Matematical structure of Information Distortion methods

Tomáš Gedeon

with

Alex Dimitrov (Neuroscience)

Albert Parker, Collette Campion (Mathematics)

Brendan Mumey (Computer Science)

Montana State University

Problems

I will discuss some common mathematical themes of these problems:

Problems

I will discuss some common mathematical themes of these problems:

Rate distortion theory (Shannon 1948)

Problems

I will discuss some common mathematical themes of these problems:

Rate distortion theory (Shannon 1948)

Deterministic annealing (Rose 1990's).

Problems

I will discuss some common mathematical themes of these problems:

Rate distortion theory (Shannon 1948)

Deterministic annealing (Rose 1990's).

Information Bottleneck (Tishby *et. al* 1999.)

Problems

I will discuss some common mathematical themes of these problems:

Rate distortion theory (Shannon 1948)

Deterministic annealing (Rose 1990's).

Information Bottleneck (Tishby *et. al* 1999.)

Information Distortion (Dimitrov, Miller 2001) .

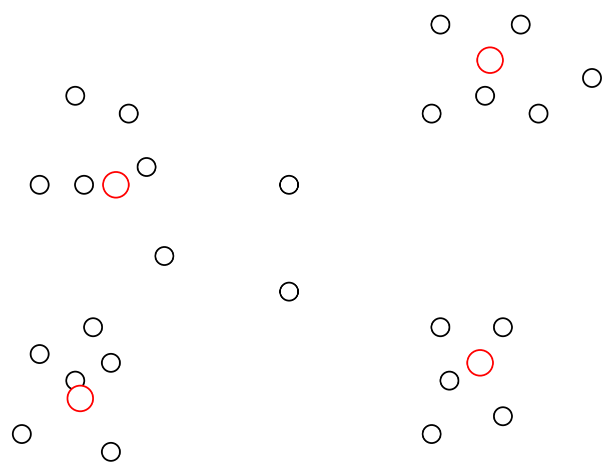
Rate distortion theory

Given:

- X a discrete random variable (source);
- N the size of reproduction variable $\hat{X}$,
- distortion function $d(\hat{x}, x)$

Goal: Find assignment $q = q(\hat{x}|x)$

$$min_{q:E_p d(x,\hat{x}) \leq D} I(X, \hat{X}).$$



Clustering via Deterministic Annealing

Given:

- X -data set
- N -number of centers of clusters in $\hat{X}$
- distortion function $d(\hat{x}, x)$, usually Euclidean distance

Goal: Find assignment $q = q(\hat{x}|x)$ and positions of centers of clusters $\hat{x}$ to

$$max_{q,\hat{x}:E_p d(x,\hat{x})<D} H(\hat{X}|X)).$$

Select $\hat{x} = \sum_x q(\hat{x}, x)x$. Then

$$max_{q:E_p d(x,\sum_x q(\hat{x},x)x)\leq D} H(\hat{X}|X)).$$

Information Bottleneck

Given:

- a **pair** of random variables X, Y with $p(x, y)$ known
- N a number of elements of reproduction variable $\hat{X}$
- distortion function is $-I(\hat{X}, Y)$

Goal: Find assignment $q = q(\hat{x}|x)$

$$min_{q:-I(\hat{X},Y)\leq D} I(X, \hat{X}).$$

Markov chain:

$$Y \to X \to \hat{X}.$$

Information Distortion

Given:

- a **pair** of random variables X, Y with $p(x,y)$ known
- N a number of elements of reproduction variable $\hat{X}$
- distortion function is $-I(\hat{X},Y)$

Goal: Find assignment $q = q(\hat{x}|x)$

$$min_{q:-I(\hat{X},Y)\leq D} - H(\hat{X}|X).$$

Markov chain:

$$Y \rightarrow X \rightarrow \hat{X}.$$

Summary

After Lagrange multipliers:

- Information distortion

$$maxH(\hat{X}|X) + \beta I(Y, \hat{X})$$

- Information Bottleneck Method

$$max - I(X, \hat{X}) + \beta I(Y, \hat{X})$$

- Rate Distortion Theory

$$max - I(X, \hat{X}) - \beta D(X, \hat{X})$$

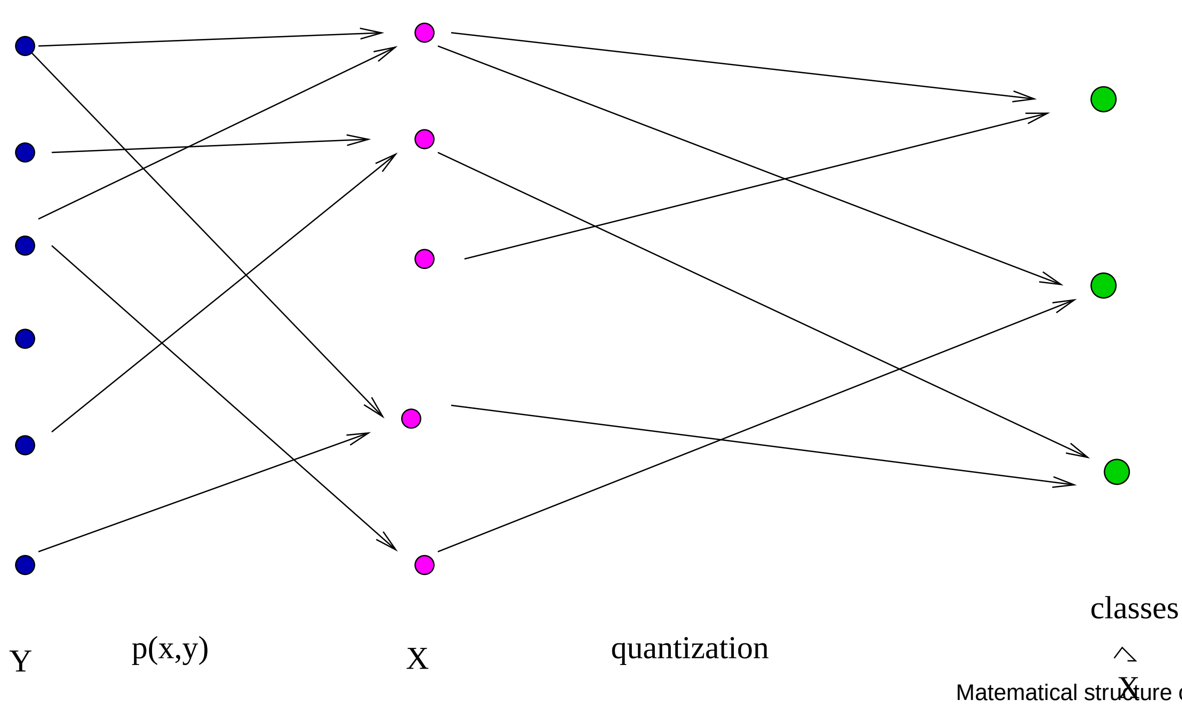
- Deterministic Annealing.

$$maxH(\hat{X}|X) - \beta D(X, \hat{X})$$

Information distortion function

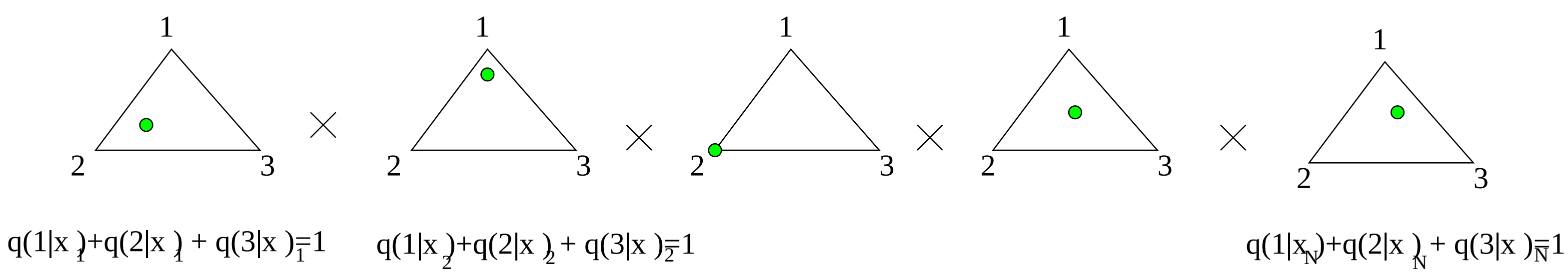
Concentrate on IB and ID: same distortion function $I(\hat{X}, Y)$.

- Information Bottleneck:
  $max_{q \in \Delta(N)} - I(\hat{X}, X) + \beta I(\hat{X}, Y)$,
- Information Distortion: $max_{q \in \Delta(N)} H(\hat{X}|X) + \beta I(\hat{X}, Y)$.



Optimization space

- correspondence between IB and ID: $-I(\hat{X}, X) = H(\hat{X}|X) - H(\hat{X})$
- In bot cases, maximum over space of conditional probabilities $\Delta(N) = \Pi_{i=1}^{k} \Delta_i^N N$ where Δ_i^N is an N-simplex,



Constrained optimization.

Goal: find solution q at some value of β

- Information Bottleneck: β is finite, represents tradeoff between sparsity of representation and goodness of reproduction.
- Information Distortion: $\beta = \infty$.

Bad news: $max_{q \in \Delta^N} I(\hat{X}, Y)$ is NP-complete for all $N \geq 2$ ($\beta = \infty$ problem).
Good news: $max_{q \in \Delta^N} H(\hat{X}|X)$ has unique solution $q = 1/N$ ($\beta = 0$ problem).
Solution: Annealing (maybe Deterministic Annealing?)!

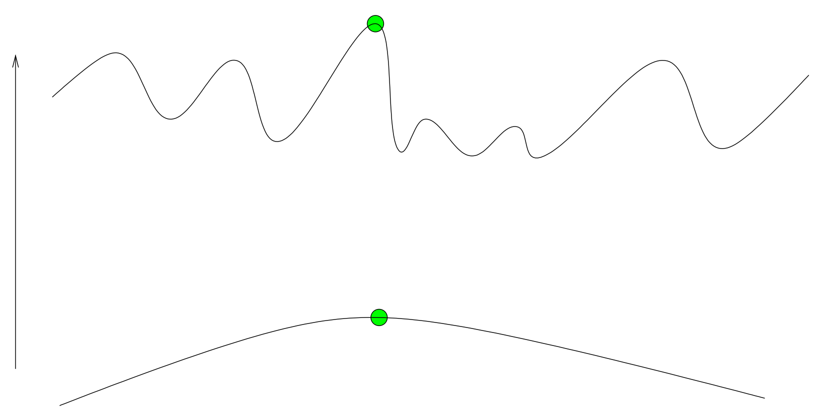
Annealing

Annealing/homotopy idea:

$$max \quad H(\hat{X}|X) + \beta I(X, \hat{X})$$

$$max \quad -I(\hat{X}, X)) + \beta I(X, \hat{X})$$

- Start at $(q, \beta) = (1/N, 0)$, continue this solution in β until $\beta =$ target



Problem: Does this find global maximum at $\beta = \beta^*$?.

Annealing IB

Degeneracy: Initial problem

$$max_{q \in \Delta(N)} - I(\hat{X}, X)$$

has $N-1$ dimensional space of solutions:

$$I(\hat{X}, X) = \sum_{\hat{x},x} q(\hat{x}|x)p(x) \log \frac{q(\hat{x}|x)p(x)}{p(\hat{x})p(x)}$$

Take $q(\hat{x}|x) = p(\hat{x}) = a(\hat{x})$ with $\sum_{\hat{x}} a(\hat{x}) = 1$.
Then $I(\hat{X}, X) = 0$.

Solution: Start wit $N = 2$ and increase N at phase transitions.

Dealing with annealing

Let

$$G(q,\beta) \;:=\; H(\hat{X}|X) + \beta I(\hat{X},Y) \text{ or}$$
$$G(q,\beta) \;:=\; -I(\hat{X},X) + \beta I(\hat{X},Y)$$

Problem:

$$max_{q\in\Delta(N)} G(q,\beta)$$

- Numerical methods.
- Phase transitions: where and what direction.
- What is being computed at phase transitions?

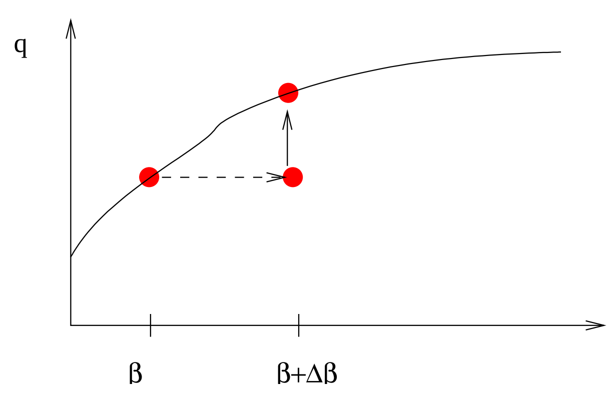
Numerical methods

Basic method:

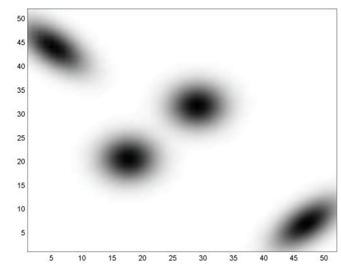
- Increase β by $\Delta\beta$
- Perturb q and **find** solution for new β.

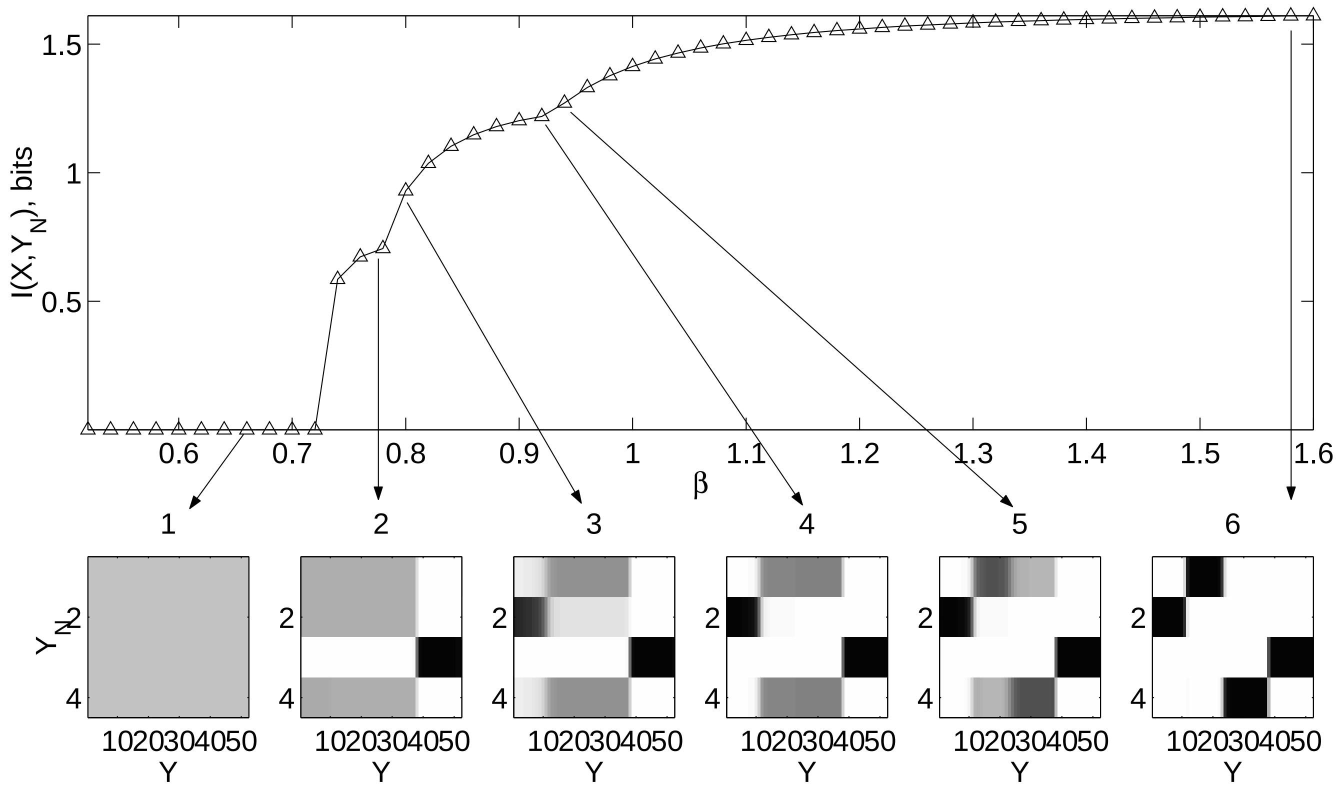
Find can use different methods:

- Blahut-Arimoto type iteration (Tishby et al.)
- Fixed point iteration (Dimitrov et al.)
- Both find only local maxima, no saddle points.



Basic method





Agglomerative Bottleneck

Agglomerative Bottleneck (Slonim, Tishby 1999):

Start at $\beta = \infty$ and decrease β.
However, problem at $\beta = \infty$ is NP-complete.

Dynamical system problem

Since the problem is constrained, we need to consider Lagrangian

$$L(q,\lambda,\beta) = G(q,\beta) + \lambda_x\left(\sum_{\hat{x}} q(\hat{x}|x) - 1\right)$$

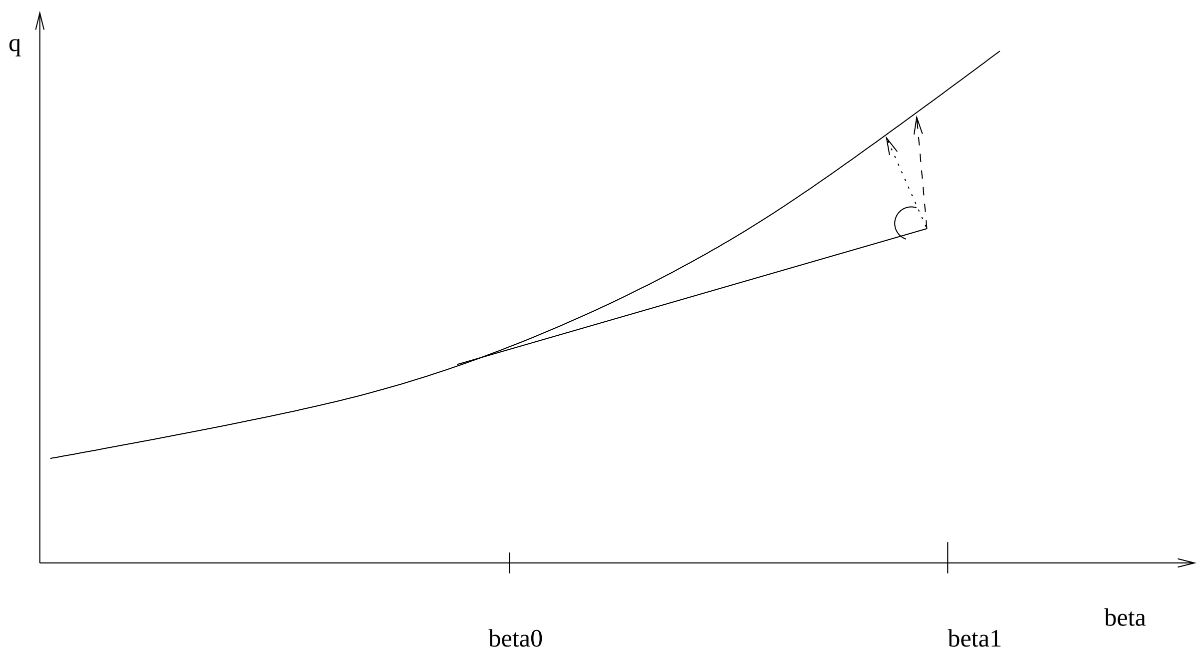
Local maxima are equilibria of the flow

$$\begin{pmatrix} \dot{q} \\ \dot{\lambda} \end{pmatrix} = \nabla_{q,\lambda} L(q,\lambda,\beta)$$

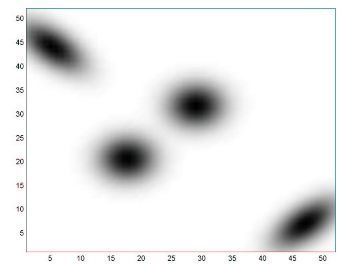
Bifurcation happens at (q^*,λ^*,β^*) if the Hessian ΔL is singular.

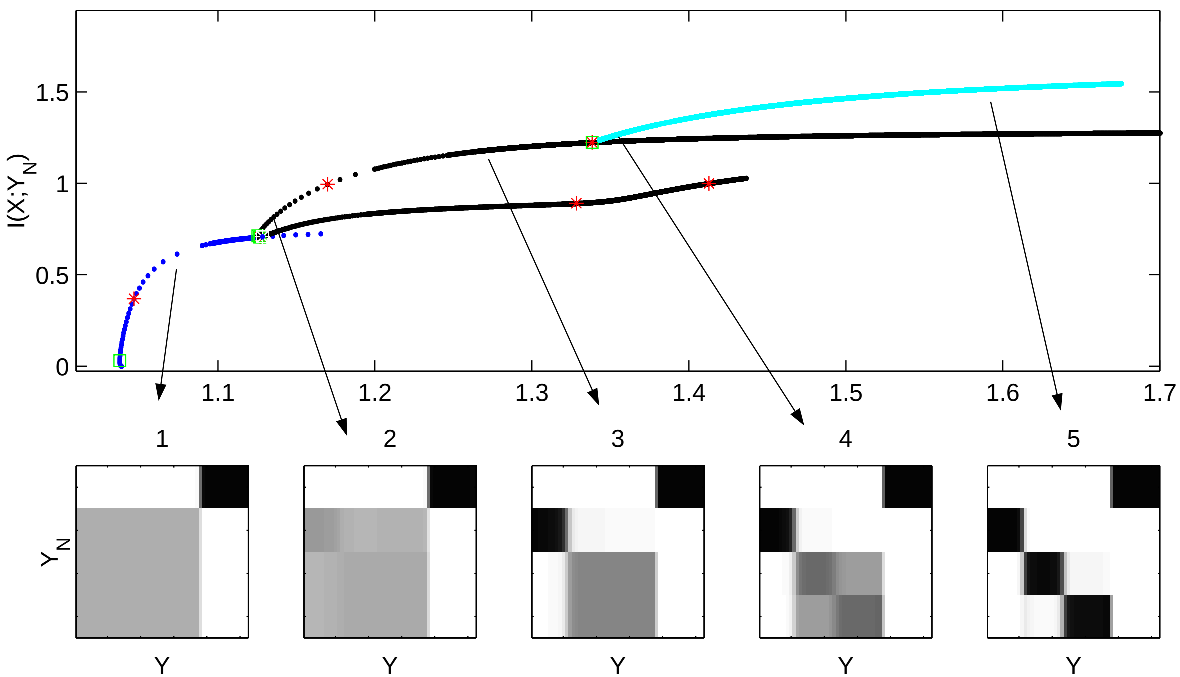
More sophisticated numerics

Not faster numerics!
Continuation algorithm for $L(q, \lambda, \beta)$ - using Newton iteration
- can find unstable solutions.



Numerics using continuation



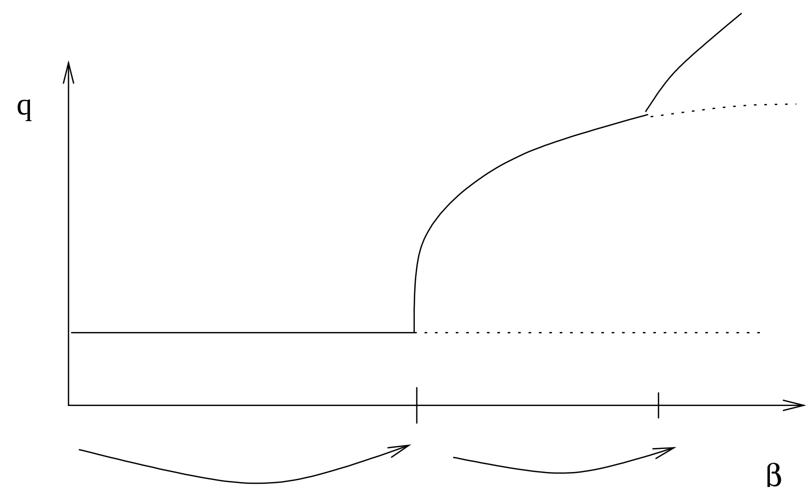


Phase transitions

Can we compute them "ahead of time"?

Then we can jump to phase transition directly and resolve phase transition.
Yes, we can, for $q = 1/N$.

This is analogous to Deterministic Annealing for Euclidean distortion (Rose 1998)



Deterministic annealing

Phase transitions - zero eigenvalues of ΔL, eigenvector - direction of the split.

$$\Delta L = \begin{pmatrix} \Delta G & J^T \\ J & 0 \end{pmatrix}$$

where J consists of N identity matrices. At $q = 1/N$:

$$\Delta G = \begin{pmatrix} B & 0 & \dots & 0 \\ 0 & B & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & B \end{pmatrix}$$

Symmetry: relabeling of the elements $\hat{x}$.

Deterministic annealing

Phase transition values β at $q = 1/N$ corresponds to existence of a null eigenvector v of block B

$$Bv = (\Delta H + \beta \Delta I)v = 0$$

Rewritten, this is

$$(\Delta H)^{-1} \Delta I v = \frac{1}{\beta} v$$

- First phase transition value $\beta \leftrightarrow$ largest positive eigenvalue of $(\Delta H)^{-1} \Delta I$

Computing phase transitions

- Matrix
$$M := (\Delta H)^{-1}\Delta I$$
has the form
$$M = Q - A$$
- Q^T is stochastic
- M has eigenvalue $1/\beta = 0$ with eigenvector $(1, 1, 1, \ldots, 1)$ - not interesting !
- All other eigenvalues of M are eigenvalues of Q
- Q^T is stochastic implies largest eigenvalue of Q is ≤ 1
- $1/N$ looses stability at $\beta \geq 1$

Phase transitions for IB

Degeneracy problem again:

- for IB ΔG has for all β and all q $N-1$ dimensional null space.
- Phase transition - dimension of null space $\geq N$.

Phase transitions for IB

Instead of

$$(Q - A)v = 1/\beta v$$

we get

$$(Q - A)v = (I - A)1/\beta v$$

Same result:

- has solution $1/\beta = 0$ with eigenvector $(1, 1, 1, \ldots, 1)$ - not interesting !
- All other solutions are eigenvalues of Q

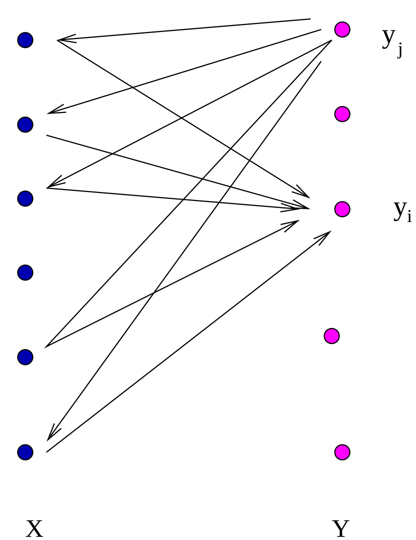
Bottom line: Bifurcations for IB and ID at $q = 1/N$ happen at the same values of β and in the same direction.

Phase transitions

Phase transitions at $q = 1/N \Leftrightarrow$ eigenvalues of stochastic matrix Q^T

The matrix Q^T is a transition matrix for a graph G:

- Vertices are patterns y_i
- edge $y_j \to y_k$ has weight $\sum_i p(y_k|x_i)p(x_i|y_j)$



Digression-Normal cut

Given a graph G with weights $w(a,b)$ divide into 2 groups A and B so that

$$\frac{cut(A,B)}{assoc(A,G)} + \frac{cut(A,B)}{assoc(B,G)}$$

is minimized

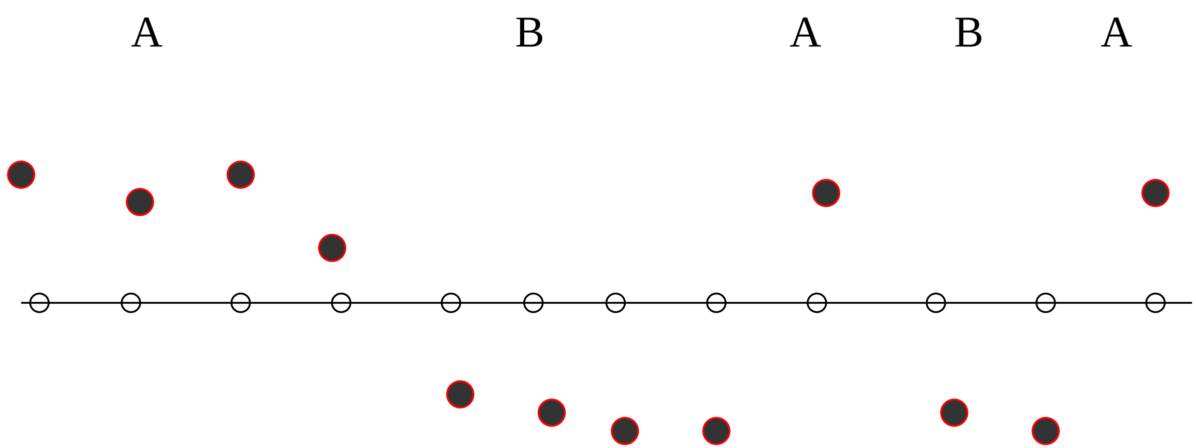
- $cut(A,B) = \sum_{a \in A, b \in B} w(a,b)$
- $assoc(A,G) = \sum_{a \in A, e \in G} w(a,e)$
- Finding N-cut is NP-complete problem.

Approximate Normal Cut

Approximate Normalized cut (Shi and Malik (2000))
Find second smallest eigenvalue of

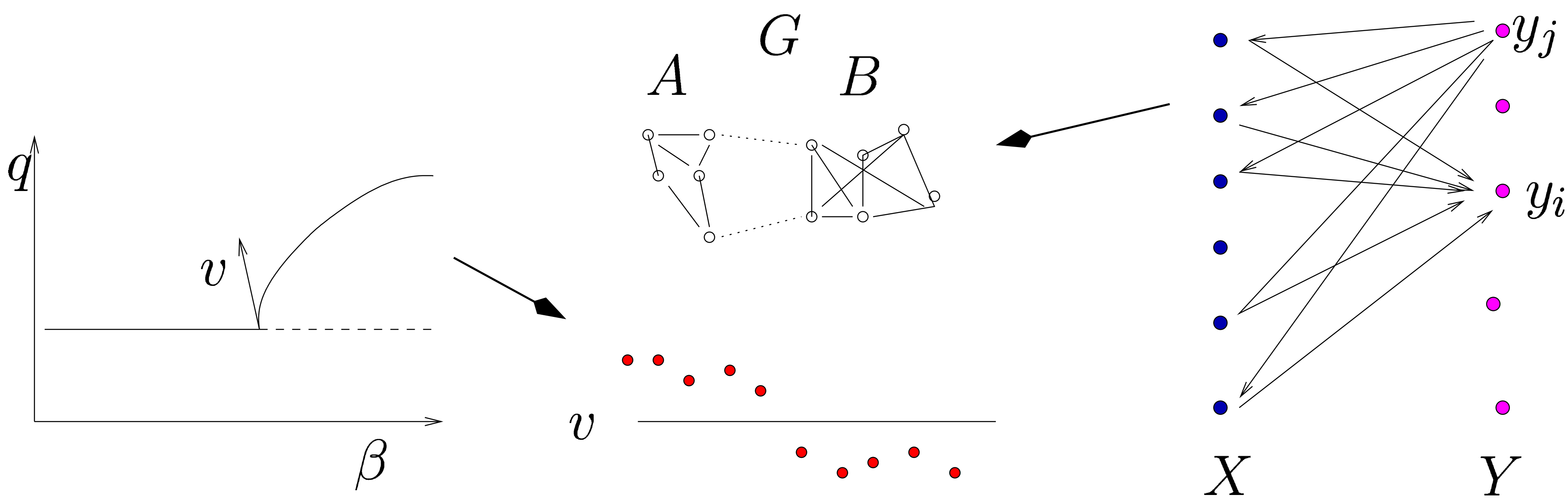
$$(D - W)y = \lambda Dy.$$

After y is computed, Approximate Normalized Cut is **If** $y_i > 0$**,** $i \in A$**, if** $y_i \leq 0$**, then** $i \in B$



Correspondence

- Bifurcation direction v at first bifurcation at $q = 1/2$ computes Approximate Normal cut for the graph G:
- Vertices V correspond to the set of patterns Y;
- Weight $w(y_i, y_j) = \sum_i p(y_i|x_i)p(x_i|y_j)$



Correspondence

Take $|\hat{X}| = 2$ (two classes) After bifurcation, the probability of x to belong to

- class A: $1/2 + \epsilon v_i$;
- class B: $1/2 - \epsilon v_i$,

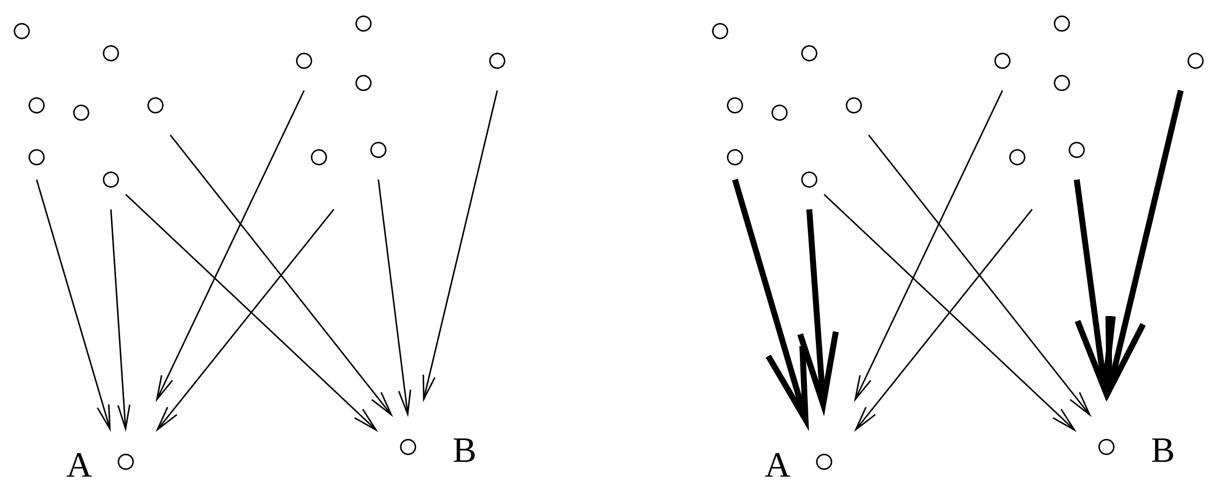
v is bifurcating direction ("soft push")

Correspondence

Take $|\hat{X}| = 2$ (two classes) After bifurcation, the probability of x to belong to

- class A: $1/2 + \epsilon v_i$;
- class B: $1/2 - \epsilon v_i$,

v is bifurcating direction ("soft push")

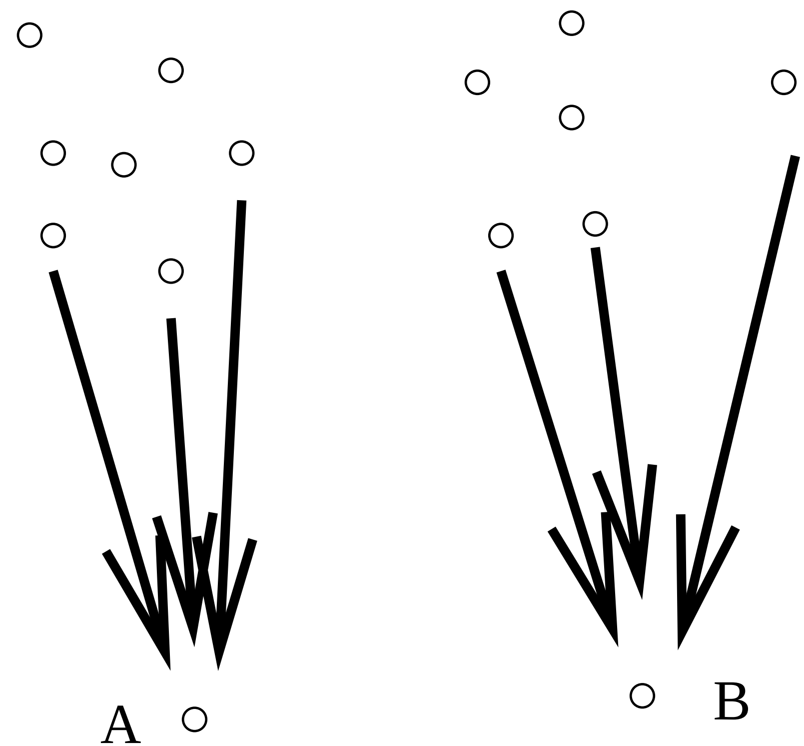


Correspondence

It would be nice if, as $\beta \to \infty$ probabilities converge to 0 or 1 ="hard clusters" of the N-cut.

Correspondence

It would be nice if, as $\beta \to \infty$ probabilities converge to 0 or 1 ="hard clusters" of the N-cut.



Correspondence

TRUE, for slightly different cost function: replace $H(q) + \beta I(q)$ by $H(q) + \beta U(q)$

$$I(X, Y_N) = \sum_{x,\mu} p(x,\mu) \log(\frac{p(x,\mu)}{p(x)p(\mu)})$$

$$U(X, Y_N) = \sum_{x,\mu} p(x,\mu)(\frac{p(x,\mu)}{p(x)p(\mu)} - 1).$$

- Bifurcation direction v at first bifurcation at $q = 1/2$ computes Approximate Normal cut for the graph G':
- Weight $w(y_i, y_j) = \sum_i p(y_i|x_i)p(x_i, y_j)$
- As $\beta \to \infty$ solution converges to Normal Cut of G'.

Summary

- There are similarities and differences between Information Bottleneck, Information Distortion, Rate distortion theory and Deterministic Annealing.
- We reviewed numerical methods used to solve IF and ID: Basic algorithm, agglomerative bottleneck and continuation.
- $max I(\hat{X}, Y)$ is NP-complete
- Phase transitions can be explicitely computed for $q = 1/N$.
- First phase transition computes an Approximate Normal Cut of a certain graph.

Questions

Mathematics

- Global stability of branches
- Extensions to X, Y continuous random variables, multivariate bottleneck.

Computer Science

- Given a graph $G = (Y, E)$, is there a random variable X and a probability distribution $p(X, Y)$ such that annealing $H + \beta\hat{X}$ will compute both Approximate N-cut and N-cut of G?

Neuroscience:

- Use Information distortion as a tool to compare different models of sensory systems (cricket sensory system).