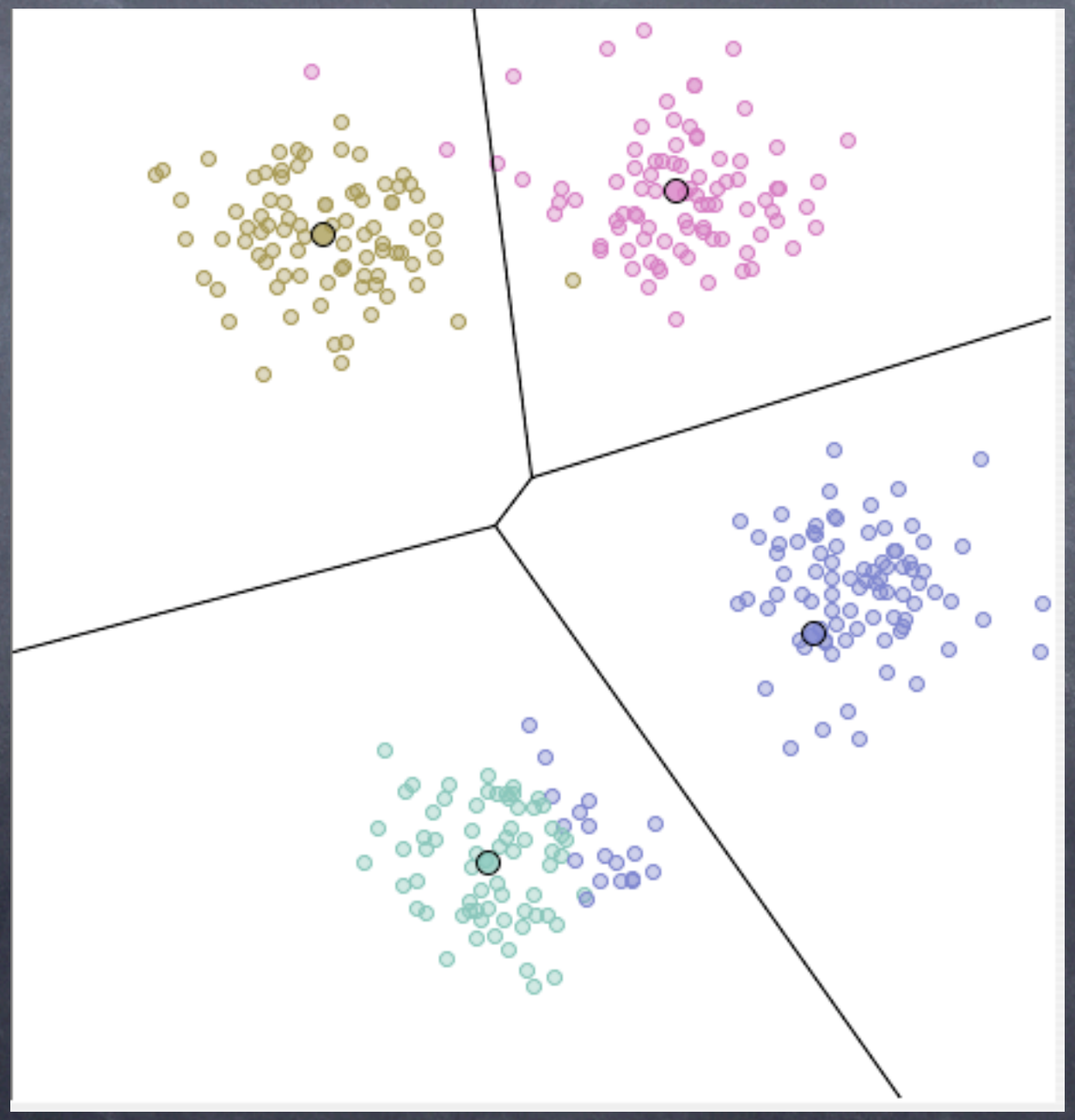
The algorithm

1. Initialize K cluster centers. Here K=4.

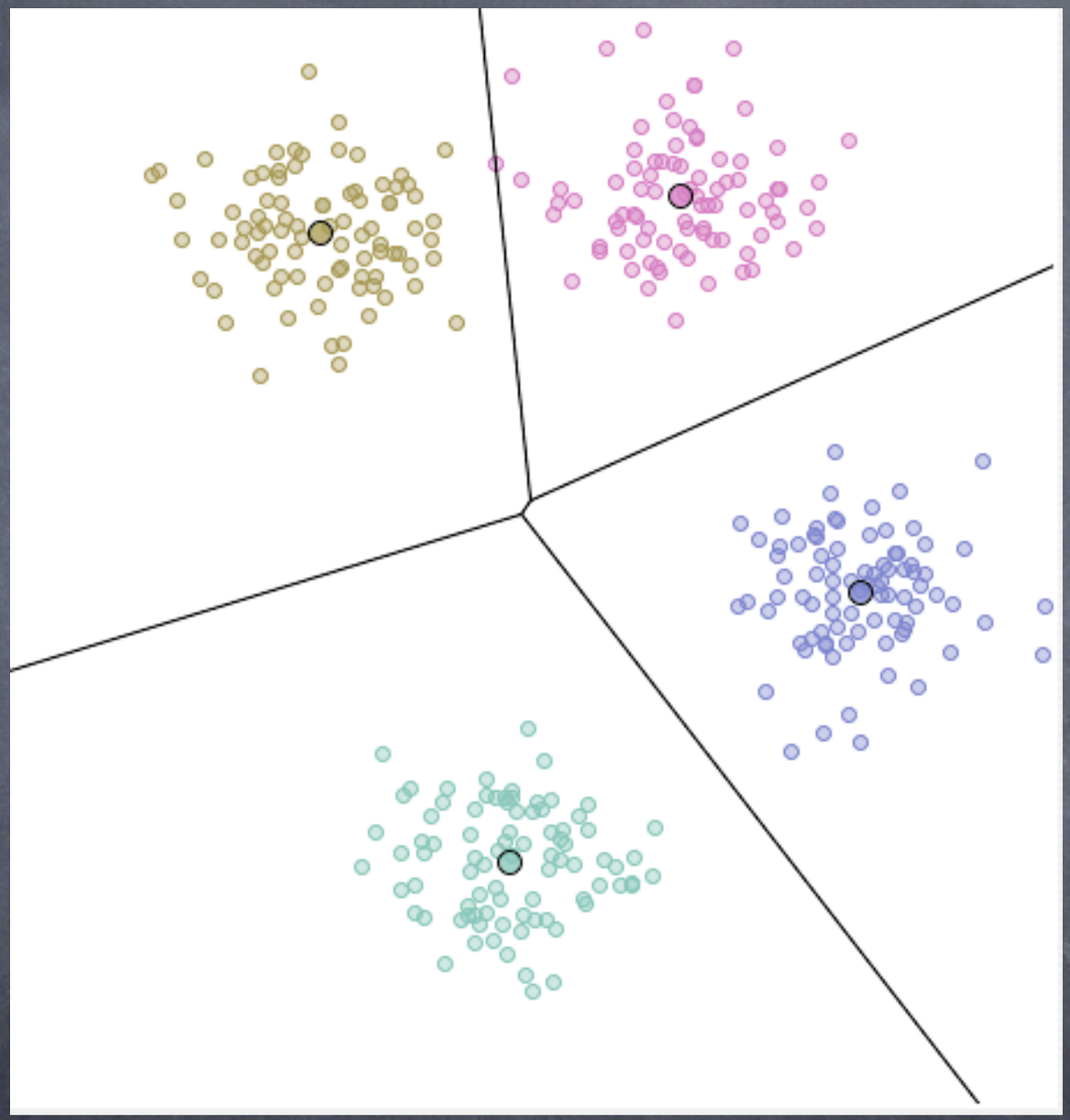


2. Assign each point to nearest cluster center

3. Re-compute cluster centers as the means of the points assigned to each cluster



- Repeat until nothing changes anymore



- Fast convergence! (here after only 2 steps)

Objective function

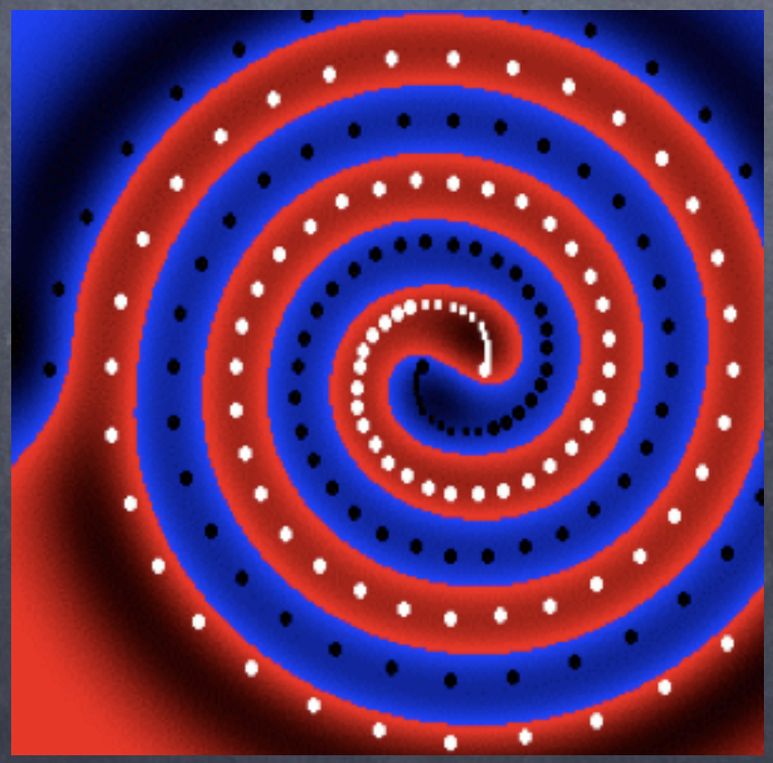
- Can show that K-means algorithm minimizes the squared error function (c = 1, .., K cluster centers; x: N data points)

$$\sum_{j=1}^{K} \sum_{i=1}^{N} \|x_i^{(j)} - c_j\|^2$$

- Other objective functions could be used!
- It is not always obvious which one to choose.

Similarity measure

- Squared distance (euclid norm) $\|x_i^{(j)} - c_j\|^2$
- Not good for data distributions with non-euclidean structure. Example:



http://www.kernel-methods.net

Other similarity measures

- Metric, i.e. dis-similarity (j: dimension; l,m: object index)
  - ✳ Minkowski $d_{lm} = (\sum_{j=1}^{D} w_j^\lambda |x_l^j - x_m^j|^\lambda)^{\frac{1}{\lambda}}$
  - ✳ City block $d_{lm} = \sum_{j=1}^{D} w_j |x_l^j - x_m^j|$
- Similarity
  - ✳ Correlation coefficient; $\bar{x}_l = \sum_{j=1}^{D} x_l^j / D$

$$s_{lm} = \frac{\sum_{j=1}^{D}(x_l^j - \bar{x}_l)(x_m^j - \bar{x}_m)}{(\sum_{j=1}^{D}(x_l^j - \bar{x}_l)^2 \sum_{j=1}^{D}(x_m^j - \bar{x}_m)^2)^{1/2}}$$

Measures cosine of angle between two vectors, originating at the mean of the data.

- and many more...

Constructing objective functions

1. Measures of heterogeneity and isolation for each cluster:

- Averages (sum of squares as before; average L1 norm |x-c|)
- Extrema (Diameter: measures the dissimilarity between the most dissimilar points in a cluster; Split: smallest dissimilarity between an object in a class and an object outside the class)

2. Combine those by {sum, min, max} -> Clustering Criterion (objective function, OF)

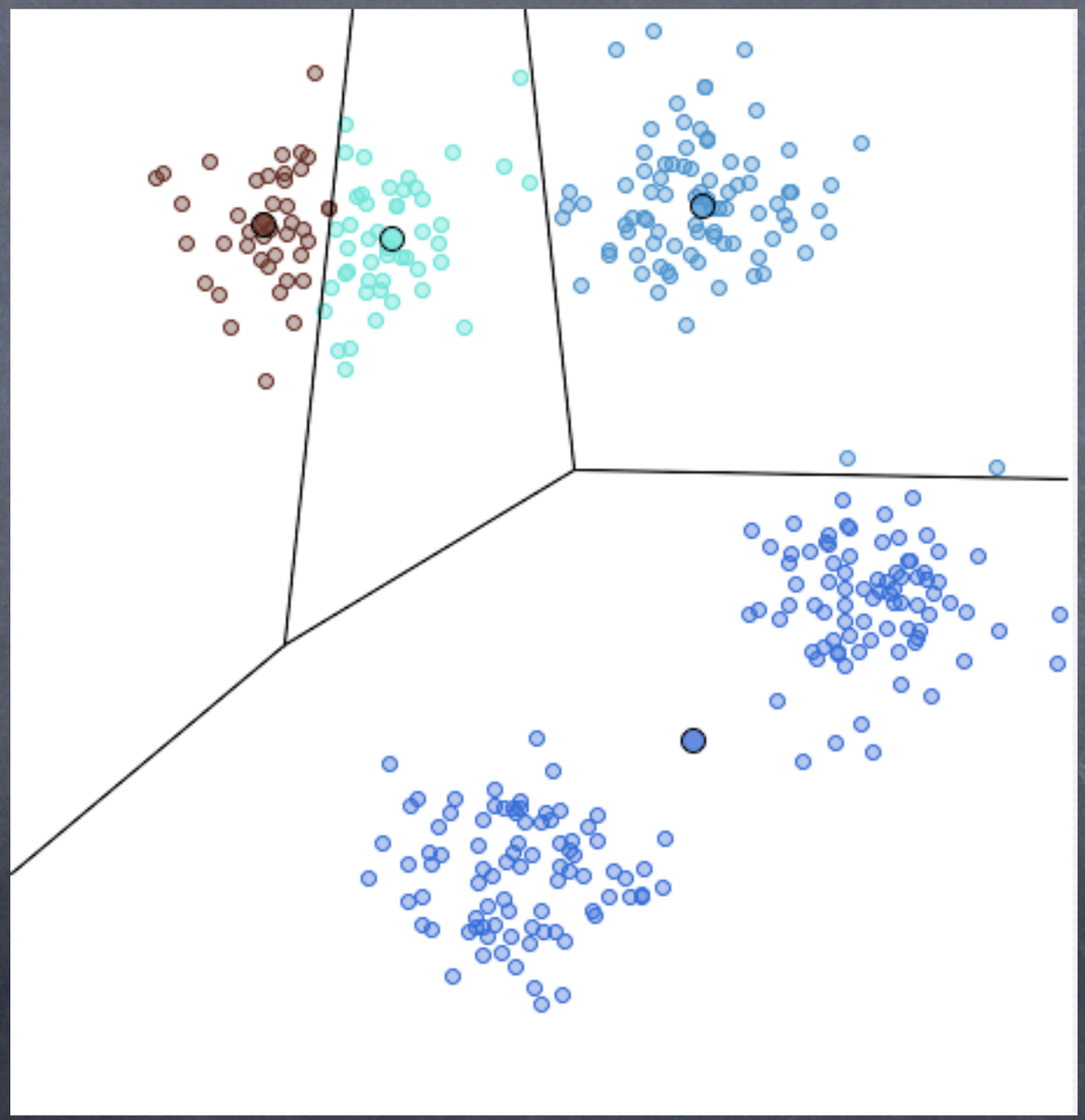
3. Optimize objective function (min or max)

- lots of freedom to choose; often no principle behind choice, just "common sense" -> hand craft OF for data set.
  Problem: Results of the analysis depend on OF.

Clustering algorithms

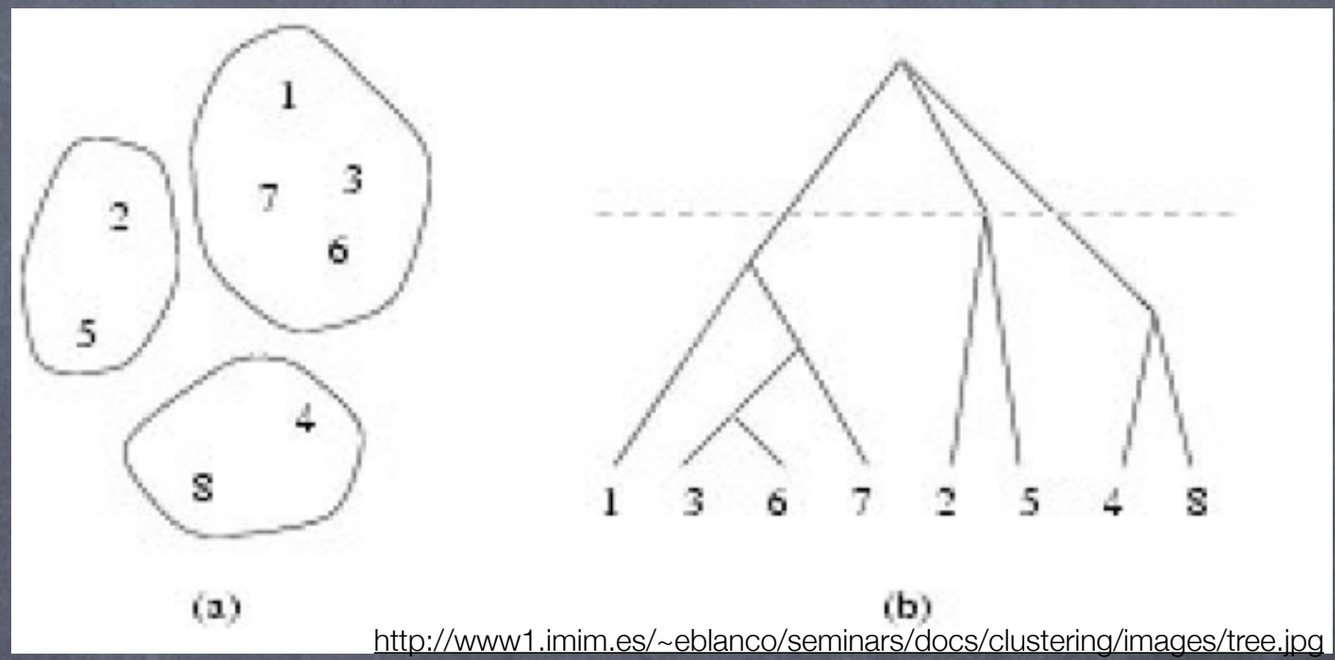
- Iterative reallocation (e.g. K-means)
- Other option: hierarchical clustering algorithms
- Problem with K-means: Local minima

- A different initialization leads to a sub-optimal solution.



Hierarchical Clustering

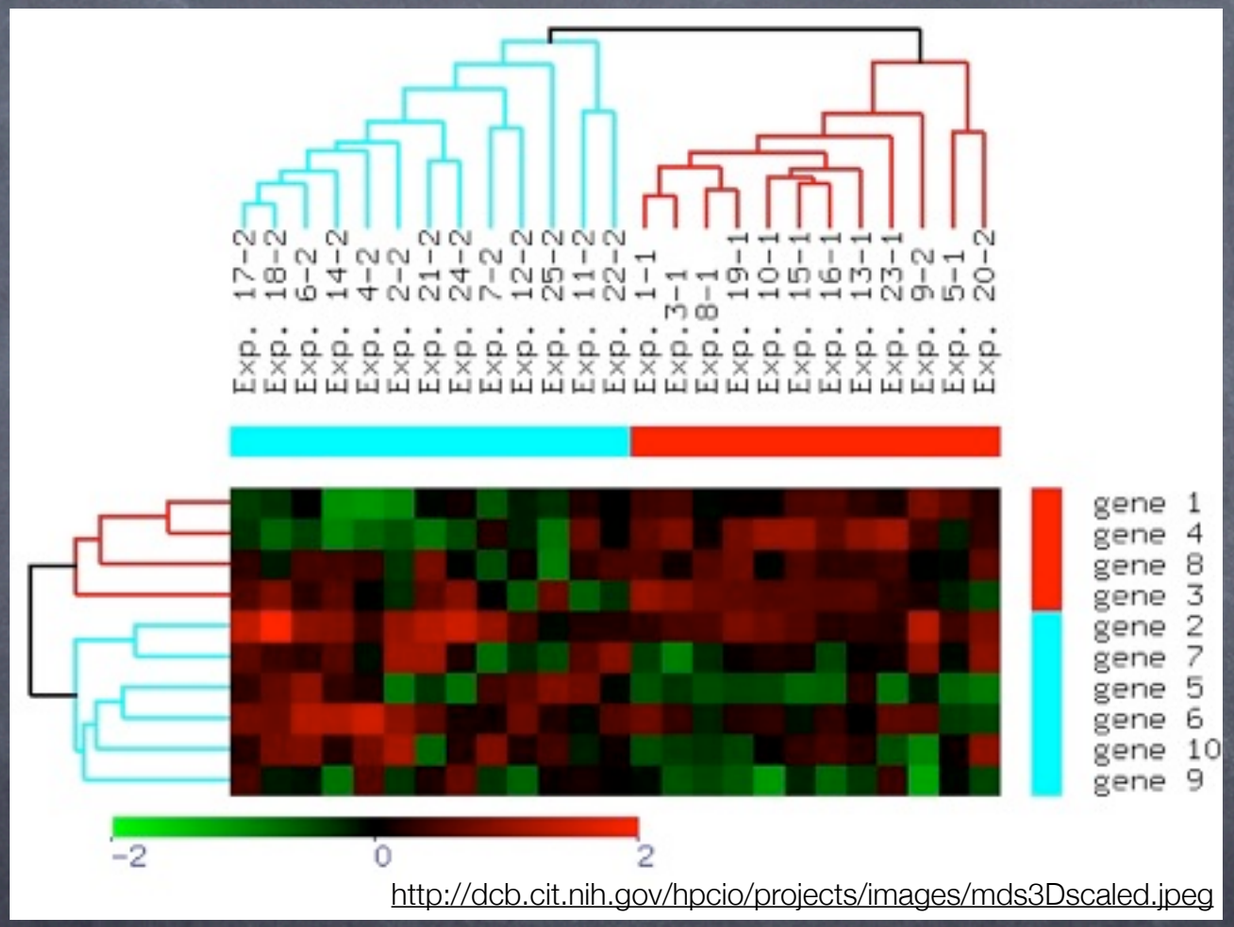
- Build a hierarchy of similarities (i.e. a rooted tree structure)



http://www1.imim.es/~eblanco/seminars/docs/clustering/images/tree.jpg

- Algorithms:
- Direct optimization, transform similarity into height on the tree (can be NP-hard).
- Agglomerative algorithms, greedy amalgamation.

- Hierarchical Clustering makes sense when we expect a hierarchy in the data (e.g. taxonomy)
- Careful when we can not assume a priori that a hierarchical structure is present in the data.



http://dcb.cit.nih.gov/hpcio/projects/images/mds3Dscaled.jpeg

- Instead we could use iterative reallocation algorithm for different numbers of classes, and observe if a hierarchy emerges...

Number of clusters

- Has to be set a priori in K-means
- The solution (optimal partition) depends on K!
- In general, we don't know K.
- How do we determine K?

- We can always do "better" (in terms of minimizing average distances) with more K
- However: in the extreme K = N, we are not summarizing the data anymore.
- What level of detail is appropriate?
- Model complexity control!

Number of clusters

- Many attempts in the statistical literature to find a "goodness" criterion for the best K-cluster fit. Use this criterion to determine K.

  Examples:

- Compare within-cluster sum of squared distances if cluster is split or not (Duda, Hart 1973).
- GAP-statistic (Tibshirani et, 2000): compare change in within cluster dispersion to that expected under a uniform null distribution.
- Problem: As before, arbitrary ad hoc measures.

Can we overcome the arbitrariness?

Problems:

- Similarity measure
- Clustering Criterion
- Complexity Control

Different approaches:

- Statistical modeling (use bayesian estimation); makes assumptions about distributions
- Stability arguments
- Information theoretic approach

Mixture models

- Assume that the underlying probability is a mixture of K specified probability functions.
- Those functions may be parameterized
- Most often Gaussians are used. Parameters: Number of gaussians (K); means; and covariances
- Find the most likely parameters (Bayesian estimation)
- Nice: the number of clusters becomes a parameter -> no need to introduce extra statistic
- Problem: the mixture model may not fit the data well (or: did we choose the right hypothesis class?)

Uses the EM algorithm

- Expectation-maximization. Iterative algorithm:
- Initialize the parameters

1. (E-step): Compute expectation values for the membership variables of each data point, given the current parameters
2. (M-step): Recompute the parameters from the membership

- Repeat until convergence

Stability arguments

- Idea: Robustness of the partition is important!
- Any reasonable data clustering has to be stable under sample fluctuations.
- See for example (on reading list) Buhmann et al. (2000) Ben David (2004, 2005)
- Similar idea also used in information theoretic context (e.g. Still and Bialek, 2004).

Information theoretic clustering

- Clustering = lossy compression (idea: communicate more efficiently)
- Minimal assumptions and no arbitrary ad hoc definitions of various statistics.
- Have to specify what is relevant to the analysis (with respect to what do we want to compress)
- Everything else follows from this decision and information theoretic principles: Objective; similarity measure; algorithm.
- Complexity control via stability arguments

Homework

- Implement and/or test K-means.
- Make artificial data, e.g. drawn from m Gaussians.
- Analyze the performance. Some measures: % correct (global optimimum); # iterations till convergence.
- Play with changing the input data. Make the task more or less difficult. Example: increasing class overlapp -> more difficult.

Break

Learning and lossy data compression

- When we build a model of a data set, we map observations to a representation that summarizes the data in an efficient way.
- Example: K-means. Map N data points to K clusters, with centroids c. If K << N, then we get a substantial compression! log(K) << log(N).

Lossy Compression

- Idea: Summarize data by keeping only relevant information and throwing away irrelevant information.

- Needs:
  - a measure for information -> Shannon
  - a notion of relevance

Relevance historically

- Shannon (1948): Define a function that measures the distortion between the original signal and its compressed representation.
- Note that this is related to the similarity measure in unsupervised learning/cluster analysis.
- Distortion function is a choice: which function to use is up to the experimenter.
- It is not always obvious what function should be used, especially if the data do not live in a metric space, and so there is no "natural" measure. Example: Speech.
- Tishby, Pereira, Bialek (1999): measure relevance directly via mutual information about variable of interest (relevant quantity, which we have to know)
- Then the appropriate "distortion" or "similarity" measure arises naturally.
- But we have to know about the correlations between data and relevant quantity.

How "efficient" is the representation?

- Entropy can be used to measure the compactness of the model. Sometimes called "statistical complexity" (J. P. Crutchfield.)
- This is a good measure of complexity if we are searching for a deterministic map (in clustering called a "hard partition")
- In general, we may search over probabilistic assignments (clustering: "soft partition")

Trade-off between compression and accuracy

- Think about a continuous variable.
- To describe it exactly, you need infinitely many bits.
- But finite bits to describe it up to some accuracy.

Rate distortion theory used for clustering

- Find assignments $p(c|x)$ from data $x \in X$ to clusters $c = 1, ..., K$, and find class-representatives x_c ("cluster centers"), such that the average distortion $D = \langle d(x, x_c) \rangle$ is small.

- Compress the data by summarizing it efficiently: Minimize the coding rate, the information which the clusters retain about the raw data (least effort).

$$I = \left\langle \frac{p(x,c)}{p(x)p(c)} \right\rangle$$

Constrained optimization

$$\min_{\substack{p(c|x) \\ x_c}} \left[I(x,c) + \beta \langle d(x, x_c) \rangle \right]$$

Solution: $$p(c|x) = \frac{p(c)}{Z(x,\beta)} \exp\left[-\beta d(x, x_c)\right]$$

x_c from $$\langle \frac{d}{dx_c} d(x, x_c) \rangle_{p(x|c)} = 0$$

(centroid condition)

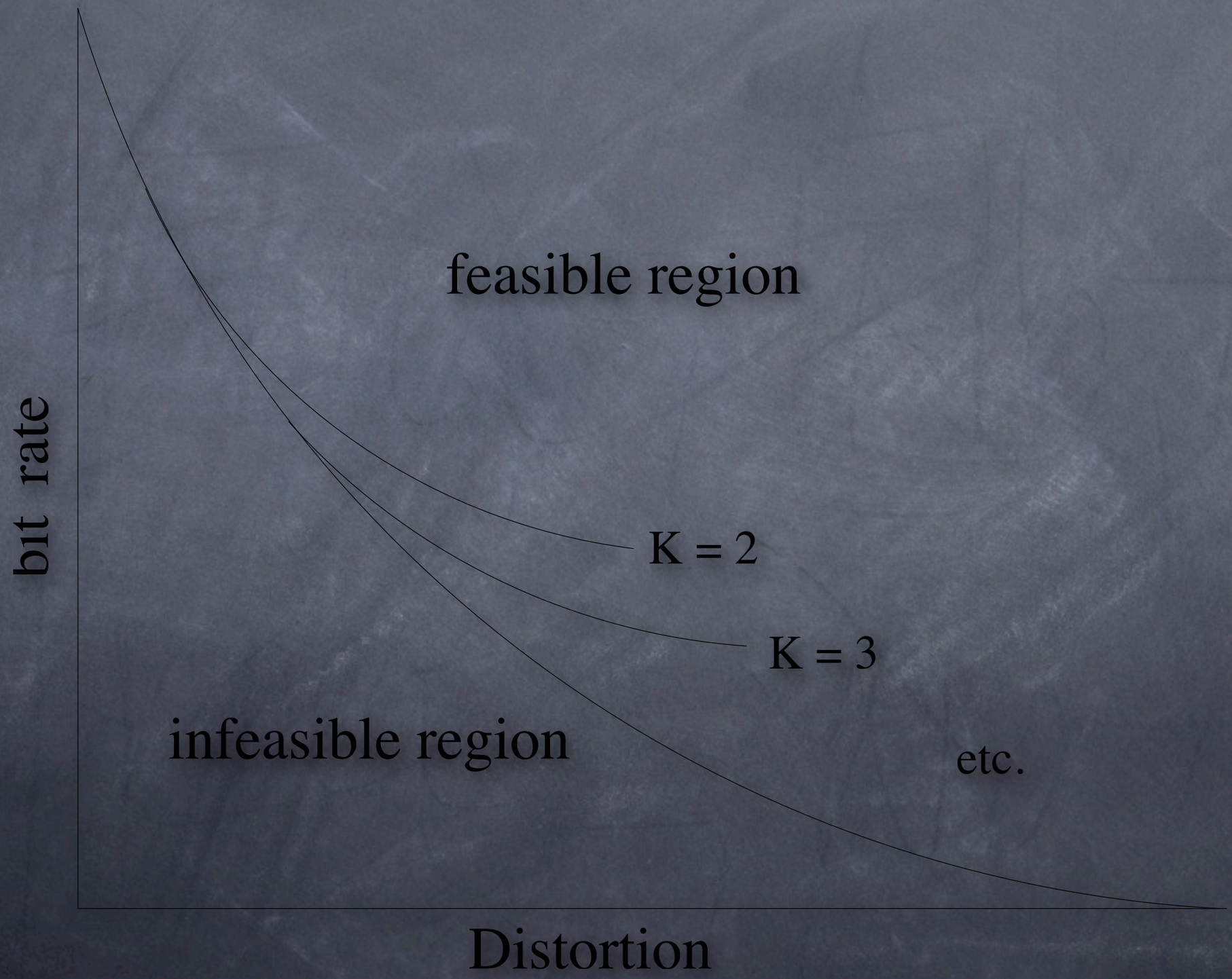
Rate distortion curve

- Family of optimal solutions, one for each value of the Lagrange multiplier beta. This parameter controls the trade-off between compression and fidelity.

$$p(c|x) = \frac{p(c)}{Z(x,\beta)} \exp\left[-\beta d(x, x_c)\right]$$

- Evaluate objective function at the optimum for each value of beta and plot I vs D.
  => Rate-distortion curve.

Rate-distortion curve



Soft K-means

- for squared error distortion, $d = (x - x_c)^2/2$, the centroid condition reduces to

$$x_c = \langle x \rangle_{p(x|c)}$$

- This is called "soft K-means" because assignments can be probabilistic (fuzzy):

$$p(c|x) = \frac{p(c)}{Z(x,\beta)} \exp\left[-\beta d(x, x_c)\right]$$

Deterministic Annealing

- in the limit, $\beta \to \infty$, we have deterministic (or "hard") assignments because

$$c^* := \arg\min_c d(x, x_c);\ D(x, c) := d(x, x_c) - d(x, x_{c^*}) > 0$$

$$p(c^*|x) = \frac{\exp(-\beta d(x, x_{c^*}))}{\sum_c \exp(-\beta d(x, x_c))} = \left(1 + \sum_{c \neq c^*} \exp(-\beta D(x, c))\right)^{-1} \to 1$$

- the analogy to thermodynamics inspired Deterministic Annealing (K. Rose, 1990)

Algorithm

- Choose a distortion measure.
- Fix the "temperature", T, to a very large value (corresponds to small $\beta = 1/T$)
- Solve iteratively, until convergence:

Assignments: $p(c|x) = \frac{p(c)}{Z(x,\beta)} \exp\left[-\beta d(x, x_c)\right]$

Centroids: $\left\langle \frac{d}{dx_c} d(x, x_c) \right\rangle_{p(x|c)} = 0$

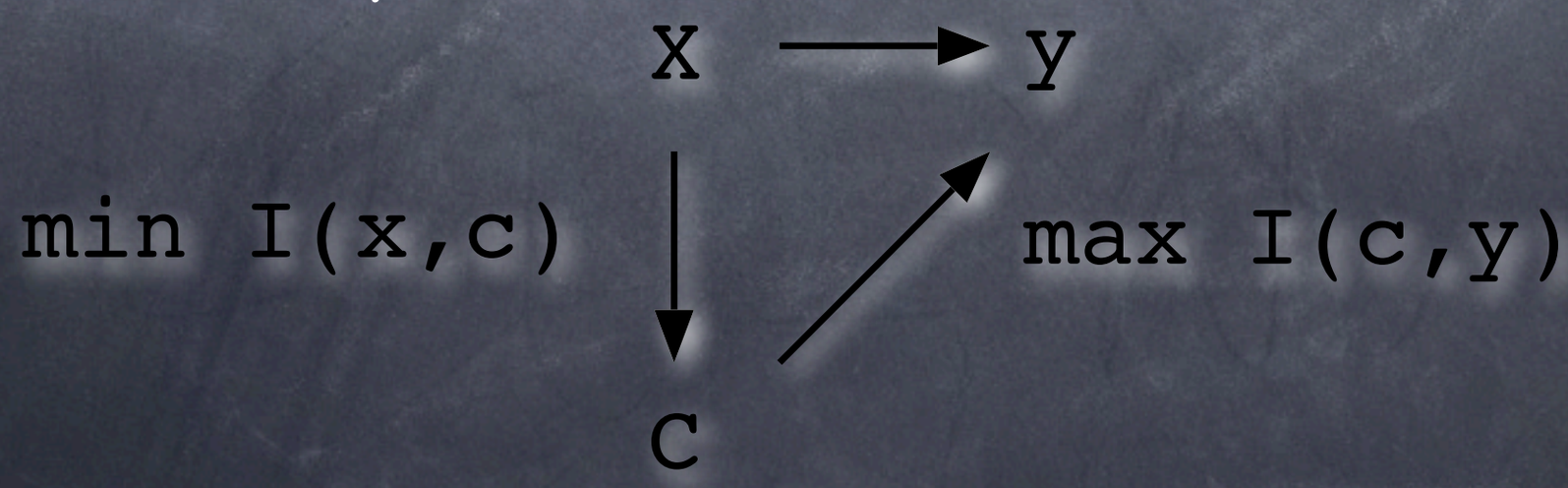
- Lower temperature: T <- aT, and repeat.
- a = "annealing rate", a small, positive number.

Choice of distortion function can be difficult

- Example: Speech
- Cluster speech such that signals which encode the same word will group together
- May be extremely difficult to define a distortion function which achieves this
- Intelligibility criterion?
- Should measure how well the meaning is preserved.

Information Bottleneck Method

- Tishby, Pereira, Bialek, 1999
- Instead of guessing a distortion function, define **relevant information** as information the data carries about a quantity of interest (Example: phonemes or words.)
- Data is compressed such that relevant information is kept maximally.



Constrained optimization

$$\max_{p(c|x)} \left[I(y,c) - \lambda I(x,c) \right]$$

- Optimal assignment rule

?

Constrained optimization

$$\max_{p(c|x)} \left[I(y,c) - \lambda I(x,c) \right]$$

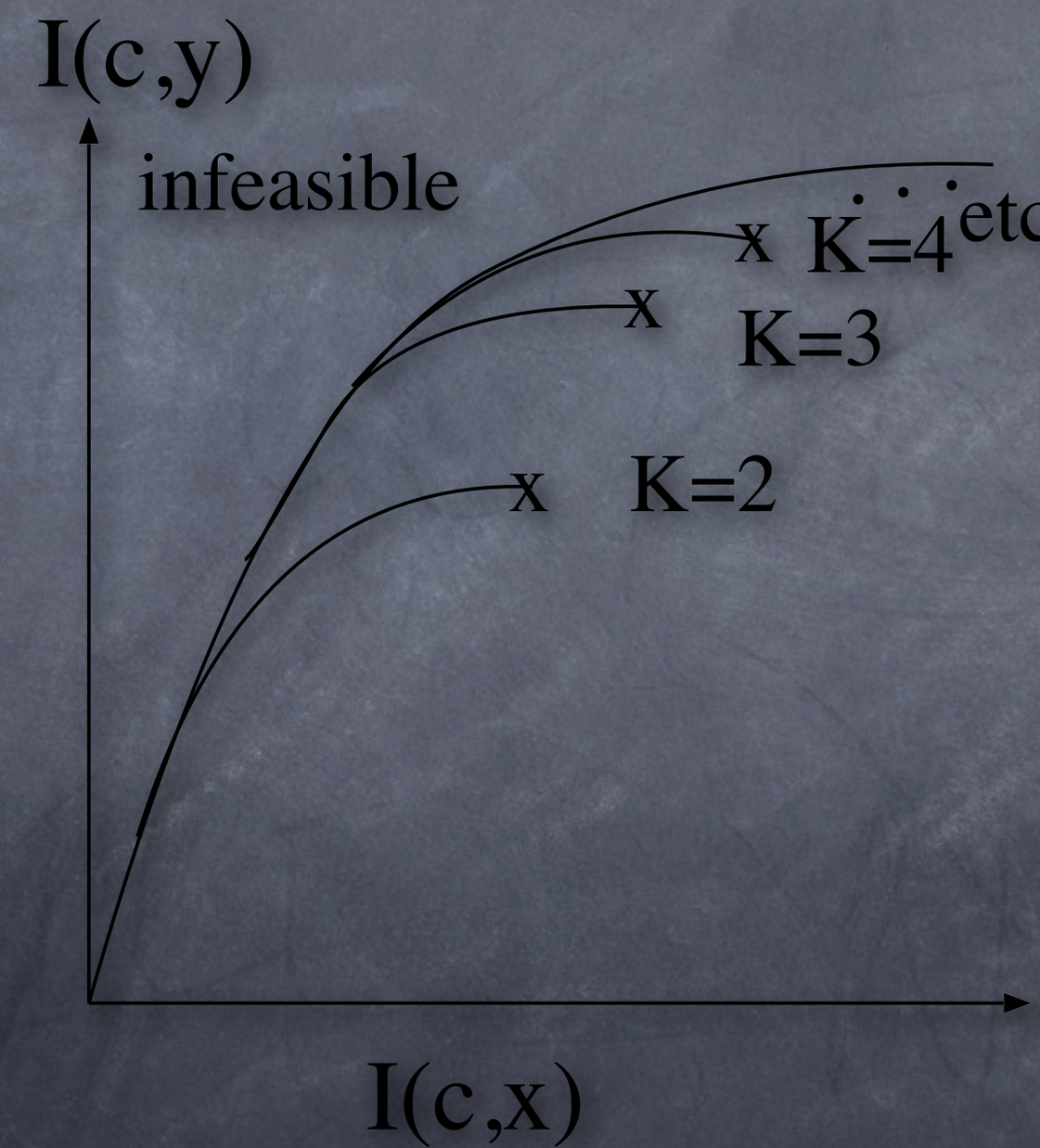
- Optimal assignment rule

$$p(c|x) = \frac{p(c)}{Z(x,\lambda)} \exp\left(-\frac{1}{\lambda} D_{KL}[p(y|x) \| p(y|c)] \right)$$

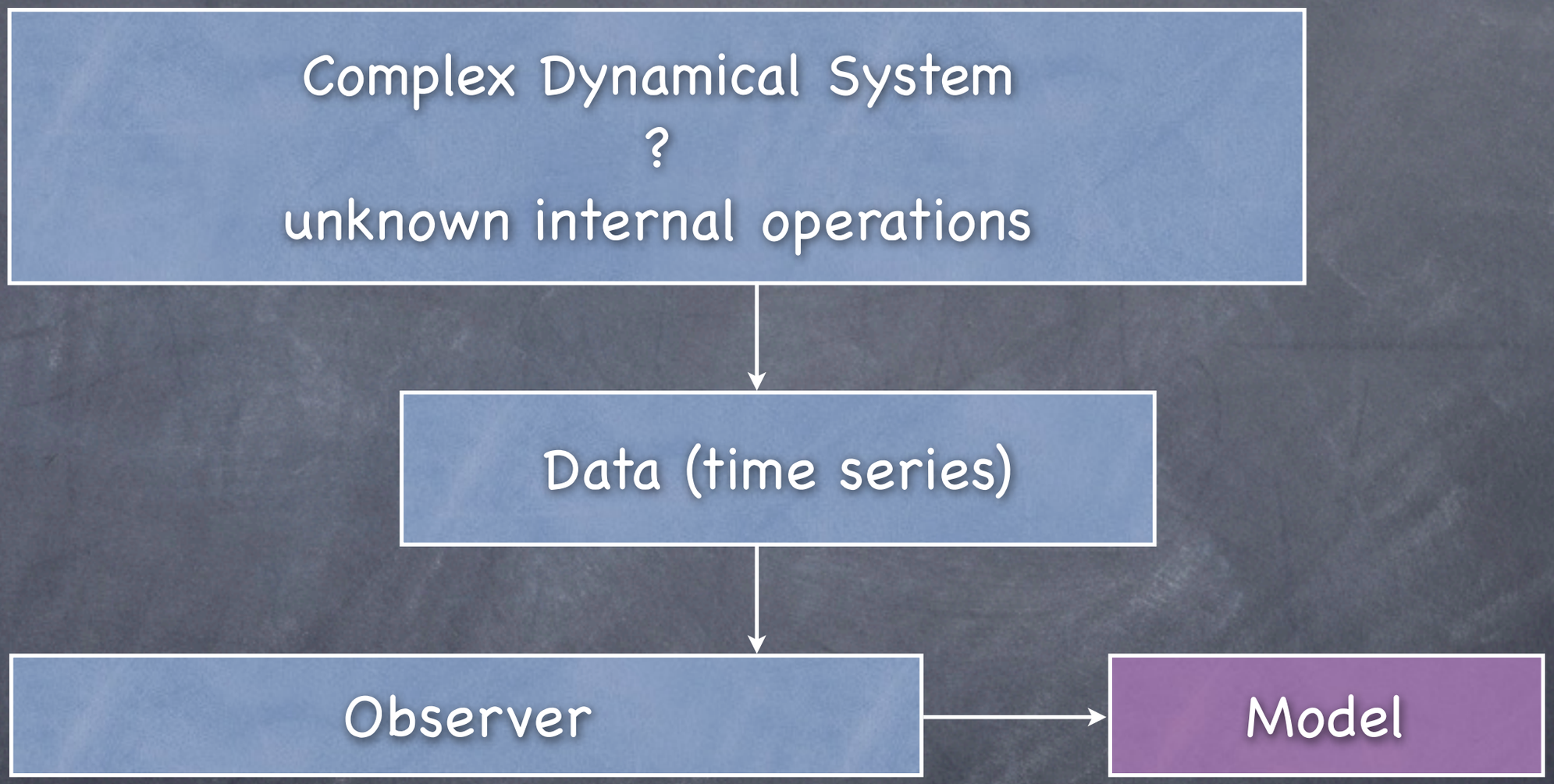
- Kullback-Leibler divergence emerges as the distortion function:

$$D_{KL}[p(y|x) \| p(y|c)] = \sum_y p(y|x) \log_2 \left[\frac{p(y|x)}{p(y|c)} \right]$$

Information Plane

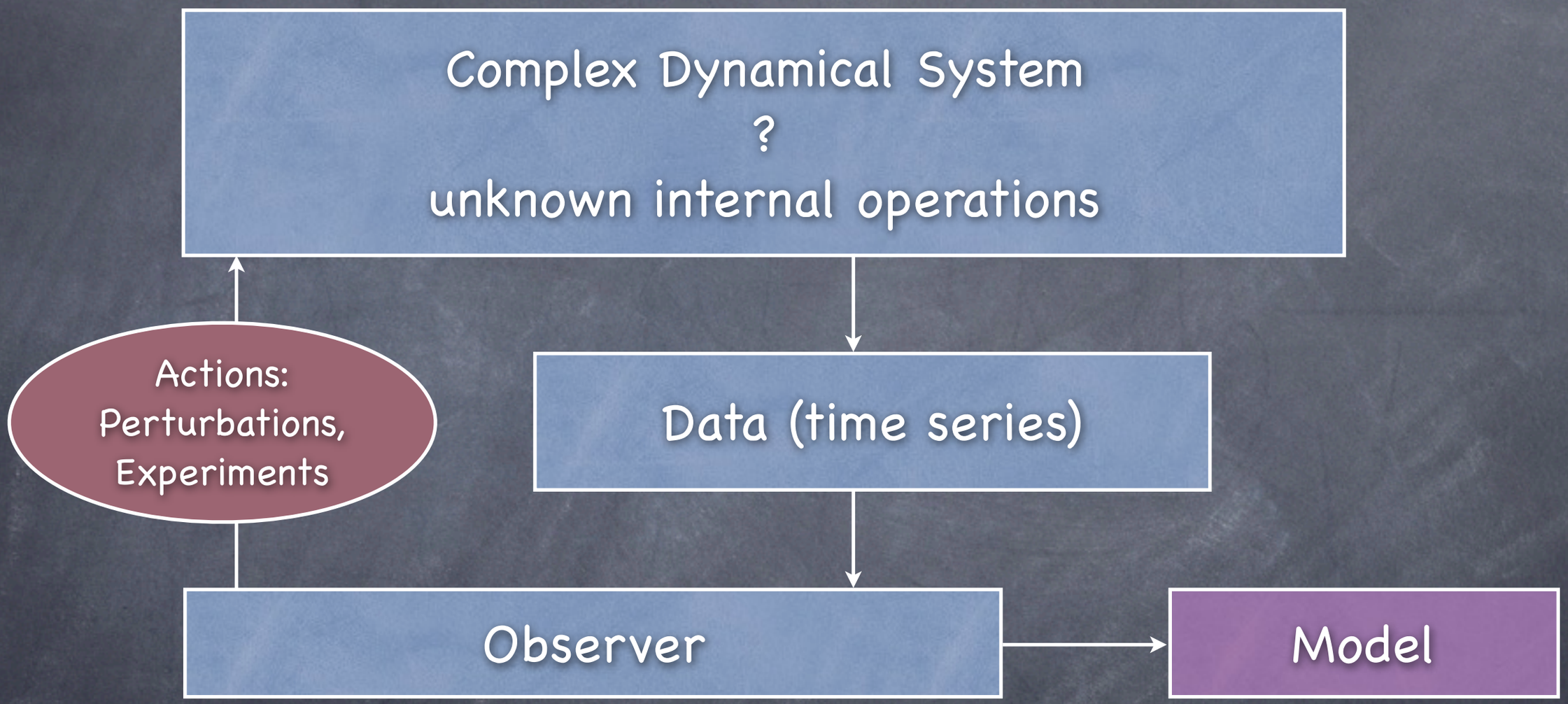


Apply to time series data



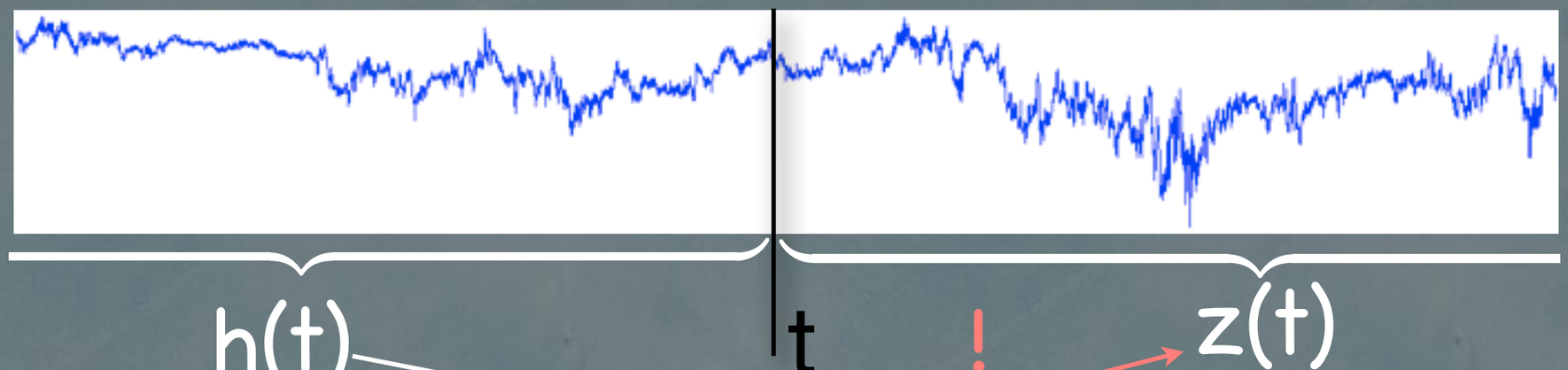
What is a good model?

More general motivation



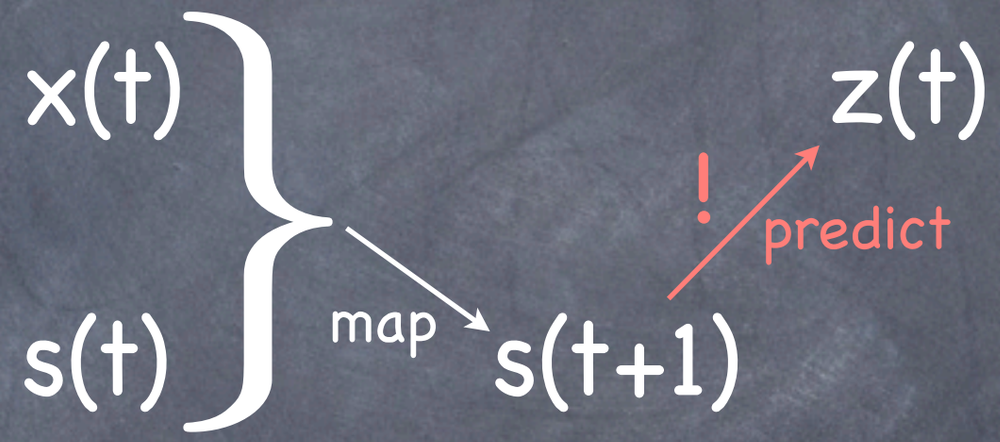
1. What is a good model?
2. What action strategies allow one to extract a good model?

1a. Signal: x(t)



find: $p(s|h)$

1b. Dynamics

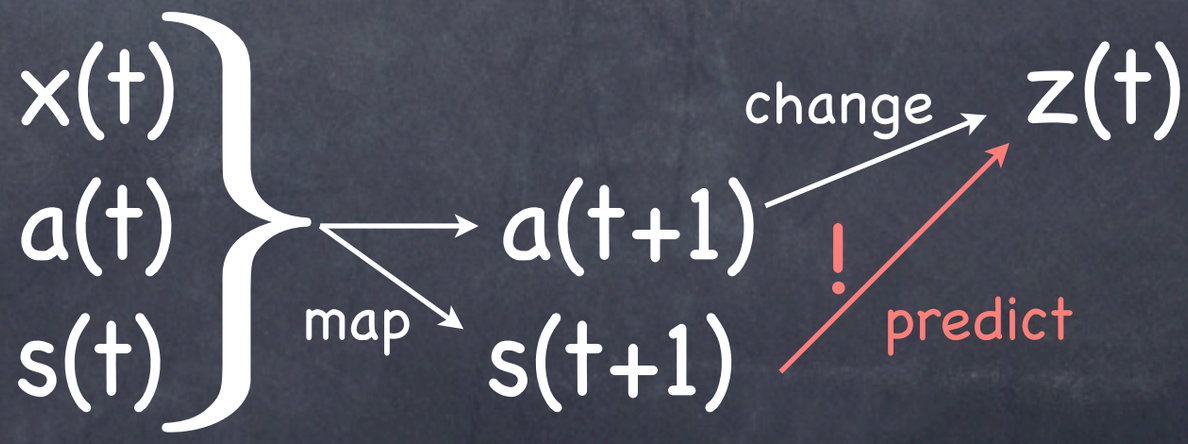


find: $p(s_{t+1}|x_t, s_t)$

---

2. include feedback

find: $p(s_{t+1}|x_t, s_t, a_t)$
$p(a_{t+1}|x_t, s_t, a_t)$



What is a good model?

- Model = a representation (summary, s) of the data we have seen (history, h).
- A good model predicts well.
  -> find maximally predictive model.
  Keep predictive information!
- A good model is compact.
  -> do not keep irrelevant information.

$$\max_{p(s_t|\overleftarrow{x}_t)} \left(I[s_t; \overrightarrow{x}_t] - \lambda I[s_t; \overleftarrow{x}_t] \right)$$

Theorem: In the regime $\lambda \to 0$ the causal state partition is found.

(S. Still, J. P. Crutchfield, C. Ellison. Optimal Causal Inference. arXiv:0708.1580)

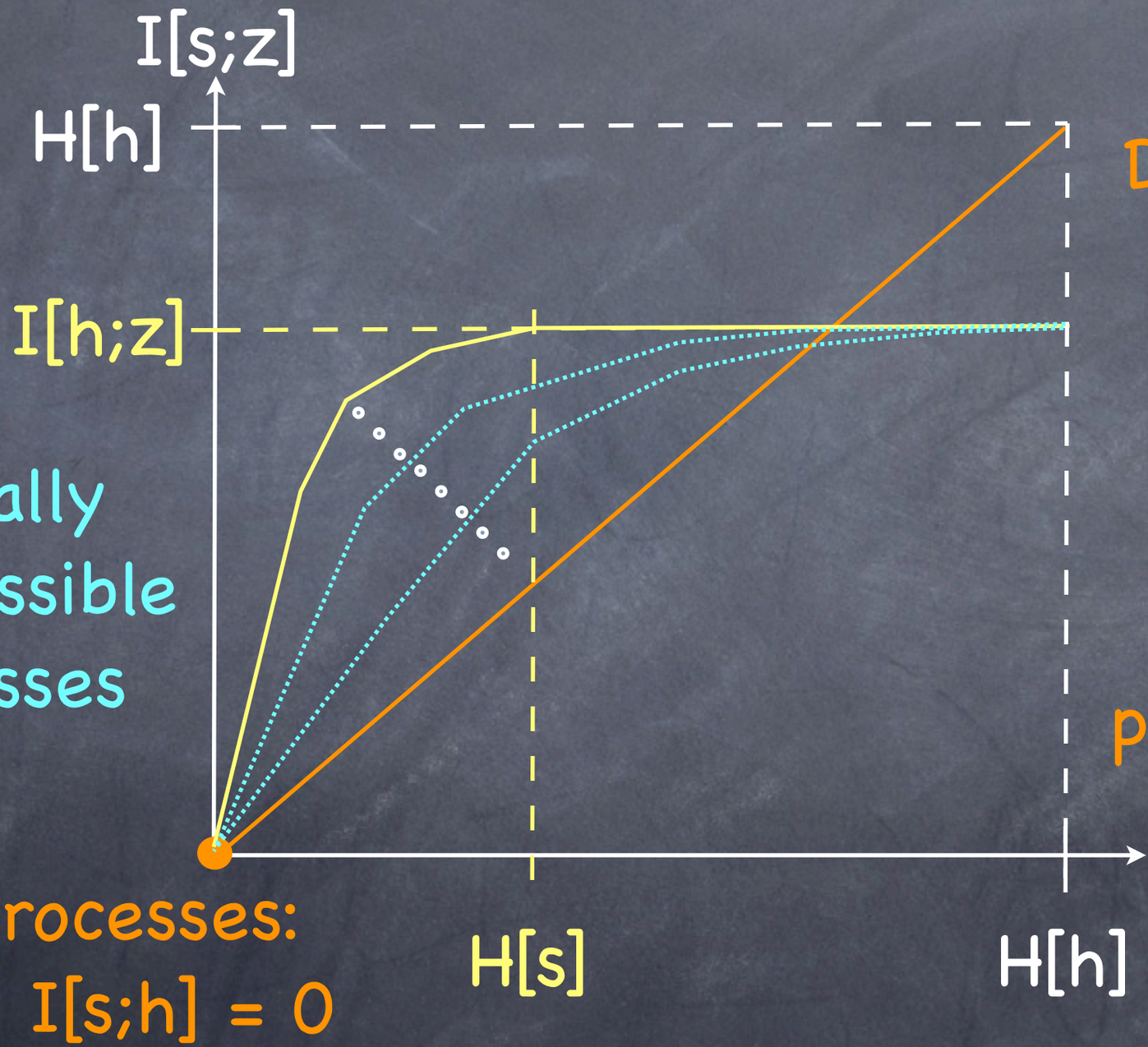
- Definition of causal states (= c(h)). Equivalence relation: p(z|h) = p(z|h′) => c(h) = c(h′). Note: deterministic map!

- The causal states reflect the underlying causal structure of the system.

  (J. P. Crutchfield and K. Young (1989) PRL 63:105–108)

- Causal states are unique and minimal sufficient statistics.
  (J. P. Crutchfield and C. R. Shalizi (1999) Phys.Rev.E 59(1):275–283)

Causal compressibility



Causally compressible processes

i.i.d. processes:
I[s;z] = I[s;h] = 0

Deterministic and time reversible processes:
I[s;z] = I[s;h]
I[h;z] = H[h]
Example: periodic processes

S. Still, J. P. Crutchfield. Structure or Noise? http://lanl.arxiv.org/abs/0708.0654

- Not causally compressible:
  - Deterministic and time reversible processes (analytical result: RD curve lies on the diagonal)
  - i.i.d. processes (analytical result: RD curve is degenerated - point at the origin)

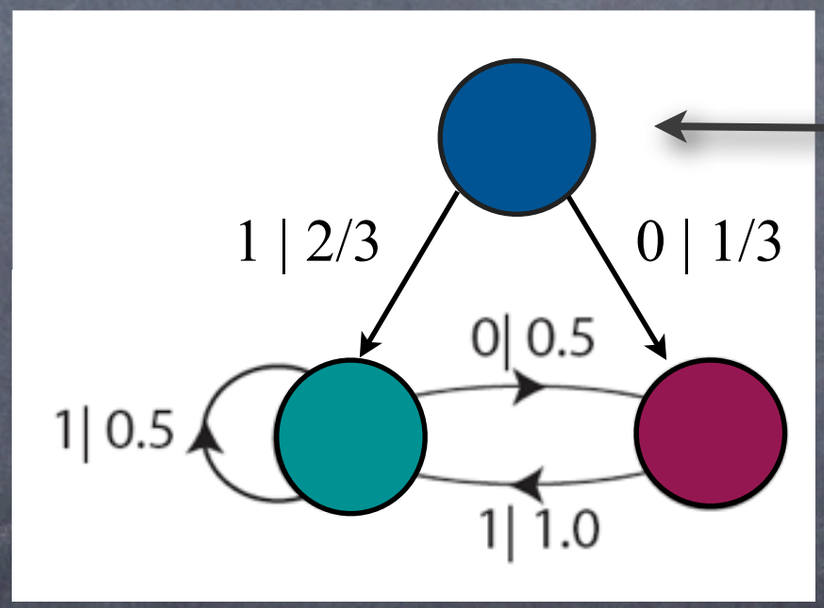
- Weakly causally compressible: RD curve is close to a straight line (small curvature).

- Strongly causally compressible: large curvature.

- Fully causally compressible: all predictive information can be kept with a model that has a complexity smaller than H[h].

Two examples of fully causally compressible processes:

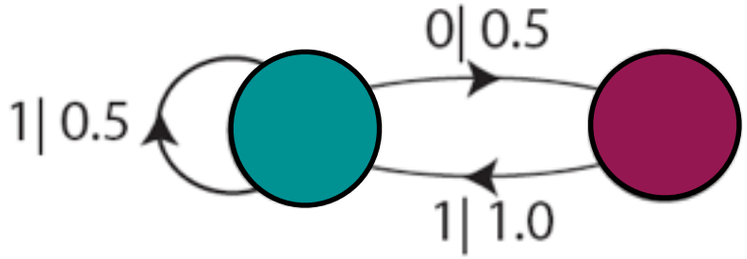
Example 1: Golden Mean Process

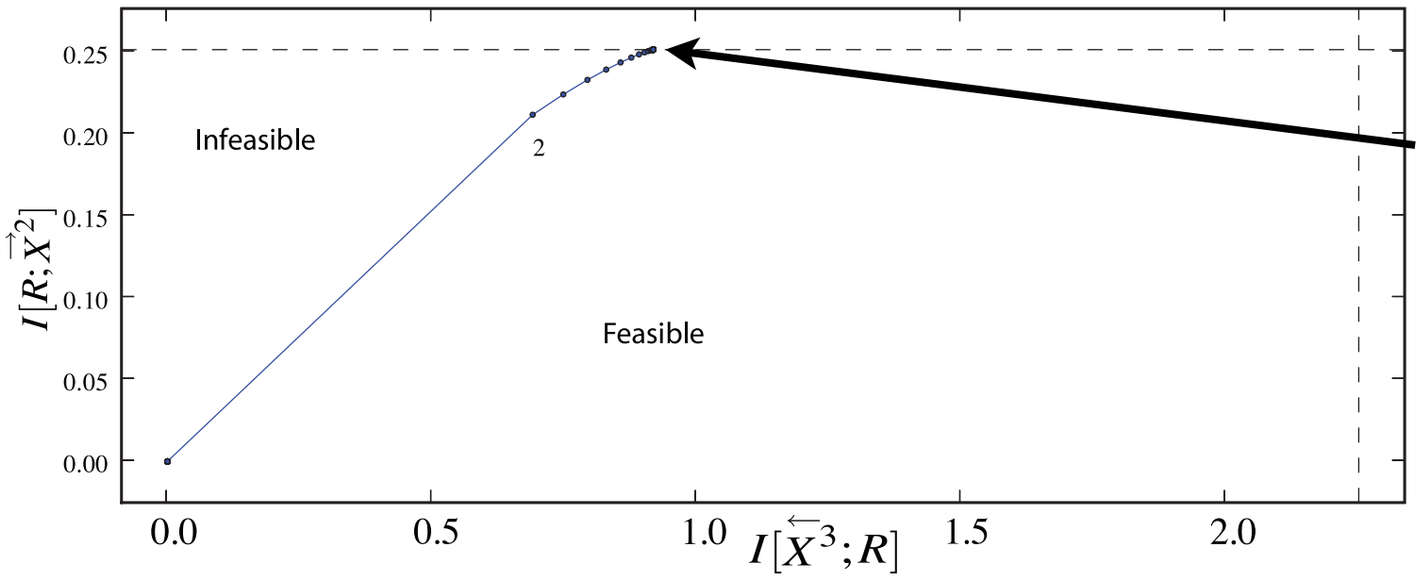


- Produces all binary sequences except those with 00
- Transient state vanishes for history lengths > 0.
- Can be modeled by a 2nd order Markov process.

Example 1: Golden Mean

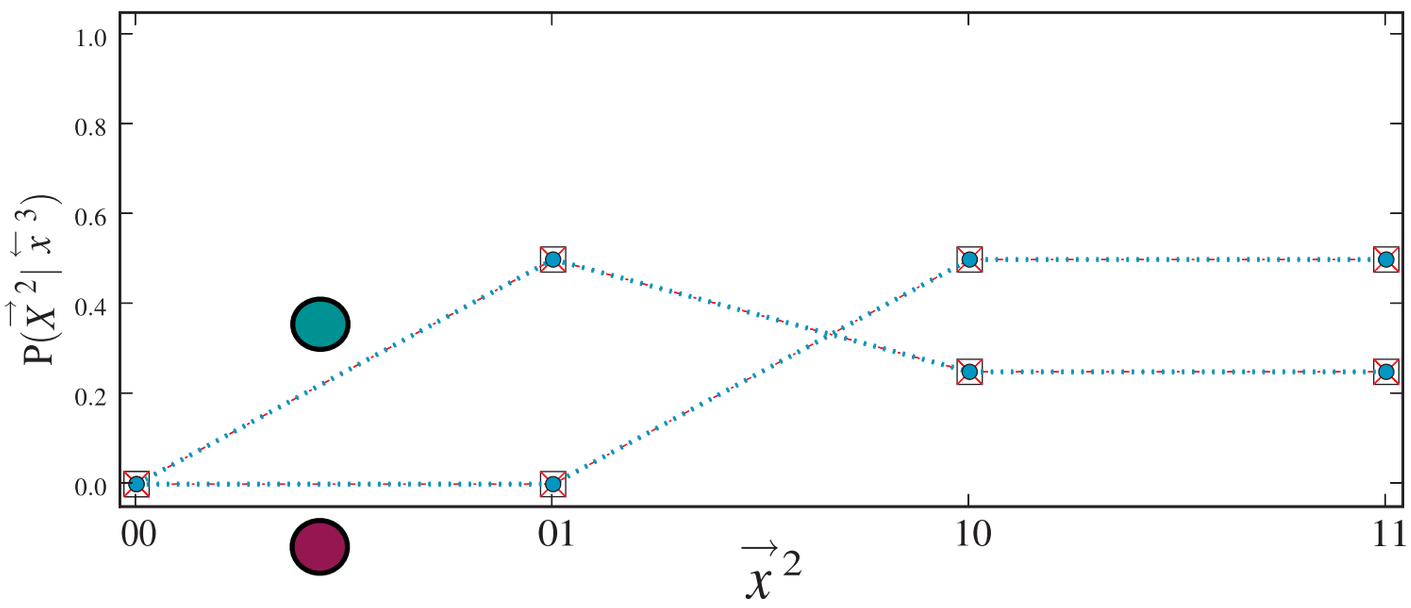
Produces all binary sequences except those with 00





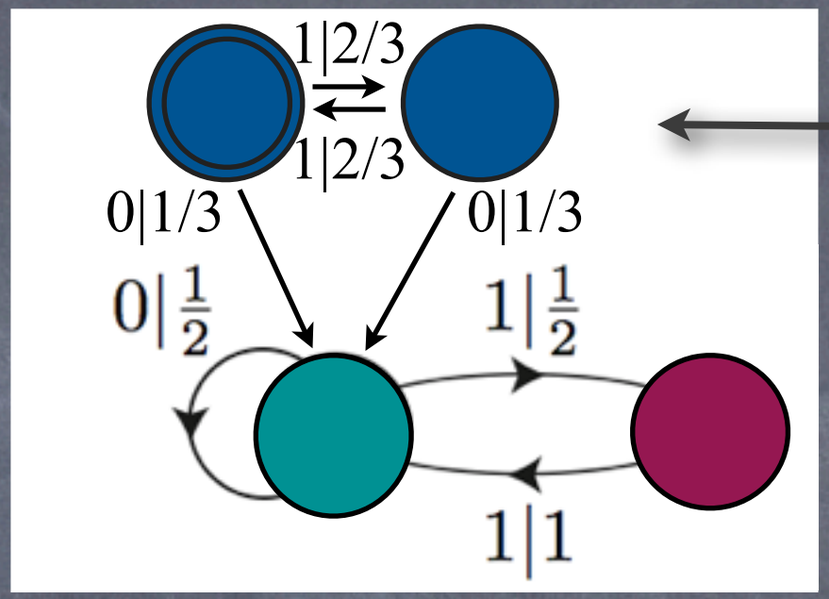
Algorithm finds the 2 states that describe the process fully - at a substantial compression.

Conditional future distributions associated with the best 2 state model



Example 2: Even Process

Even blocks of 1s.

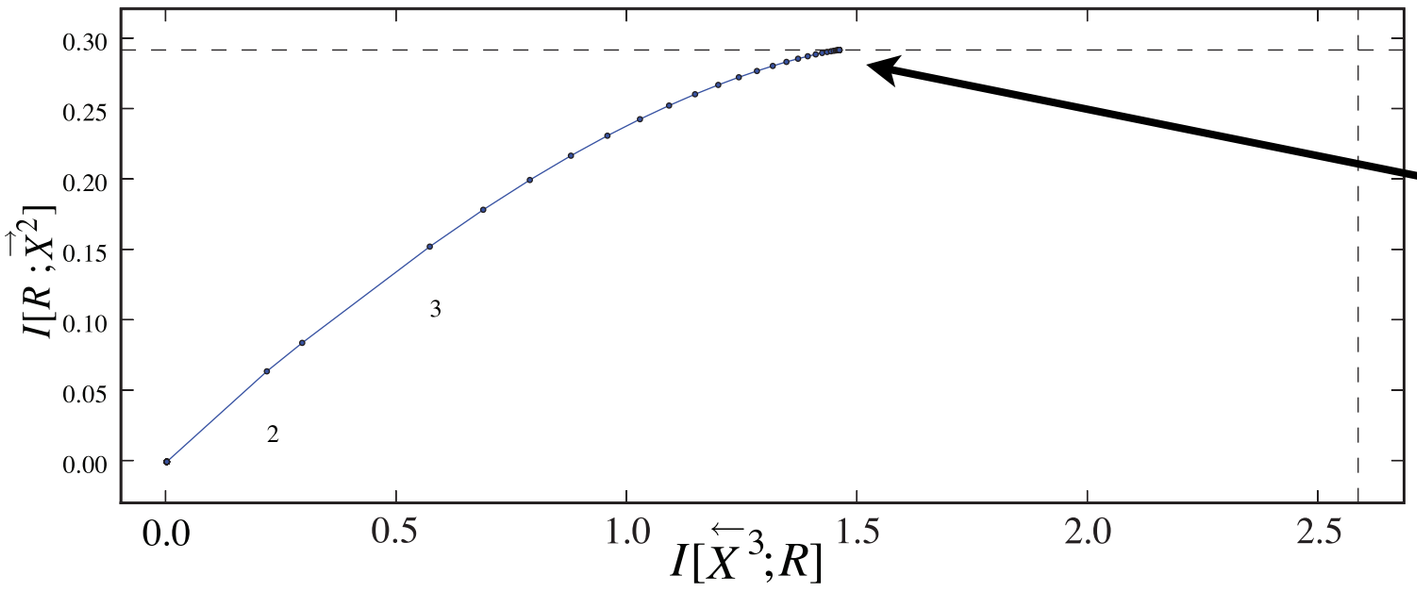


transient states

- Irreducible forbidden words: {010, 01110, 0111110, ...} - countably infinite
- No finite order Markov process can model the Even process.

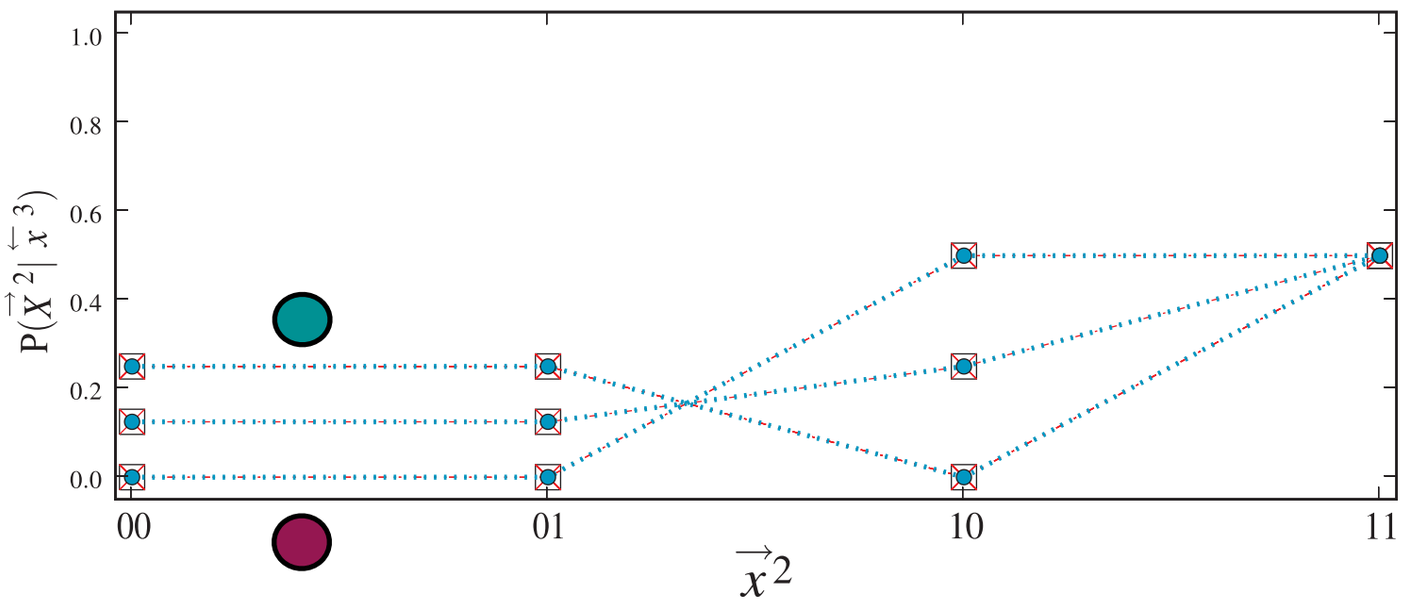
Example 2: Even Process

Generates all sequences with blocks 1s of even length



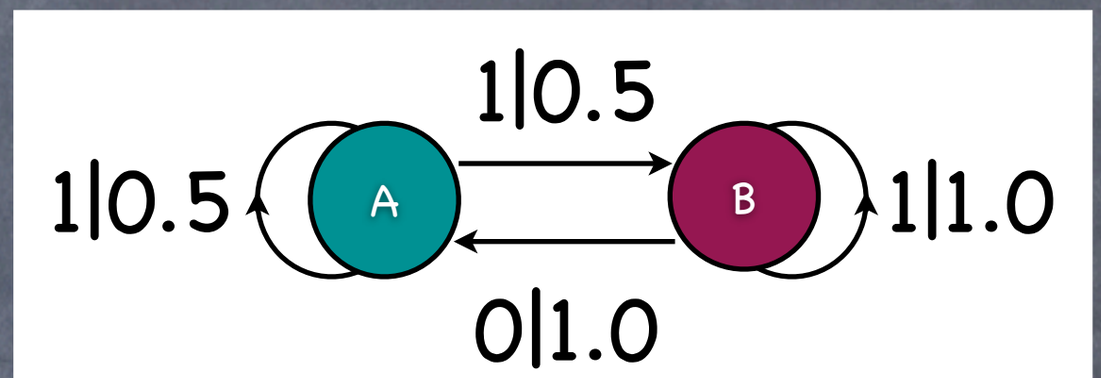
Algorithm finds the 3 states that describe the process fully - at a substantial compression.

Conditional future distributions associated with the best 3 state model



Example of a strongly causally compressible process: "Simple Nondeterministic Source" (SNS)

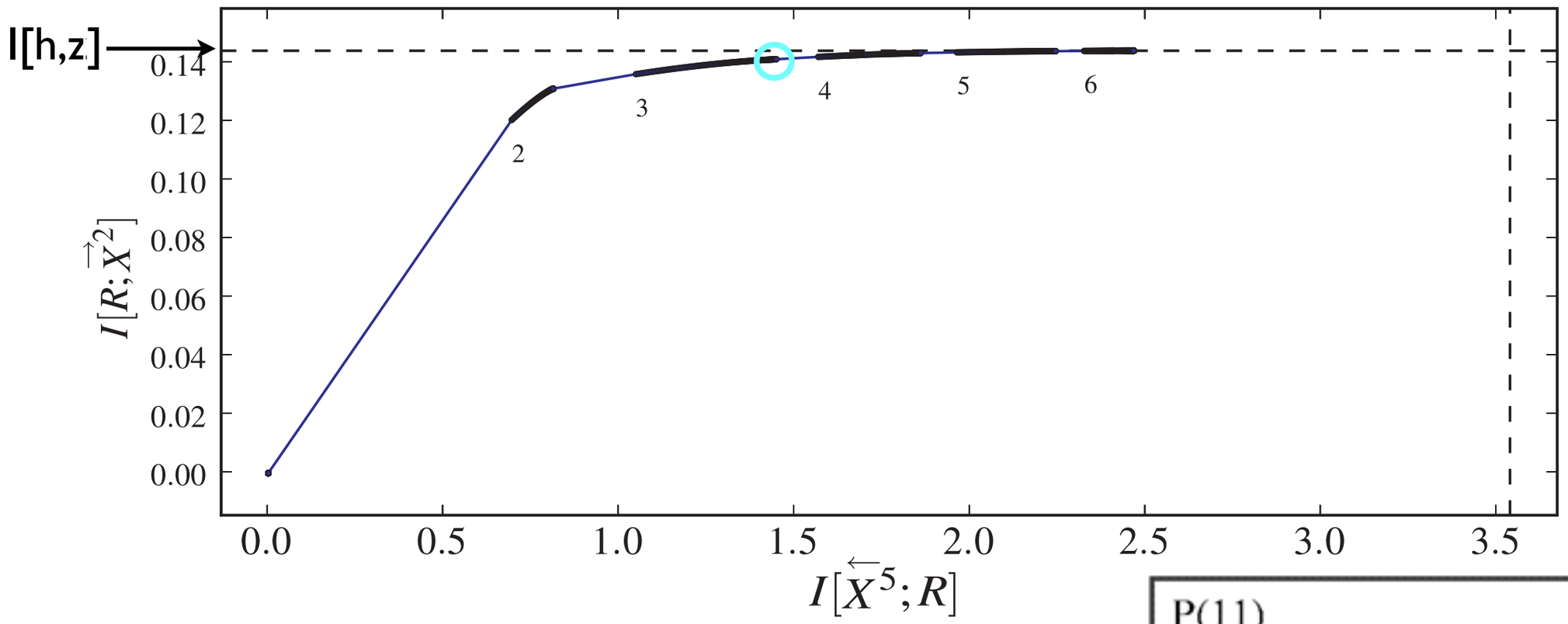
- HMM description:



- State A emits a 1 with probability 1.0 – next state is A with probability 0.5

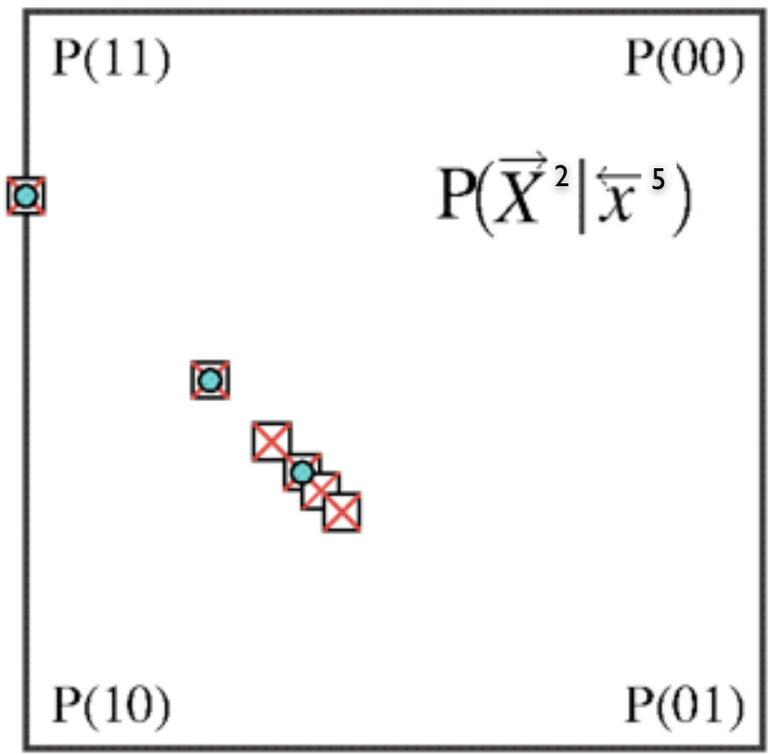
- State B emits a 1 with probability 0.5
  - if a 0 is emitted, then the next state is A.
  - if a 1 is emitted, then the next state is B.

Example (SNS)



Future distributions:

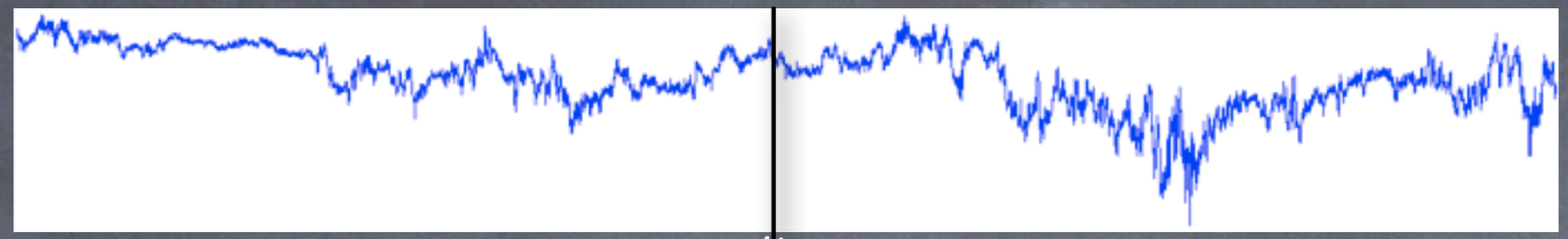
- ✗ full historical information
- □ best 6-state model
- ● best 3-state model



Extension to dynamic learning

- Dynamical learning system finds (S. Still, Entropy 2014) asymptotically the "epsilon-machine" (Crutchfield and Young,1989), a deterministic hidden Markov model that is maximally predictive.
- All characteristics of the underlying process can be computed from the epsilon machine, in many cases analytically: entropy rate, predictive/stored information,... (Crutchfield and colleagues,1989 - 2010)

1a. Signal: x(t)



find: $p(s|h)$

h(t) → map → s(t) → ! predict → z(t)

1b. Dynamics

find: $p(s_{t+1}|x_t, s_t)$

x(t), s(t) → map → s(t+1) → ! predict → z(t)

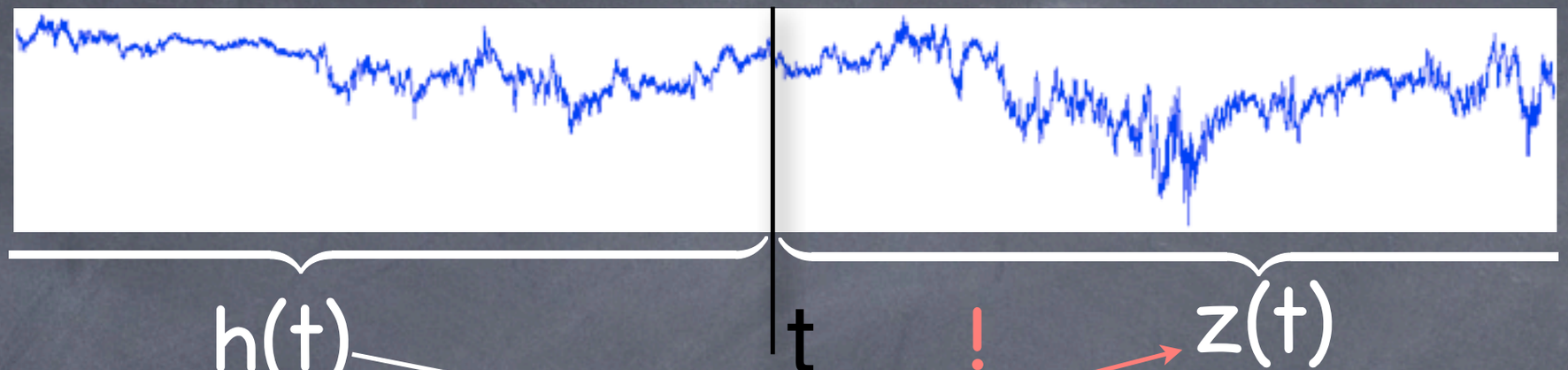
2. include feedback

find: $p(s_{t+1}|x_t, s_t, a_t)$
$p(a_{t+1}|x_t, s_t, a_t)$

x(t), a(t), s(t) → map → a(t+1), s(t+1) → change → z(t); ! predict

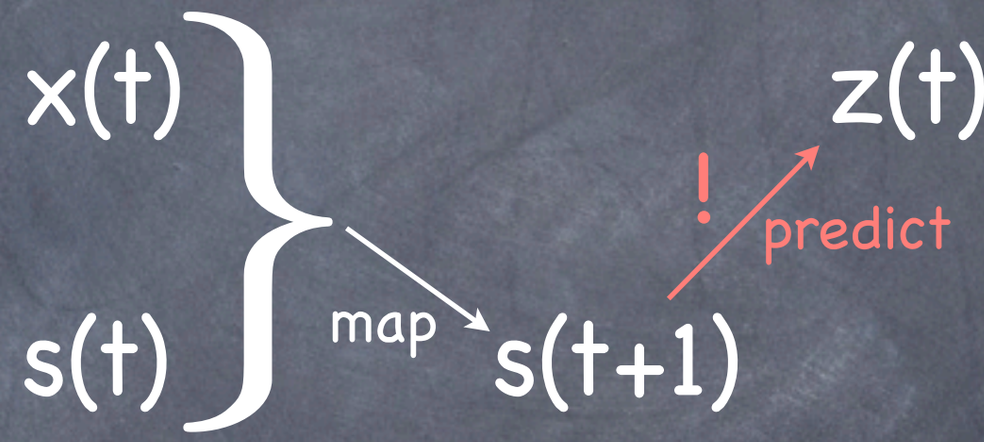
Extension to feedback ("interactive learning")

1a. Signal: x(t)



find: $p(s|h)$

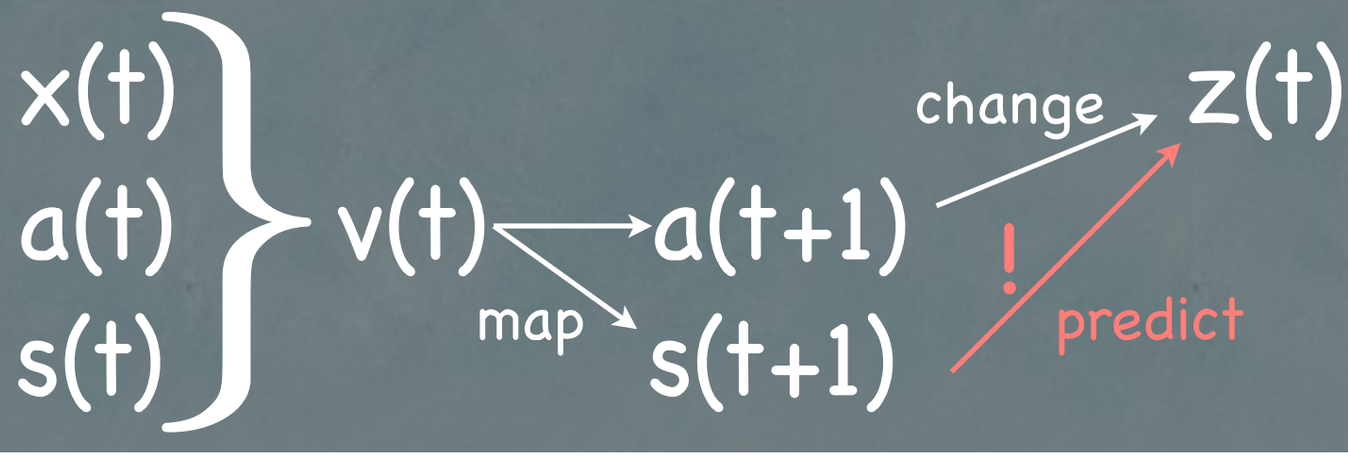
1b. dynamics



find: $p(s_{t+1}|x_t, s_t)$

2. include feedback

find: $p(s_{t+1}|x_t, s_t, a_t)$
$p(a_{t+1}|x_t, s_t, a_t)$



Here we ran out of time!

maybe more about this later?

if interested, you can read:

S. Still. **Information theoretic approach to interactive learning**. ***EPL*** 85 (2009) 28005.

Part II

Summary & Reminder

- Thermodynamic limits of information processing directly lead to a learning method known as "Information Bottleneck (IB)":

  Dissipation >= kT I[X,Y] - kT I[Y,Z]

  Therefore: min Dissipation => IB

- Also explains Shannon's rate-distortion theory (RDT) approach as the least effort required to represent the data achieving a certain fidelity

- Direct applications in clustering: Soft K-means, deterministic annealing, distributional clustering

K means

- Soft K-means and deterministic annealing can be derived from both IB and RDT with choice of euclidean distance as metric (= "similarity measure" = distortion function).
- Probabilistic assignments, exponential fall off with distance.
- Hard K-means is limit case. Deterministic assignments, each data point belongs to one cluster.
- Deterministic annealing helps to prevent getting stuck in local minima.

How many clusters?

Using information theory to address this fundamental question in cluster analysis.

The problem:

- Cluster analysis typically groups N data into K classes.
- The solution depends on K.
- If K is not known, how do we estimate it?
- Is there a "correct" K?
- Can we define a "best" K?

Traditional approaches

- Introduce a "goodness" criterion to evaluate the clustering solution at different K.
  Problem: the criterion has to be chosen ad hoc.

- Bayesian Inference for mixture models (parameterized models).
  Problem: specific assumptions about underlying probability distribution.

- Stability analysis. Idea: any reasonable partition should be stable under sample fluctuations. Difficulty: estimating the error. There are several different approaches, here I discuss one of them.

Using Information Theory

1) View clustering as lossy compression: construct a compressed representation of the data. $p(c|x) : X \mapsto C$

2)Define relevant information as the information that the data carry about a variable of interest, y. $I[x; y]$

3)The clustering problem becomes: Maximize the relevant information kept under the constraint that the coding cost is fixed (Information Bottleneck).

$$\max_{p(c|x)} \left(I[c; y] - \lambda I[x; c] \right)$$

- Note that the objective function is a functional of the joint probability density p(x,y).

$$F[p(c|x); p(x,y)] = \left(I[c;y] - \lambda I[x;c]\right)$$

- In most applications, we do not know p(x,y), rather, we have to estimate it from the N data that we are given.

$$\hat{p}(x,y) = p(x,y) + \delta p(x,y)$$

estimate — true distribution — sampling error

- We can not evaluate the objective function, we only have an estimate of it, which is naively given by evaluation at the estimate we have of the distribution.

Intuition

- The sampling error induces a systematic upward bias in the estimation of relevant information.

$$I[c;y] \leq I[x;y] \leq I[x;y]_{\mathrm{est}} = I[x,y]|_{\hat{p}(x,y)}$$

- It looks as if there is more relevant information than there really is -> danger of over-fitting, if we keep too much detail.

- The size of the sampling error depends on the data set size N, and so does the bias.

- Can we estimate the bias?

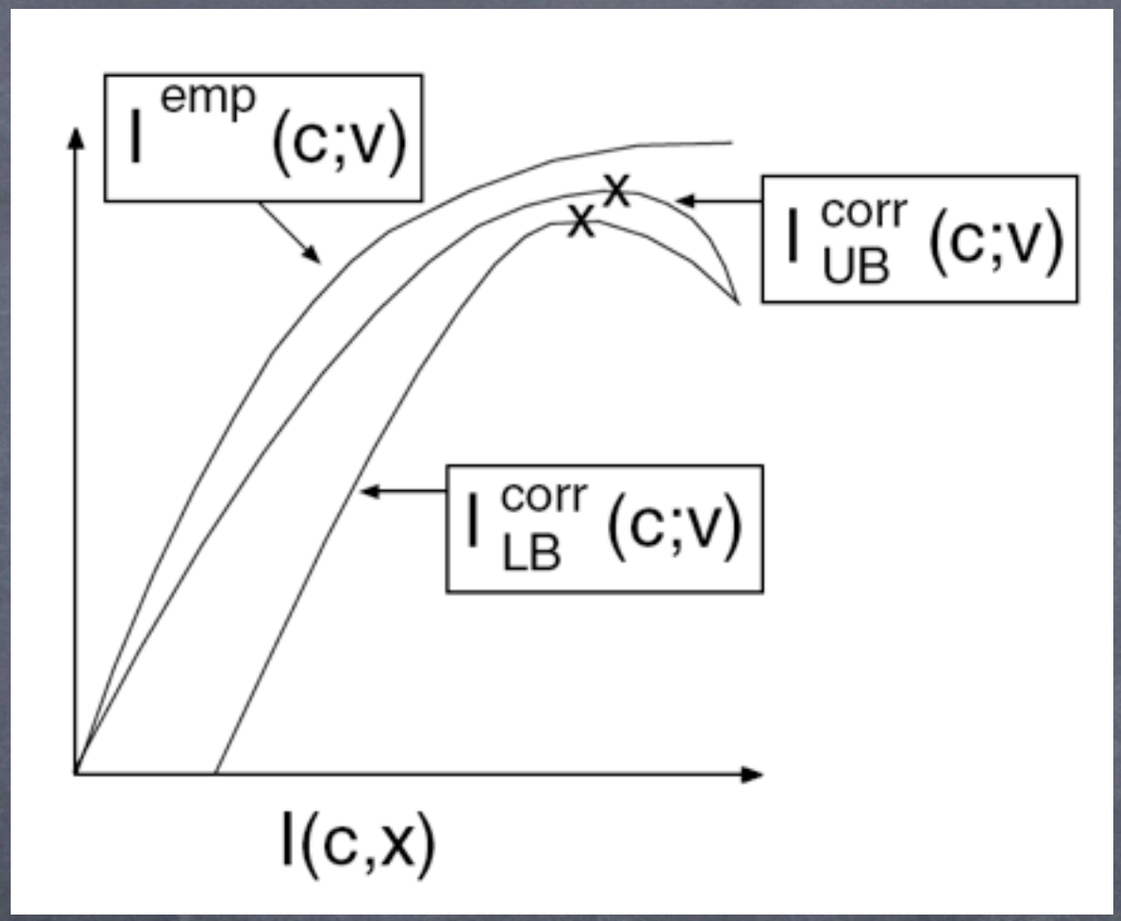
Idea

- Estimate the error: $\mathcal{E}(N) = I[c;y]_{\text{est}} - I[c;y]_{\text{true}}$
- Then, instead of maximizing $I[c;y]_{\text{est}}$ maximize: $I[c;y]_{\text{est}} - \mathcal{E}(N) \simeq I[c;y]_{\text{true}}$
- How to compute the error? Use Taylor expansion... $I[\hat{p}(x,y)] \simeq I[p(x,y)] + \mathcal{E}(N)$
- This works only for reasonably large N (the sampling error has to be reasonably small).

Intuition

- The IB trade-off is related to the amount of detail that we keep. The more detail we allow, the more relevant information do we keep - the information curve is monotonic.
- By subtracting the error we make due to under-sampling, the resulting curve should have a maximum at the optimal trade-off. This optimum should depend on the data set size N.

Optimal trade-off



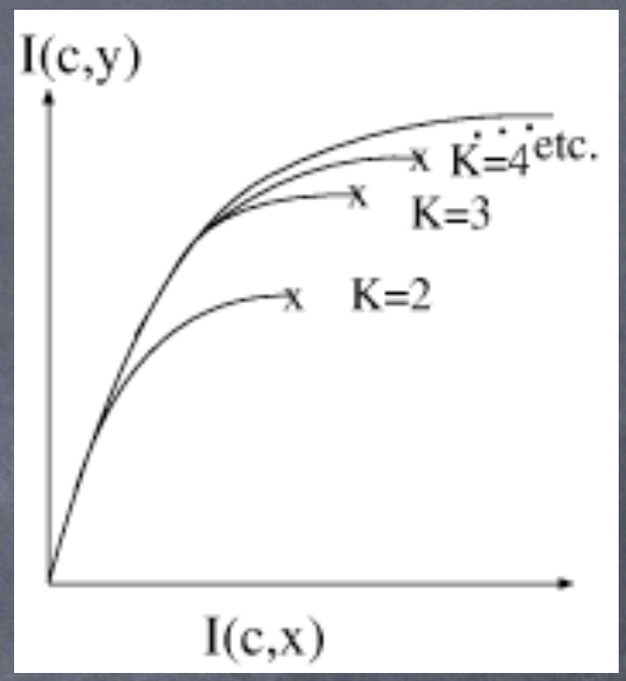
- Analytical result: upper bound and lower bound on the corrected information curve and hence on the optimal trade-off.

How many clusters?

- Any finite "temperature" still embodies a level of uncertainty in the assignment of data to clusters, and thus it is tricky to associate such a map with a number of clusters.
- A number of clusters is only well defined for deterministic assignments.

How many clusters?

- Remember: Information curve



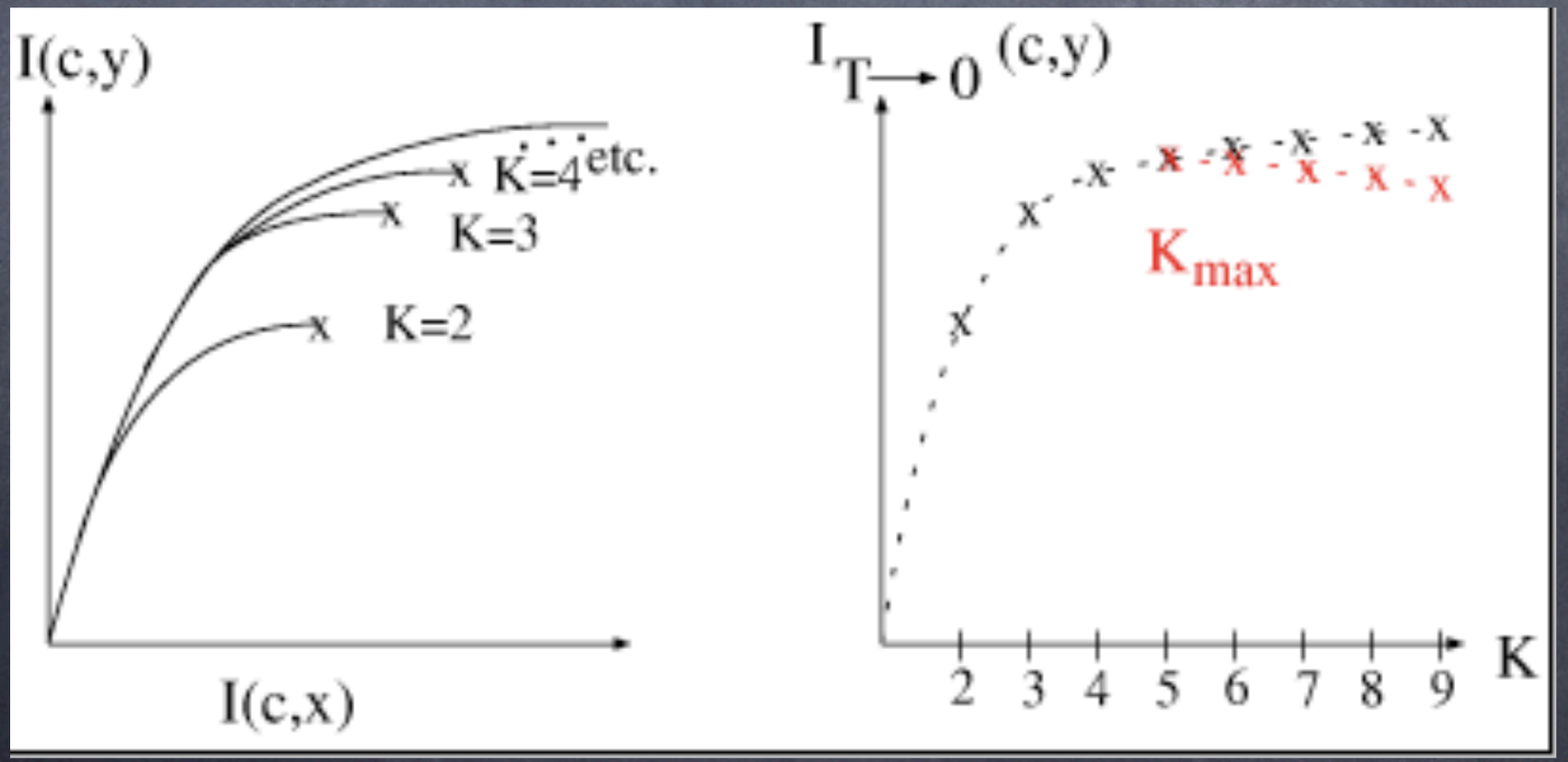
- Now: Fix the number of clusters to Nc -> anneal -> get maximal relevant information as a function of Nc: monotonically increasing

How many clusters?

- As before, it looks like the more clusters we use, the more relevant information we can keep.
- Danger of over-fitting, due to the upward bias in the relevant information because of the sampling errors!

How many clusters?

- Plot the corrected relevant information
  -> curve has maximum (or plateau)
  -> easy to detect the maximum number of clusters.



Calculations

- Taylor expansion results in a series of terms.
- Let's compute the leading order term for the simple case in which p(x) is known, and we are estimating only p(y|x).
  Example: x is an index -> p(x) = 1/N.
- assume that the perturbation has zero average $\langle \delta P(v|x) \rangle = 0$
- Then, the first order term is zero -> leading term is the 2nd order term.

Leading Error:

$$(\Delta I(c;v))^{(2)} \simeq \frac{1}{2\ln(2)N}\sum_{vc}\frac{\sum_x [P(c|x)]^2 P(v|x)P(x)}{P(c|v)P(v)}$$

For T->0 (hard clustering solution)

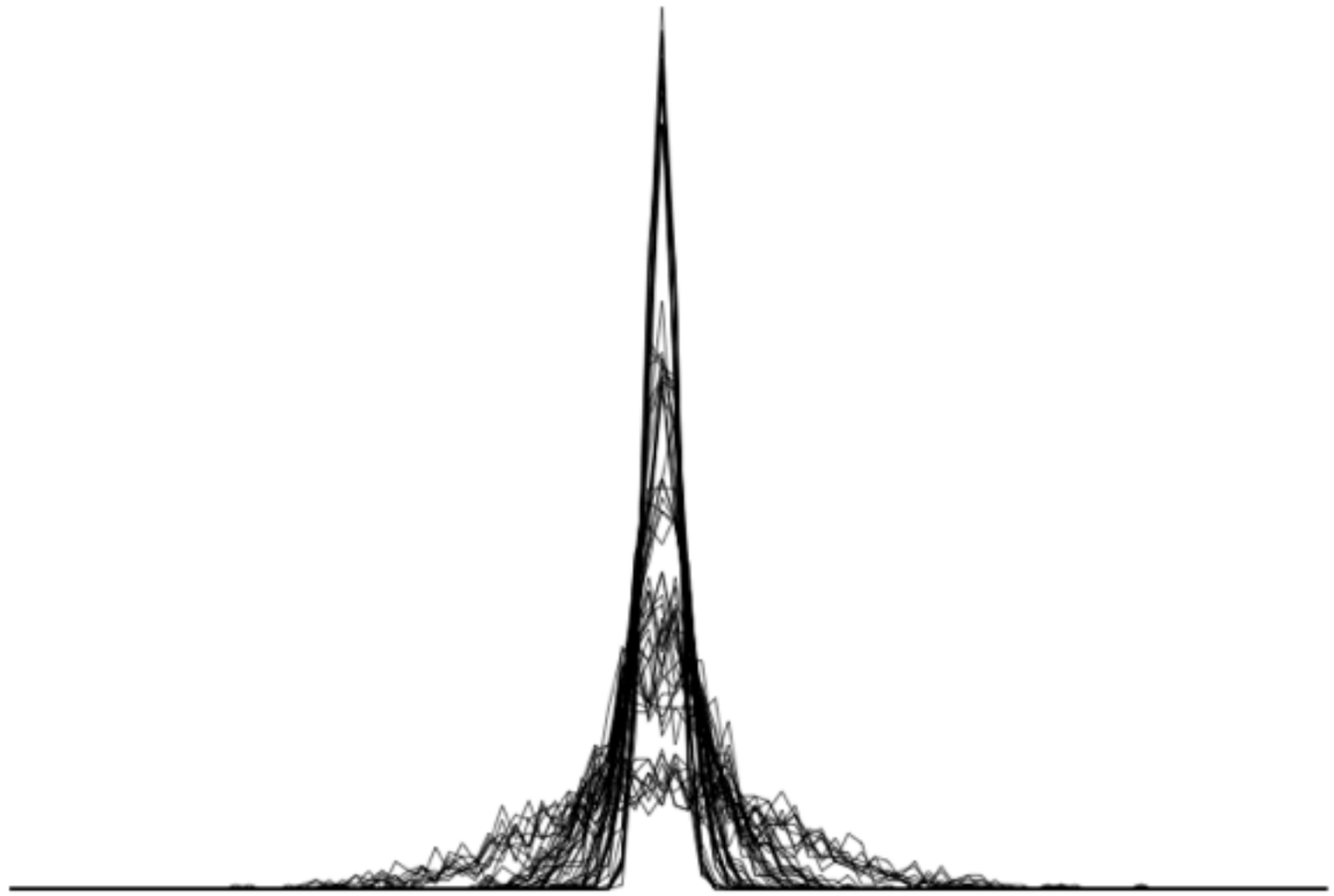
$$(\Delta I(c;v))^{(2)}_{T\to 0} = \frac{1}{2\ln(2)}\frac{K_v}{N}N_c$$

Complexity control term,
subtract from total captured relevant information

Numerical tests

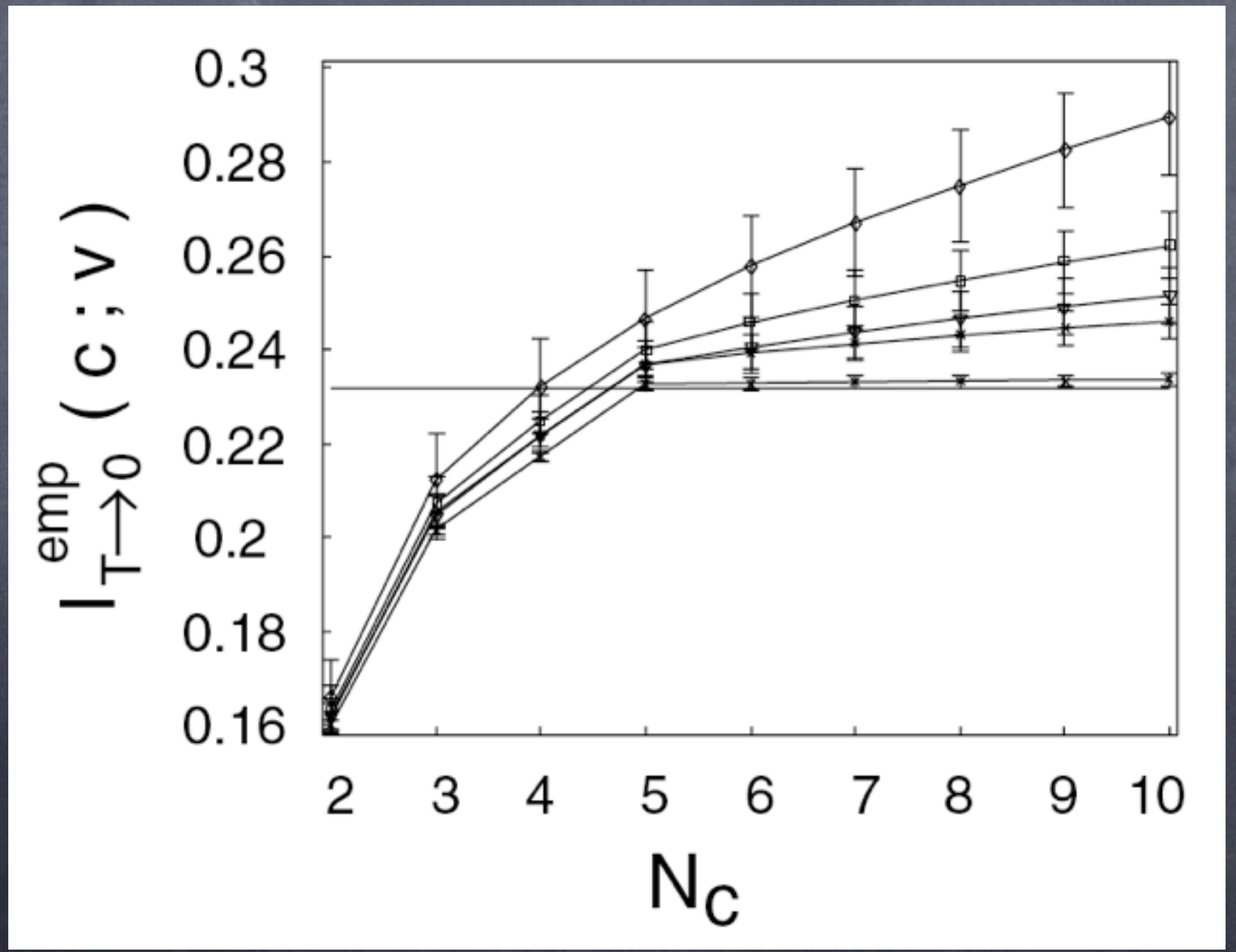
Synthetic Data - I

$P(y|x) = \mathcal{N}(0, \alpha(x))$ with 5 possible values for α, $p(\alpha) = const.$



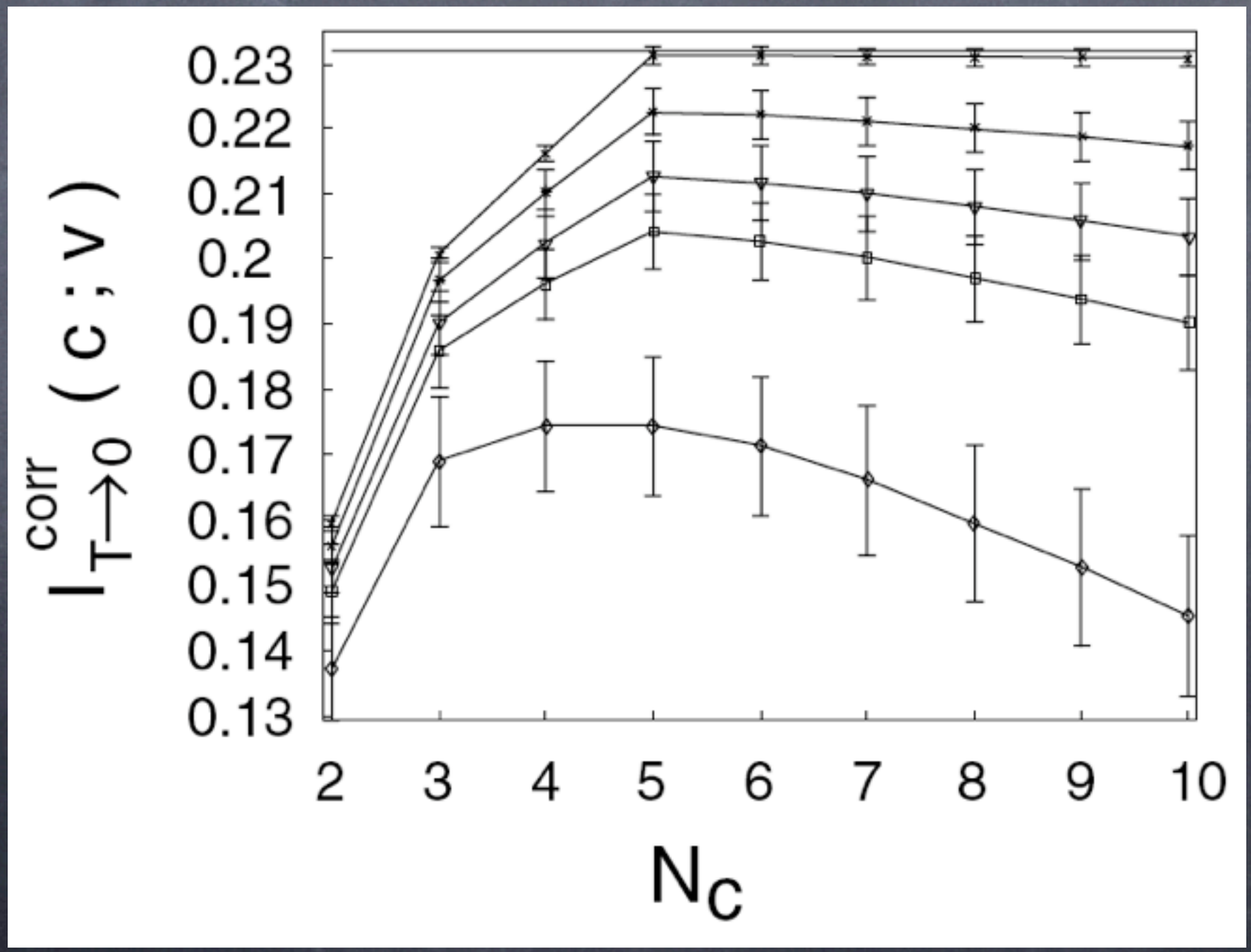
Empirical curve

Synthetic Data - I



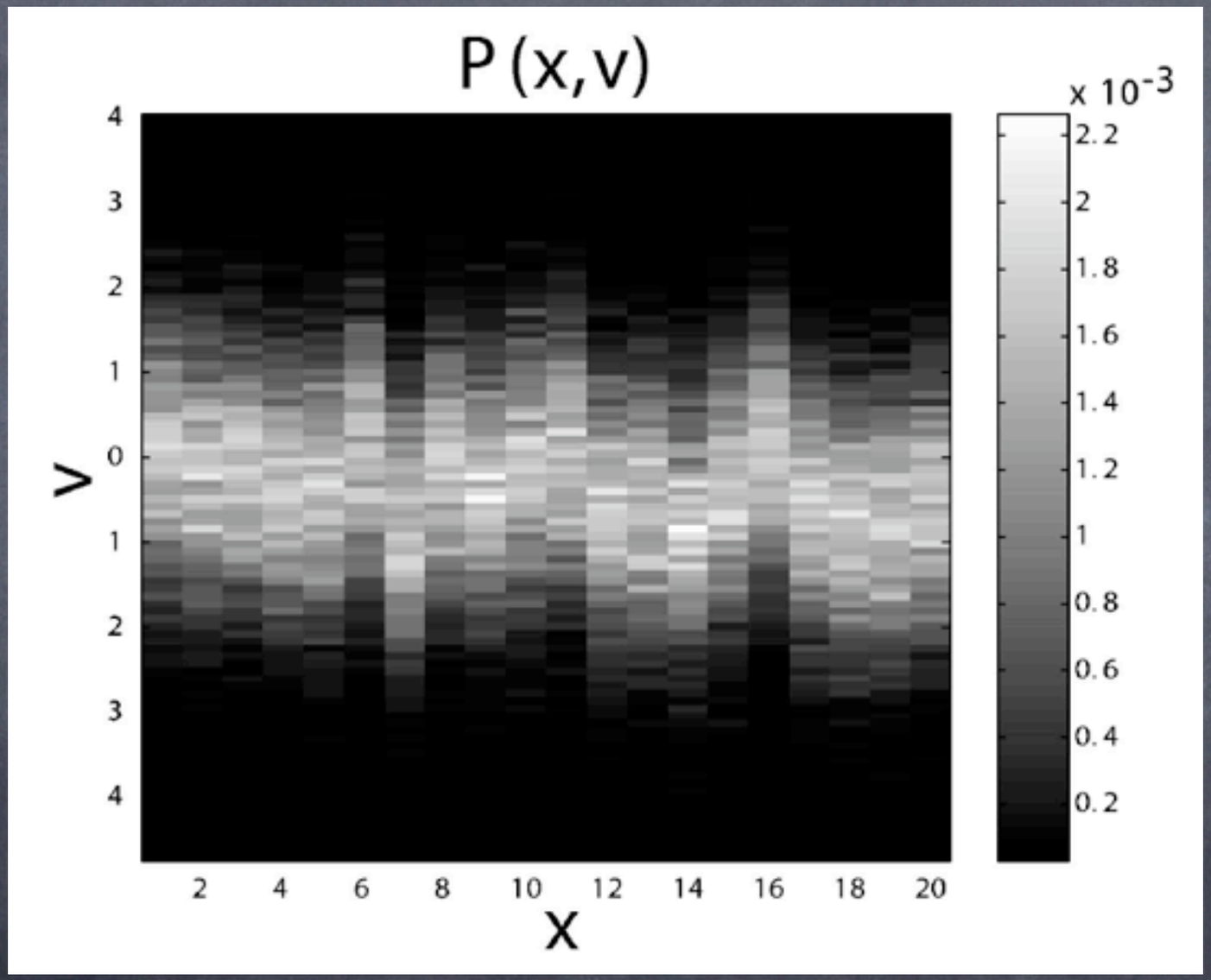
Corrected curve

Synthetic Data - I



Synthetic Data - II

K Gaussians with the same variance, but different means.

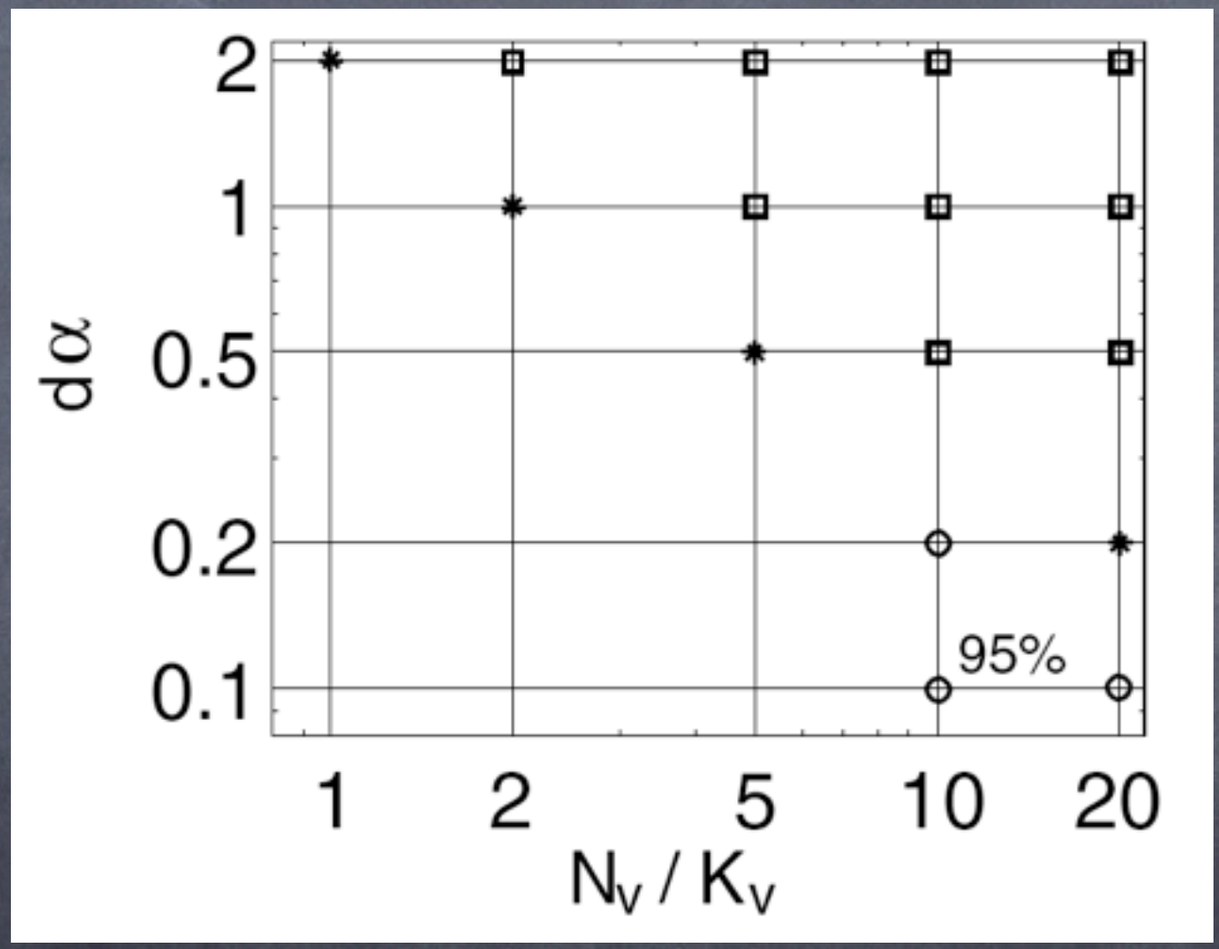


- Task is harder the closer the means are together, the larger the number of classes and the smaller N.

Results

Synthetic Data - II

Spacing of means in units of the variance



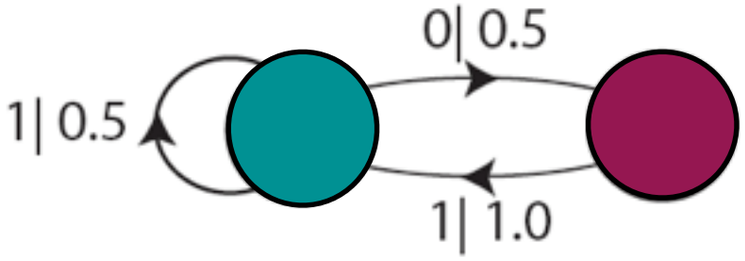
found correct # clusters for K=

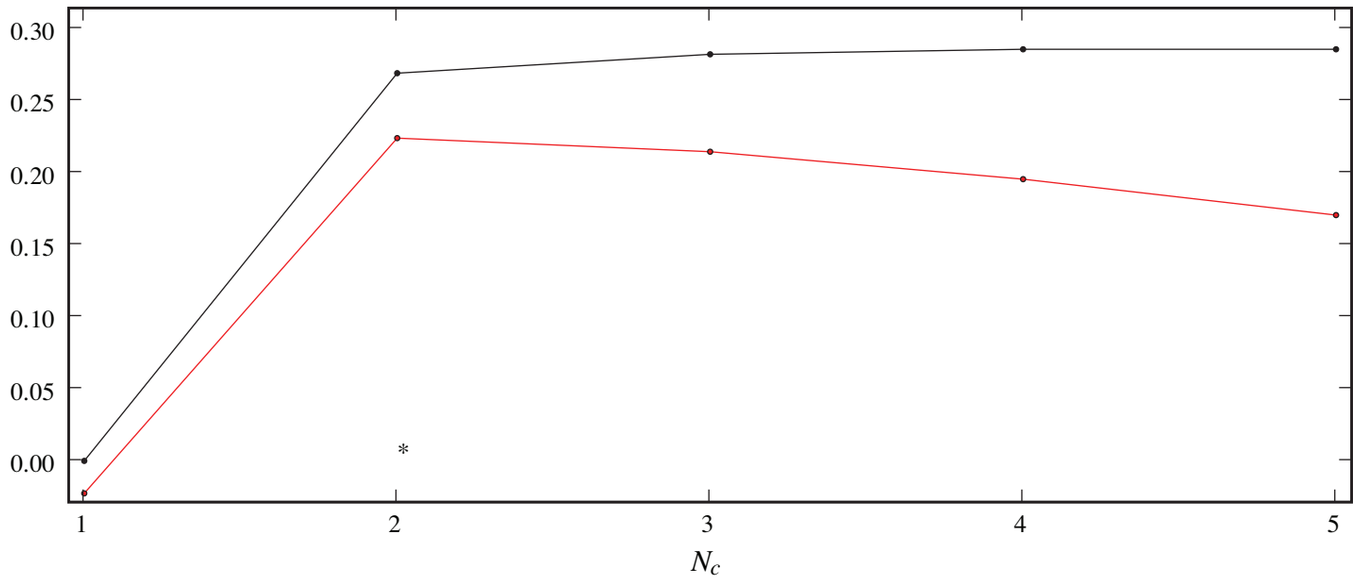
- {2,5}
- {2}
- {2,5,10}

Data points per bin (average)

Golden Mean Process

Produces all binary sequences
except those with 00

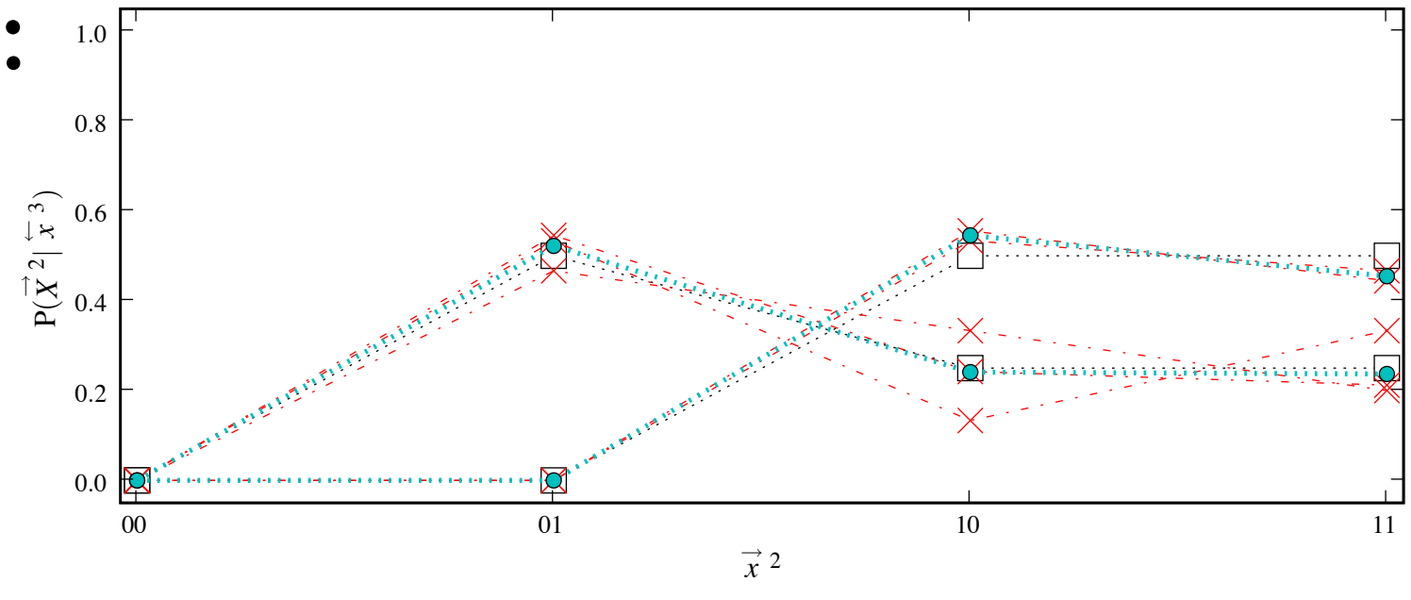




Correction finds optimal number of states (2)

Future distributions:

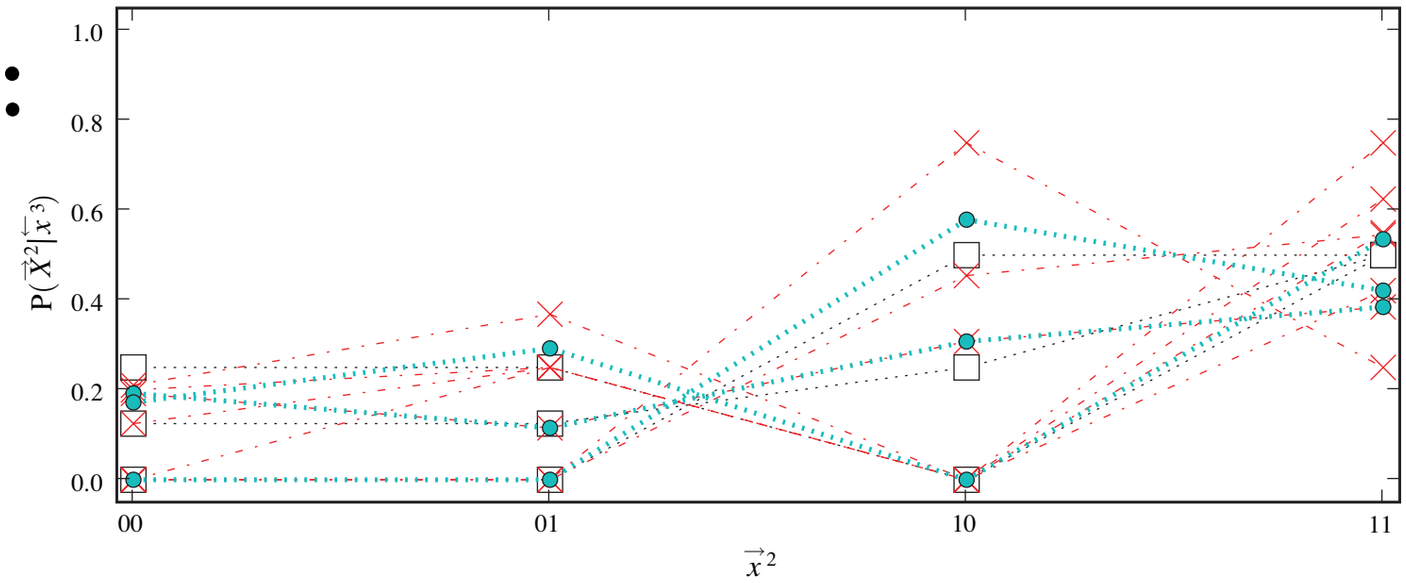
- ✗ Finite Data
- □ ideal (truth)
- ● result (algorithm)

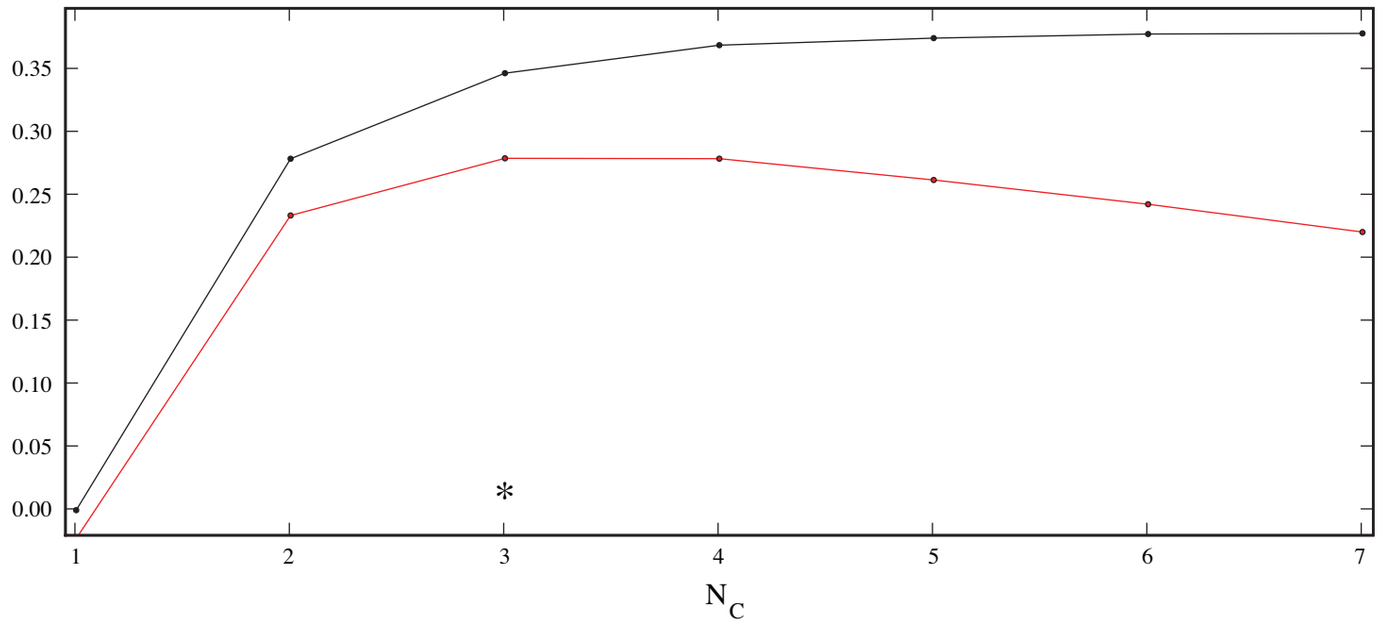


Even Process

Future distributions:

- ✗ Finite Data
- □ ideal (truth)
- ● result (algorithm)

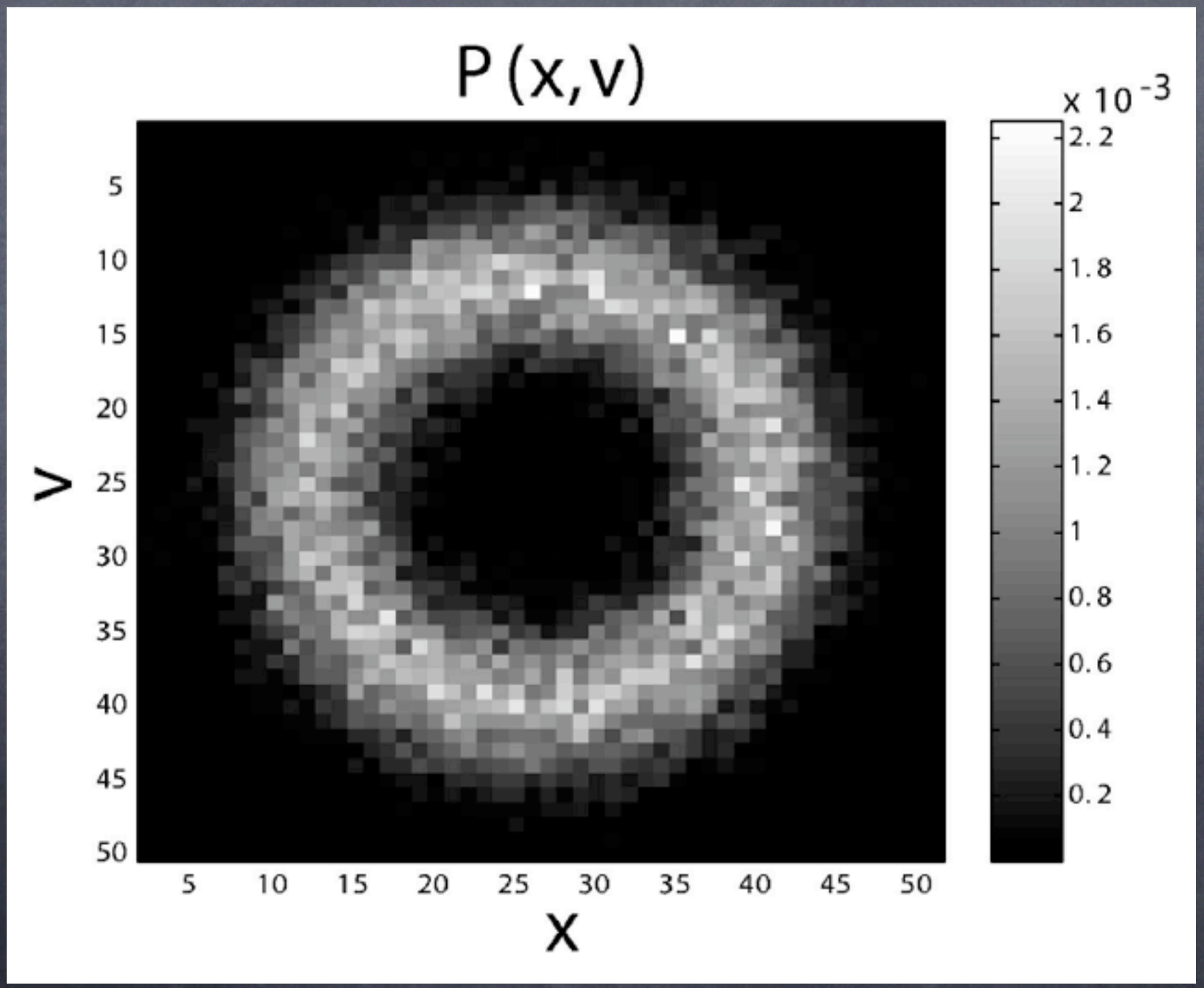




Correction finds optimal number of states (3)

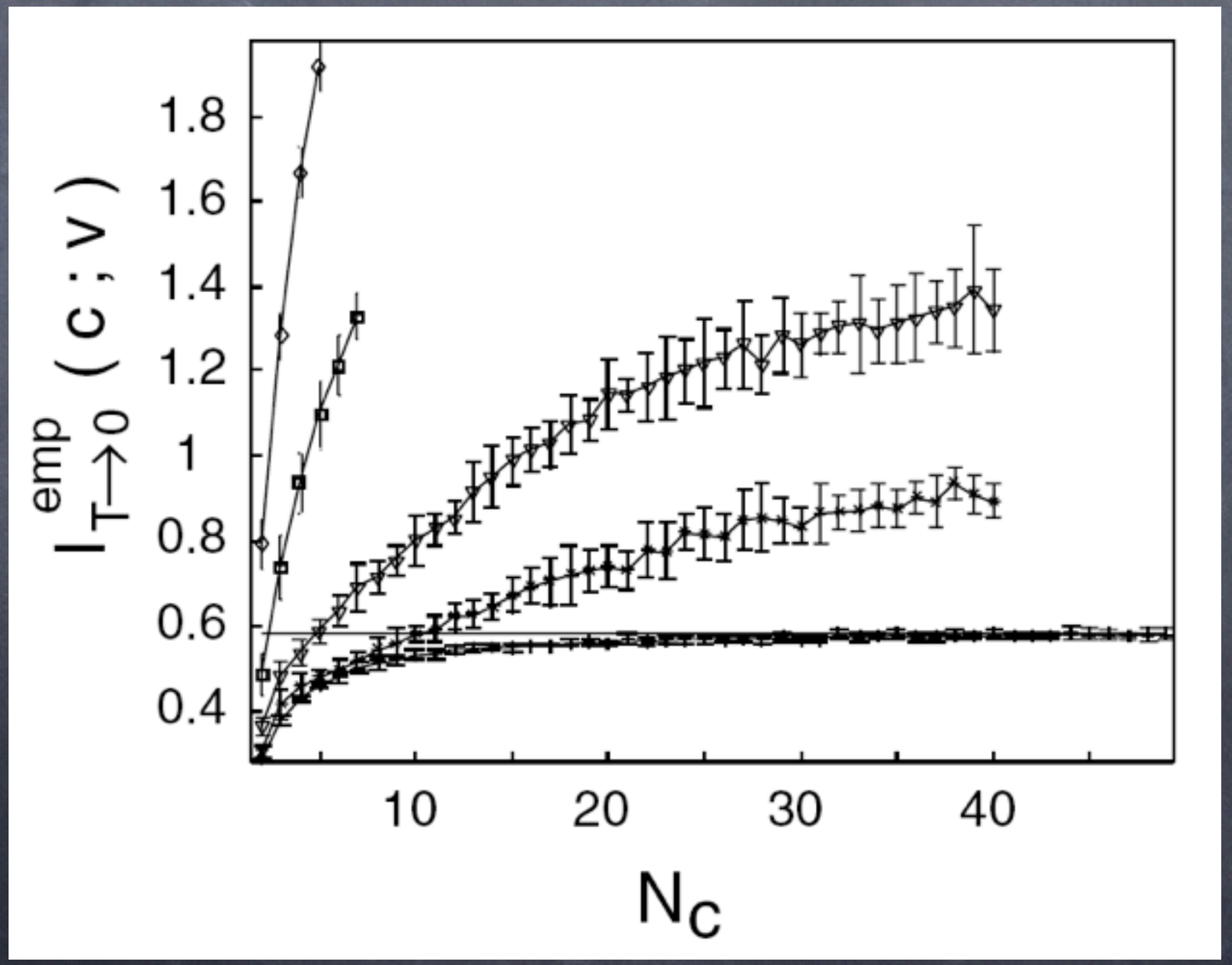
Synthetic Data - III

estimate P(x,y)



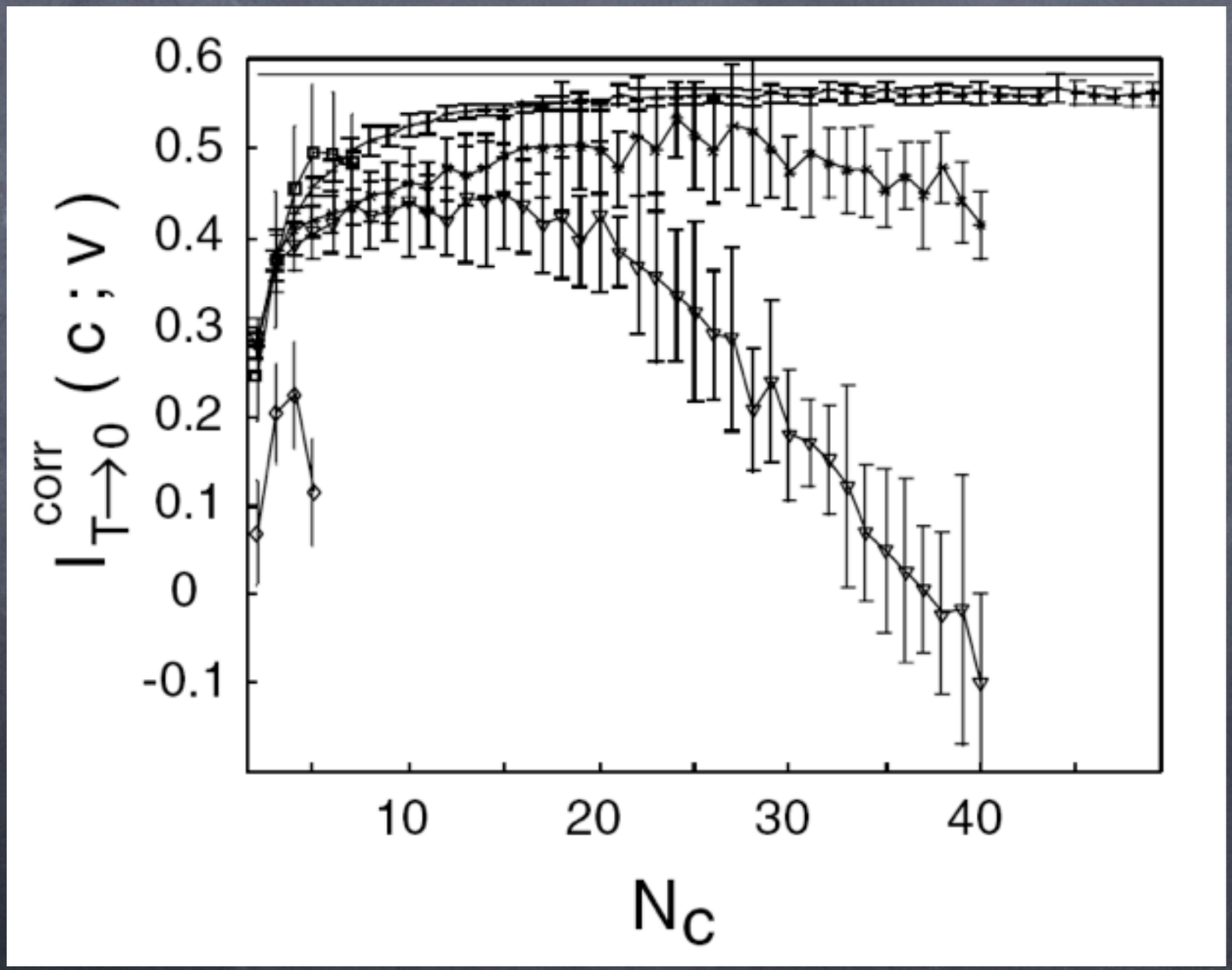
Empirical curve

Synthetic Data - III



Corrected curve

Synthetic Data - III



Conclusion

- Bottleneck idea applied to clustering offers a way to eliminate much of the arbitrary assumptions in clustering.
- Perturbation theory can be used to derive a complexity control criterion
- This renders bottleneck useful for real world applications, as a learning algorithm.

- Note that the objective function is a functional of the joint probability density p(x,y).

$$F[p(c|x); p(x,y)] = \left(I[c;y] - \lambda I[x;c]\right)$$

- In most applications, we do not know p(x,y), rather, we have to estimate it from the N data that we are given.

$$\hat{p}(x,y) = p(x,y) + \delta p(x,y)$$

estimate — true distribution — sampling error

- We can not evaluate the objective function, we only have an estimate of it, which is naively given by the evaluation of F at the estimate of p.

$$F_{\text{true}} = F[p(x,y)] \text{ (unknown)}$$

$$F_{\text{est}} = F[\hat{p}(x,y)] \text{ (known)}$$

$$\text{Error} : \mathcal{E} = F_{\text{est}} - F_{\text{true}}$$

Homework

- Implement Soft K-means algorithm.
- Helpful reading (on course website): K. Rose, "Deterministic Annealing for Clustering, Compression, Classification, Regression, and Related Optimization Problems," Proceedings of the IEEE, vol. 80, pp. 2210-2239, November 1998.