Source allocation and estimation with incomplete data

Tommi S. Jaakkola
MIT AI Lab
tommi@ai.mit.edu

joint work with Adrian Corduneanu

Source allocation problems

- Estimation problems that involve balancing one (preferred) source of information with another
- Example settings:
  - how to weigh prior model relative to the available data (i.e., the equivalent sample size)
  - model estimation from multiple possibly heterogeneous data sources (e.g., estimation with scarce complete and unlimited incomplete data sources)

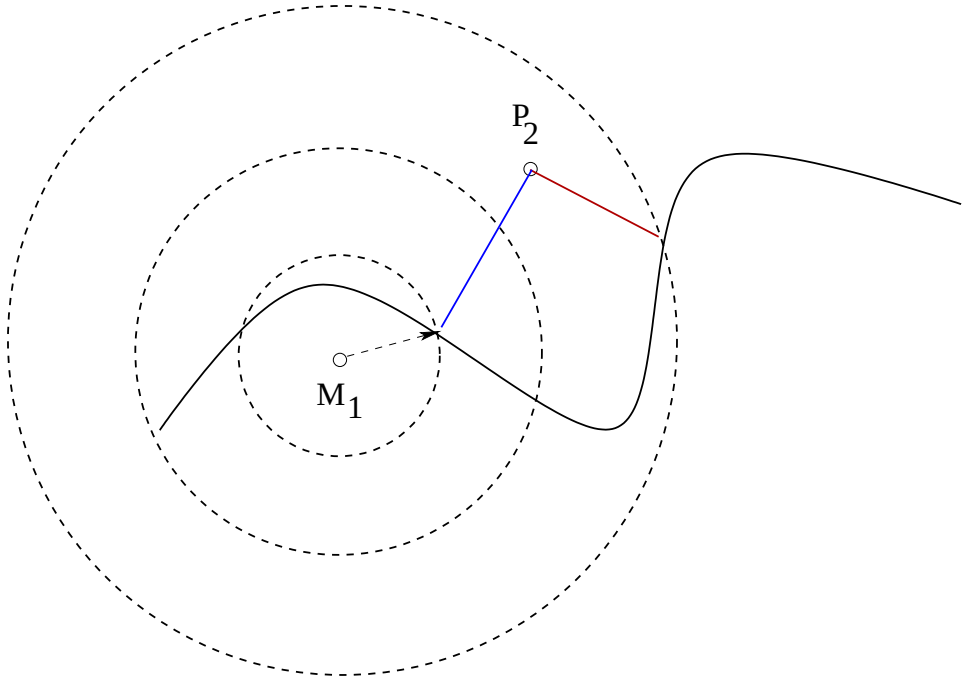
    etc.

Multiple realizations, single approach

- Many "source allocation problems" can be solved essentially in the same manner
- Key concepts: **stability** and **continuation**
- Example realizations:
  - text classification with predominantly unlabeled data
  - competitive (game theoretic) estimation of sequence motifs

Why is the problem interesting?

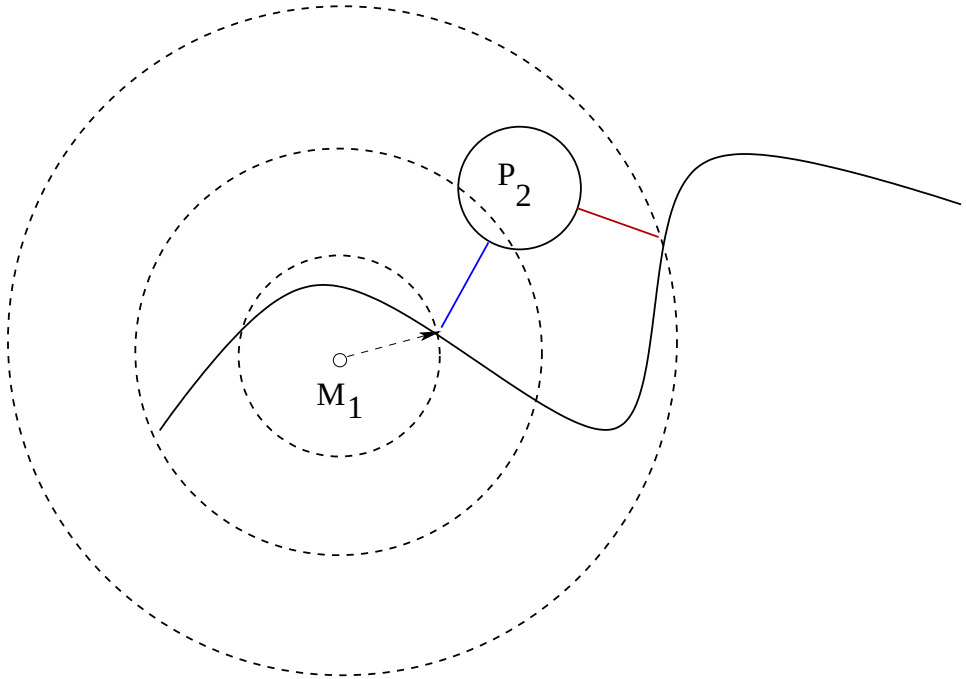
- While the estimation objective changes smoothly with varying allocation, the optimizing argument might not
  Example: two datasets (1 and 2), model M



$M_1 =$ model estimated from only dataset 1.
$P_2 =$ target distribution implied by dataset 2

Example cont'd

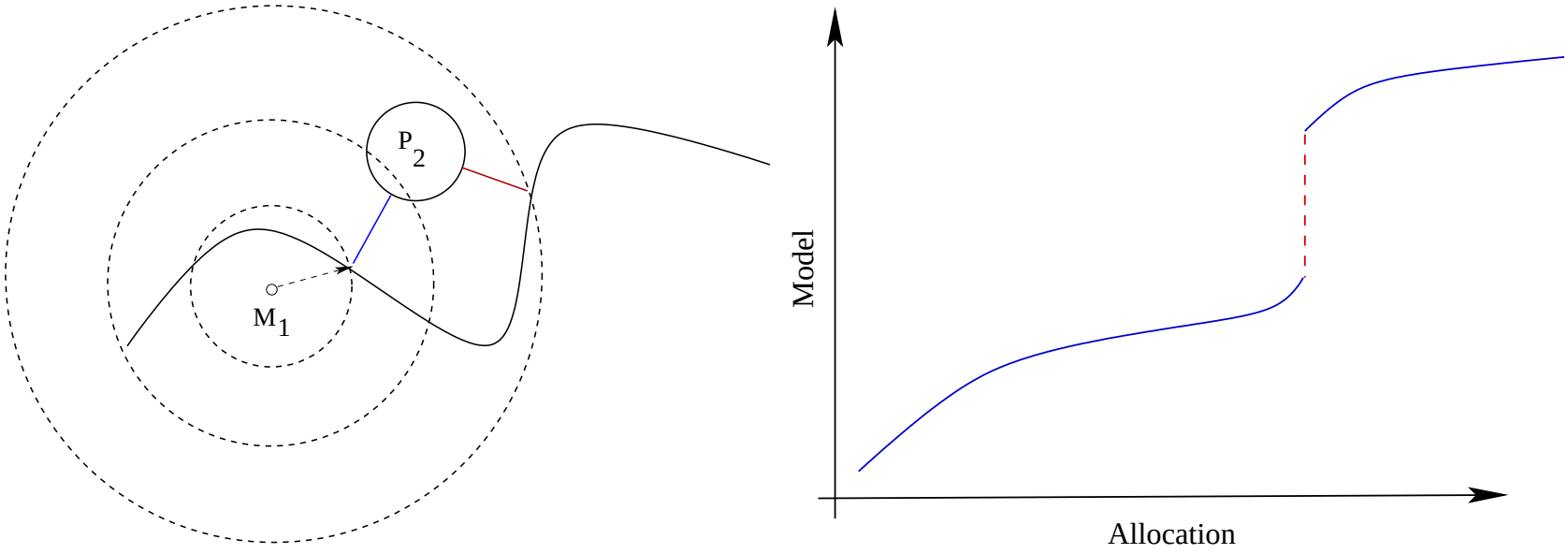
- dataset 2 is now incomplete



$M_1 =$ model estimated solely from dataset 1.

$P_2 =$ target from dataset 2 not unique

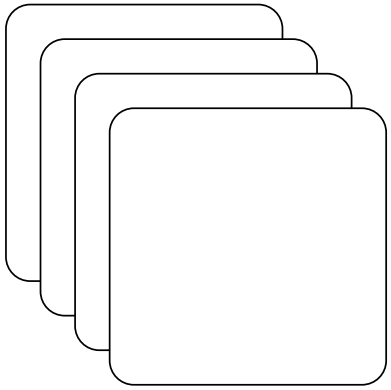
Discontinuities in estimation



- The abrupt changes will almost surely manifest themselves as discontinuities (not as bifurcations)

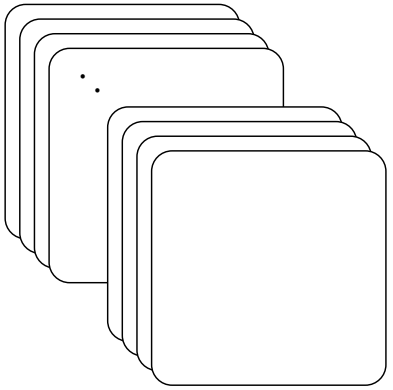
Incomplete data estimation

- E.g., document classification



few labeled examples

$\{x^t, y^t\}$, $P_l(x, y)$



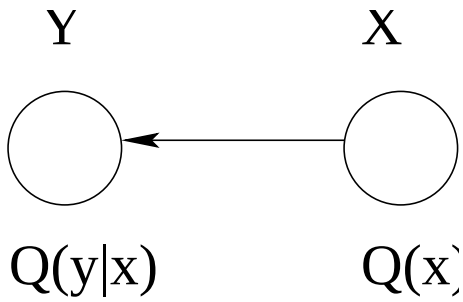
many unlabeled examples

$\{x^t\}$, $P_u(x)$

- Considerations:
  1. model structure $Q(x, y)$
  2. estimation criterion (e.g., maximum likelihood)
  3. allocation and stability

Model structure

- No constraints, no benefit

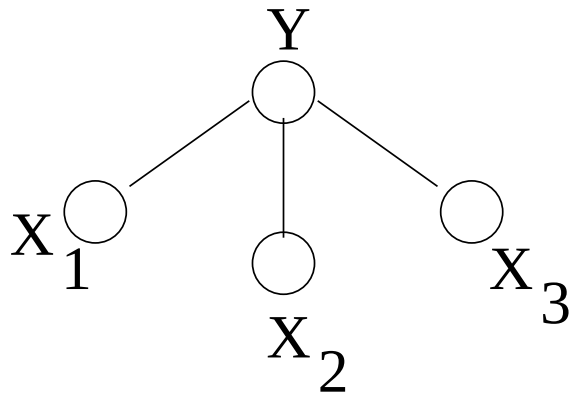


  With the maximum likelihood estimation criterion, $Q(y|x)$ and $Q(x)$ must share parameters for the unlabeled examples to help improve (affect) classification accuracy

- The fundamental estimation issues (e.g., stability) can be understood in the context of the simplest models

Naive Bayes model

- Naive Bayes model is often used in the context of document and other classification problems



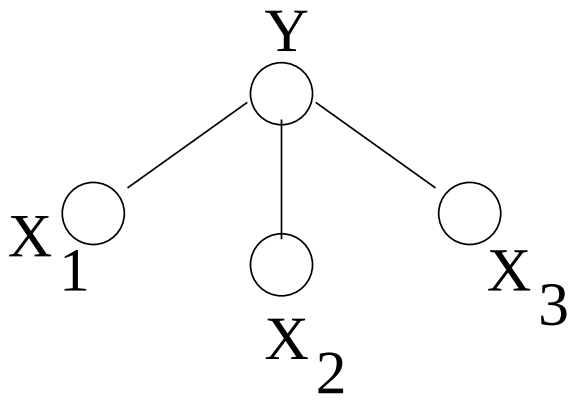
$$Q(x,y) = \left[\prod_{i=1}^{k} Q_i(x_i,y)\right] Q(y)^{1-k}$$

- The model is fully specified by pairwise marginals

$$Q_i(x_i,y), \quad i = 1,\ldots,k.$$

(this is an overcomplete parameterization)

Estimation

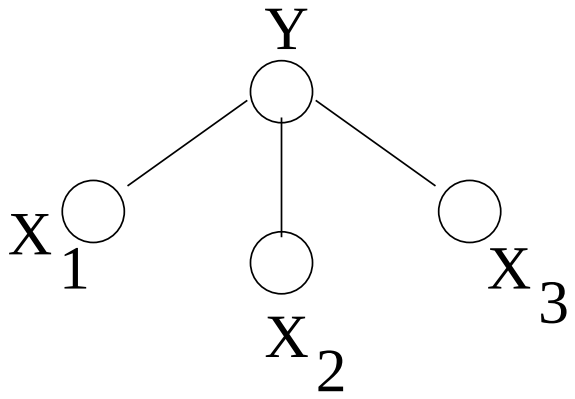


Labeled data log-likelihood: $D(P_l(x,y)\|Q(x,y))$
Unlabeled data log-likelihood: $D(P_u(x)\|Q(x))$

- Equivalent (weighted) maximum likelihood formulations

1. $\min_Q \left\{ D(P_u(x)\|Q(x)) \right\}$ s.t. $D(P_l(x,y)\|Q(x,y)) \leq d$
2. $\min_Q \left\{ (1-\lambda) D(P_l(x,y)\|Q(x,y)) + \lambda D(P_u(x)\|Q(x)) \right\}$

Estimation cont'd



- Such estimation problems reduce to solving fixed point equations (here EM)

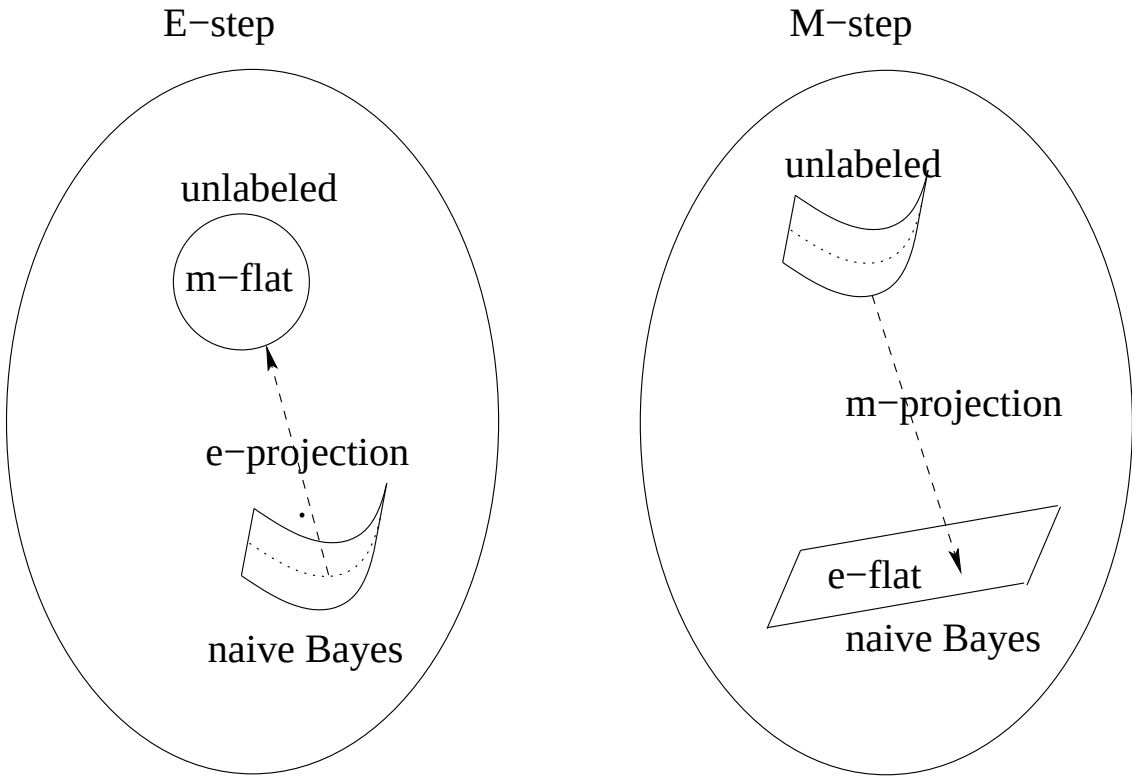
  E-step: $P(x,y) \leftarrow (1-\lambda)P_l(x,y) + \lambda Q(y|x)P_u(x)$
  M-step: $Q_i(x_i,y) \leftarrow \sum_{x \setminus x_i} P(x,y)$

  where $\lambda \in [0,1]$ is the allocation parameter.

  (we use the shorthand $Q = EM_\lambda(Q)$ for the fixed point equations)

Geometry of the EM algorithm

- Geometry for $\lambda = 1$ (unlabeled only)

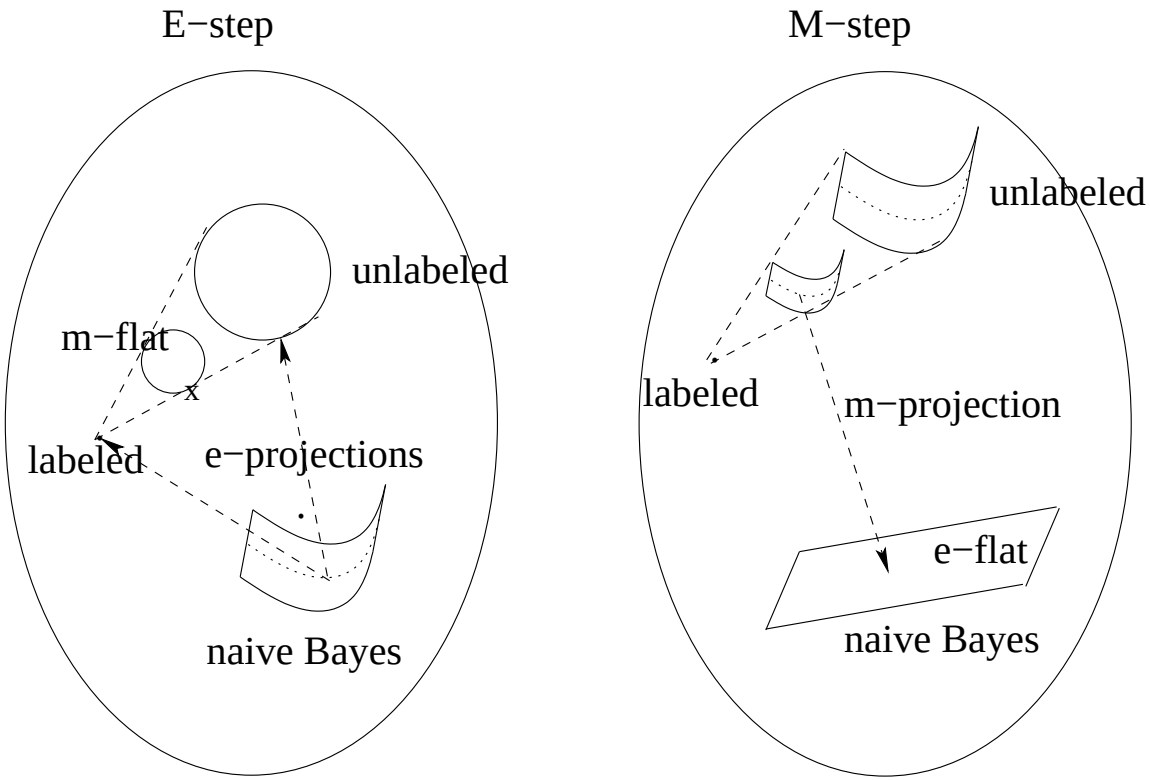


E-step: $P(x, y) \leftarrow Q(y|x)P_u(x)$

M-step: $Q_i(x_i, y) \leftarrow \sum_{x \setminus x_i} P(x, y)$

Geometry of the EM algorithm cont'd

- Geometry for $0 < \lambda < 1$



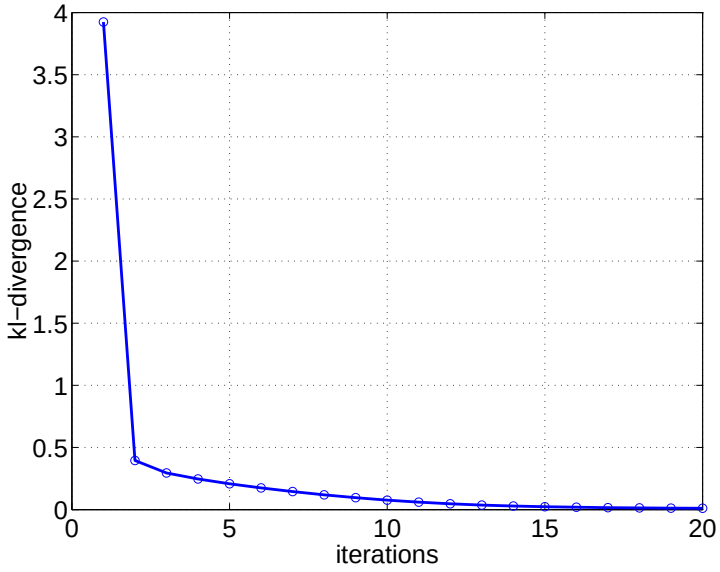
- A "geometrically correct" version of the EM algorithm in this case is the em-algorithm (Amari, 1995)

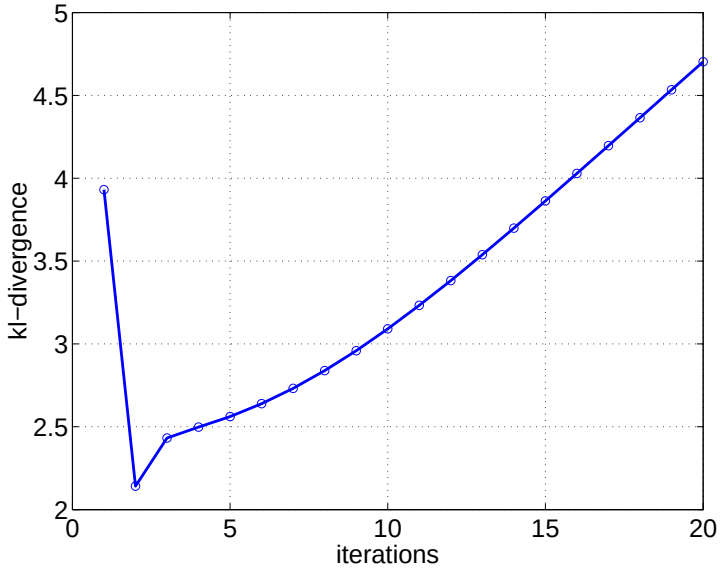
Fixed points

- The main problem with the EM-algorithm is that we have little control over the choice of the many possible fixed points

$$Q_\lambda^i = EM_\lambda(Q_\lambda^i),\ i = 1, \ldots, L$$

- Some of the fixed points are good solutions, others may be terrible



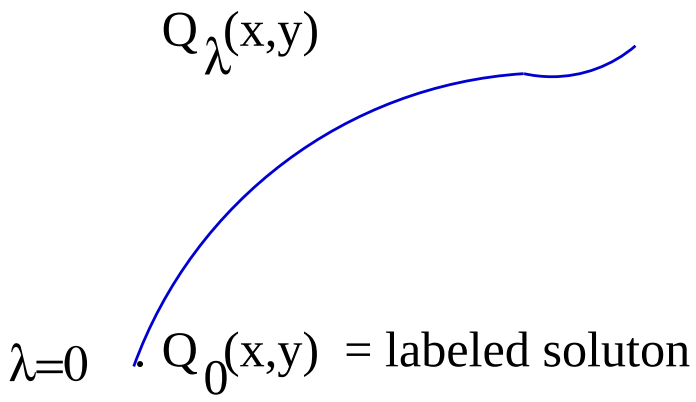


Evolution of fixed points

- Instead of finding a single fixed point, we can trace a continuous path of fixed points starting from the labeled solution

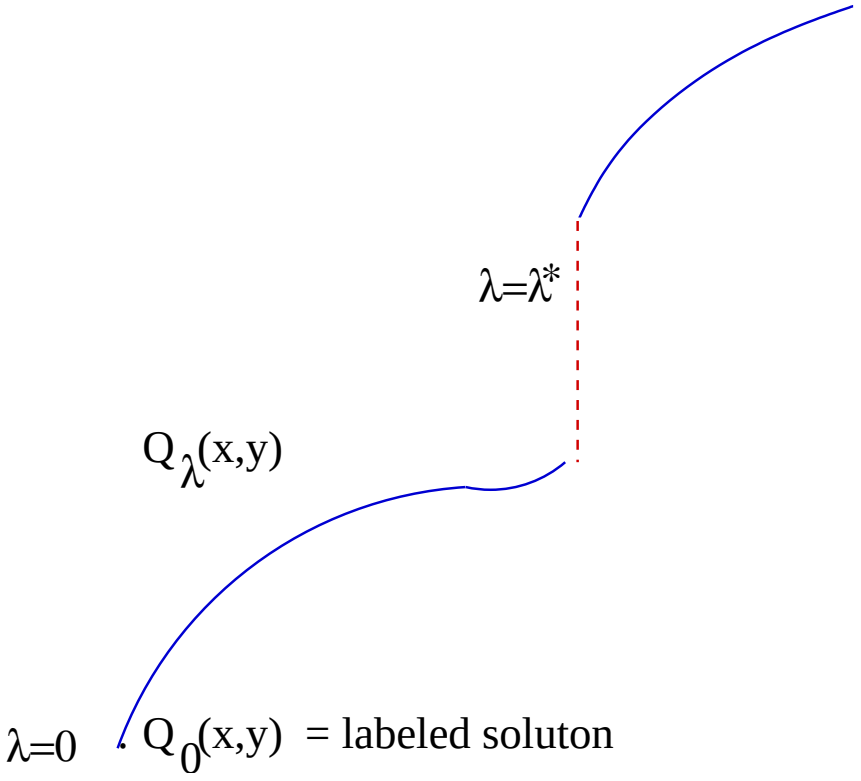
$$Q_\lambda = EM_\lambda(Q_\lambda), \quad \lambda \in [0, 1)$$

- Each fixed point in this curve is firmly rooted in the maximum likelihood solution based on the labeled examples alone



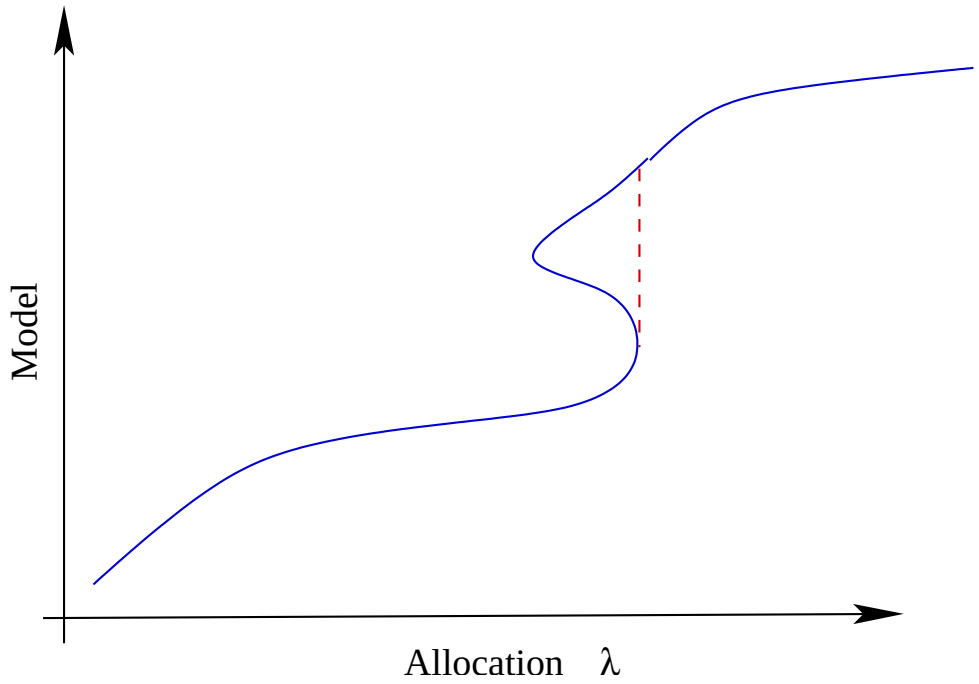
Evolution of fixed points cont'd

- We can explicitly identify any critical points where the solution changes dramatically



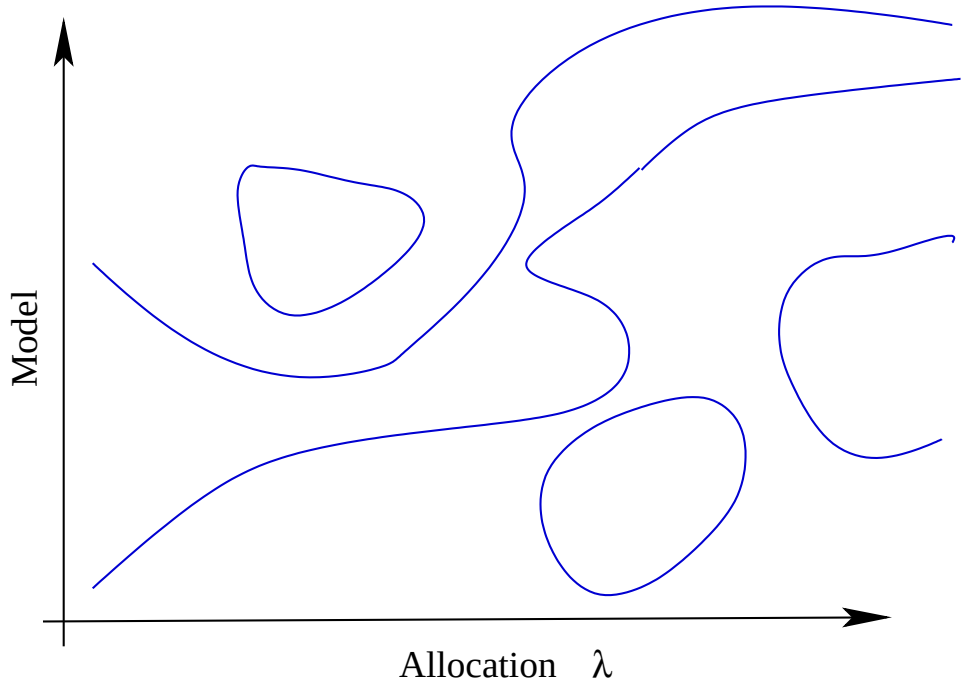
Evolution of fixed points cont'd

- We can actually find a smooth curve in the extended space (Q, λ)



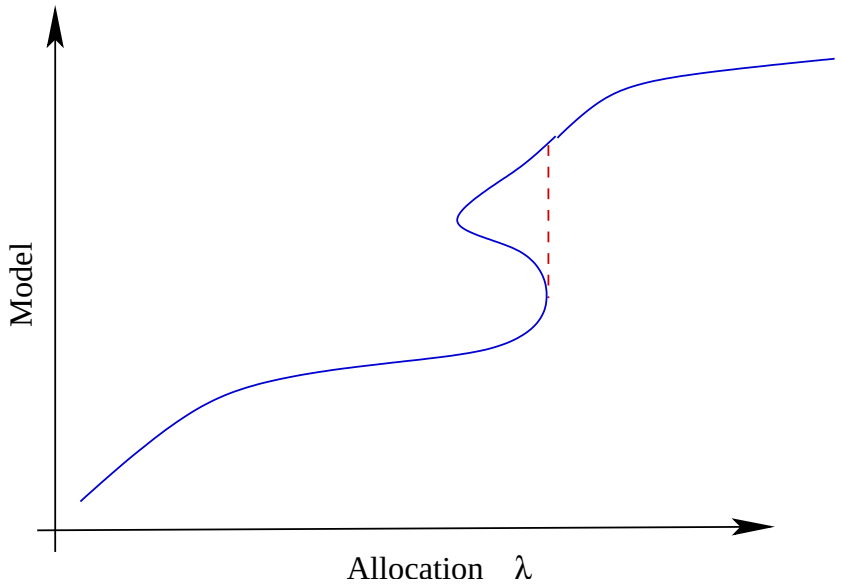
Theorem: Provided that the Jacobian of $T(Q, \lambda) = EM_\lambda(Q) - Q$ has full rank, $T(Q, \lambda) = 0$ defines a smooth 1-dim manifold in the extended (Q, λ) space.

Evolution of fixed points cont'd



- The curve starting from $(Q_0, 0)$ is *unique*

Numerically stable evolution



- In the EM case

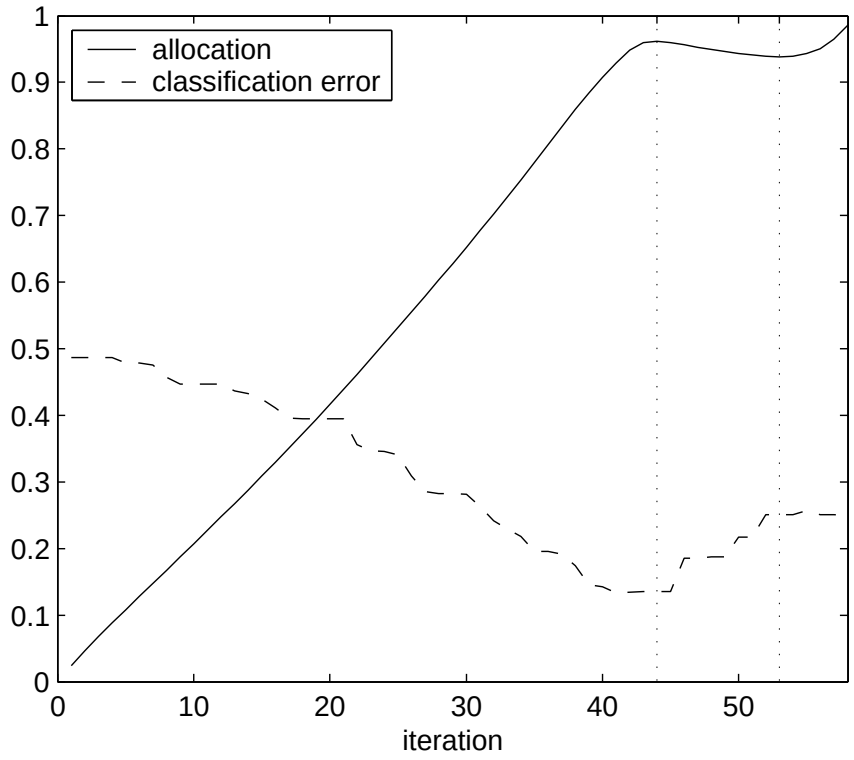
$$\left[I - \lambda \nabla_Q EM_\lambda(Q) \quad \frac{\partial}{\partial \lambda} EM_\lambda(Q) \right] \begin{bmatrix} dQ/ds \\ d\lambda/ds \end{bmatrix} = 0$$

  (s parameterizes the path)

- Each point along the curve necessarily satisfies the fixed point condition $Q = EM_\lambda(Q)$

Example results

- By identifying critical points, we can achieve nearly optimal allocation of unlabeled/labeled examples



- The critical point specifies the maximum (rational) allocation of the incomplete data source

Motif discovery

- Sequences X, empirical distribution $P_u(X)$

AGGTTGCAATTTCTTTCGGCGGTCTTTCGTCCGCTAAAAATGGGTCACGTGATCT...

TCAATTTTCCTTTCCTTATTCTACTCTTTTTATCTACCTGAAGATAAAAAACAAC...

TGGAACTTTCAGTAATACATTGCTCGGAAGACTCTCCTCCGCTATTGAAGTACGG...

...

Motif discovery

- Sequences X, empirical distribution $P_u(X)$

  AGGTTGCAATTTCTTTCGGCGGTCTTTCGTCCGCTAAAAATGGGTCACGTGATCT...

  TCAATTTTCCTTTCCTTATTCTACTCTTTTTATCTACCTGAAGATAAAAAACAAC...

  TGGAACTTTCAGTAATACATTGCTCGGAAGACTCTCCTCCGCTATTGAAGTACGG...

  ...

- Background model $P_{B_0}(X)$ (e.g., independent multinomial)

Motif discovery

- Sequences X, empirical distribution $P_u(X)$

  AGGTTGCAATTTCTTTCGGCGGTCTTTCGTCCGCTAAAAATGGGTCACGTGATCT...

  TCAATTTTCCTTTCCTTATTCTACTCTTTTTATCTACCTGAAGATAAAAAACAAC...

  TGGAACTTTCAGTAATACATTGCTCGGAAGACTCTCCTCCGCTATTGAAGTACGG...

  ...

- Background model $P_{B_0}(X)$ (e.g., independent multinomial)
- Simple (ungapped) motif model of length m

$$Q(X_t, \ldots, X_{t+m-1}) = \prod_{i=0}^{m-1} Q_i(X_{t+i})$$

  - prior that a motif is present in any sequence (e.g., 0.5)
  - prior over locations along any sequence (e.g., uniform)

$\Rightarrow$ sequence model $P_{B_0,Q}(X)$

Estimation criterion

- We minimize

$$J(Q;\lambda) = (1-\lambda)D(Q_0\|Q) + \lambda D(P_u\|P_{B_0,Q})$$

  where $D(\cdot\|\cdot)$ is the KL-divergence and λ is the allocation parameter
  - at $\lambda = 0$ we get background
  - at $\lambda = 1$ we get the usual ML criterion

Estimation criterion

- We minimize

$$J(Q;\lambda) = (1-\lambda)D(Q_0\|Q) + \lambda D(P_u\|P_{B_0,Q})$$

  where $D(\cdot\|\cdot)$ is the KL-divergence and λ is the allocation parameter
  - at $\lambda = 0$ we get background
  - at $\lambda = 1$ we get the usual ML criterion

- Fixed point equations (EM for a fixed λ)

$$\nabla_Q J(Q;\lambda) = 0$$

Estimation criterion

- We minimize

$$J(Q;\lambda) = (1-\lambda)D(Q_0\|Q) + \lambda D(P_u\|P_{B_0,Q})$$

  where $D(\cdot\|\cdot)$ is the KL-divergence and λ is the allocation parameter
  - at $\lambda = 0$ we get background
  - at $\lambda = 1$ we get the usual ML criterion

- Fixed point equations (EM for a fixed λ)

$$\nabla_Q J(Q;\lambda) = 0$$

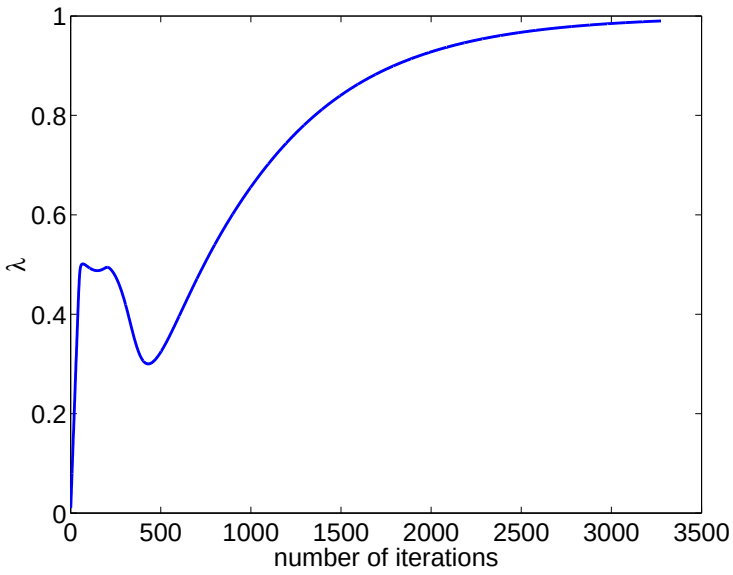
- Evolution of fixed points

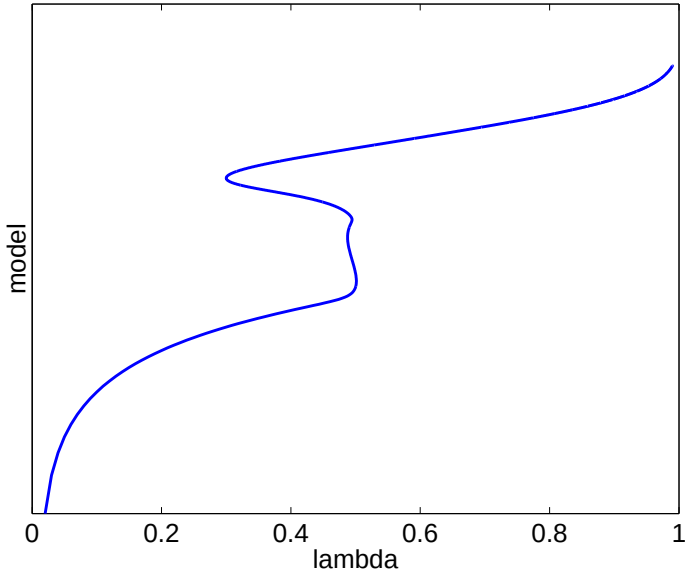
$$\left[\nabla_Q^2 J(Q;\lambda) \quad \frac{\partial}{\partial\lambda}\nabla_Q J(Q;\lambda)\right] \begin{bmatrix} dQ/ds \\ d\lambda/ds \end{bmatrix} = 0$$

  (s parameterizes the path)

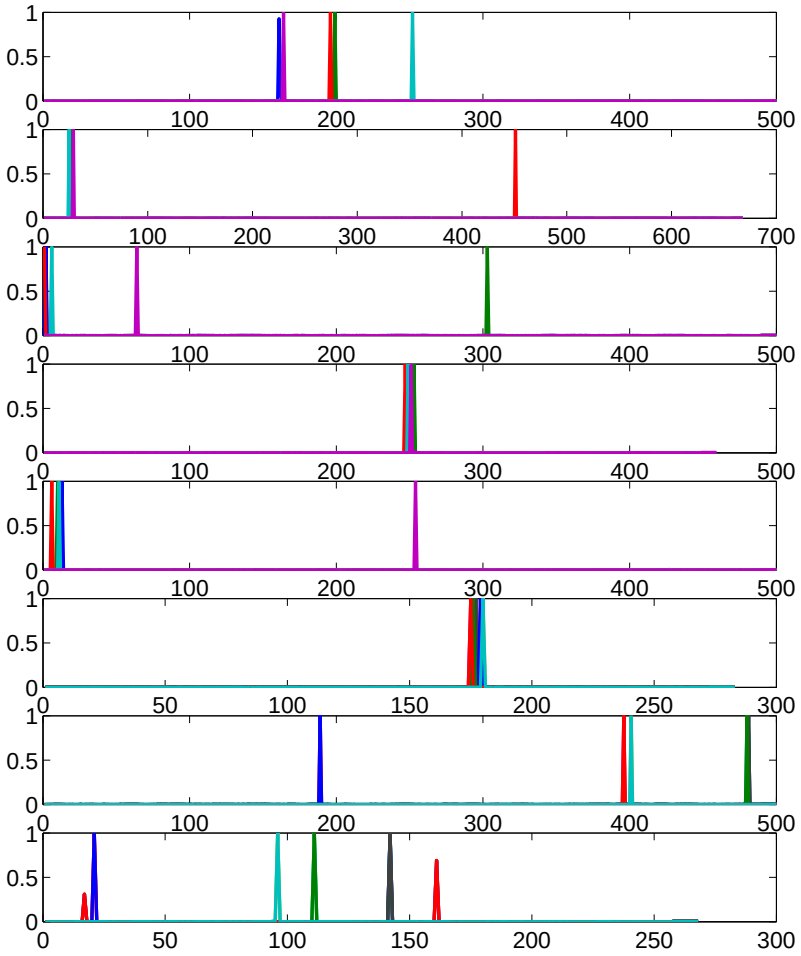
Critical points

- Critical points are typical and desirable in this context. For example, for GAL4 motif:





Iterative deterministic search



Competitive (game theoretic) formulation

- What is an appropriate background model?
- We instead find motifs that can compete against the worst case background, i.e., we find the *minimax* solution

$$\min_{Q} \max_{B} \Big\{ \; (1-\lambda) D(Q_0 \| Q) + \lambda D(P_u \| P_{B,Q}) \\ -(1-\lambda) D(B_0 \| B) - \lambda D(P_u \| P_B) \; \Big\}$$

  - solution characterized by fixed points
  - continuation applies as before
  - critical points arise as before

Summary (estimation)

- "Source allocation" problems are ubiquitous
- The abstract problem structure permits a general solution through (homotopy) continuation methods
- Any tractable model family can be used with continuation
- Work in progress:
  - critical points and estimation quality
  - experiment design based on stability
  - solving equilibria of multi-person games