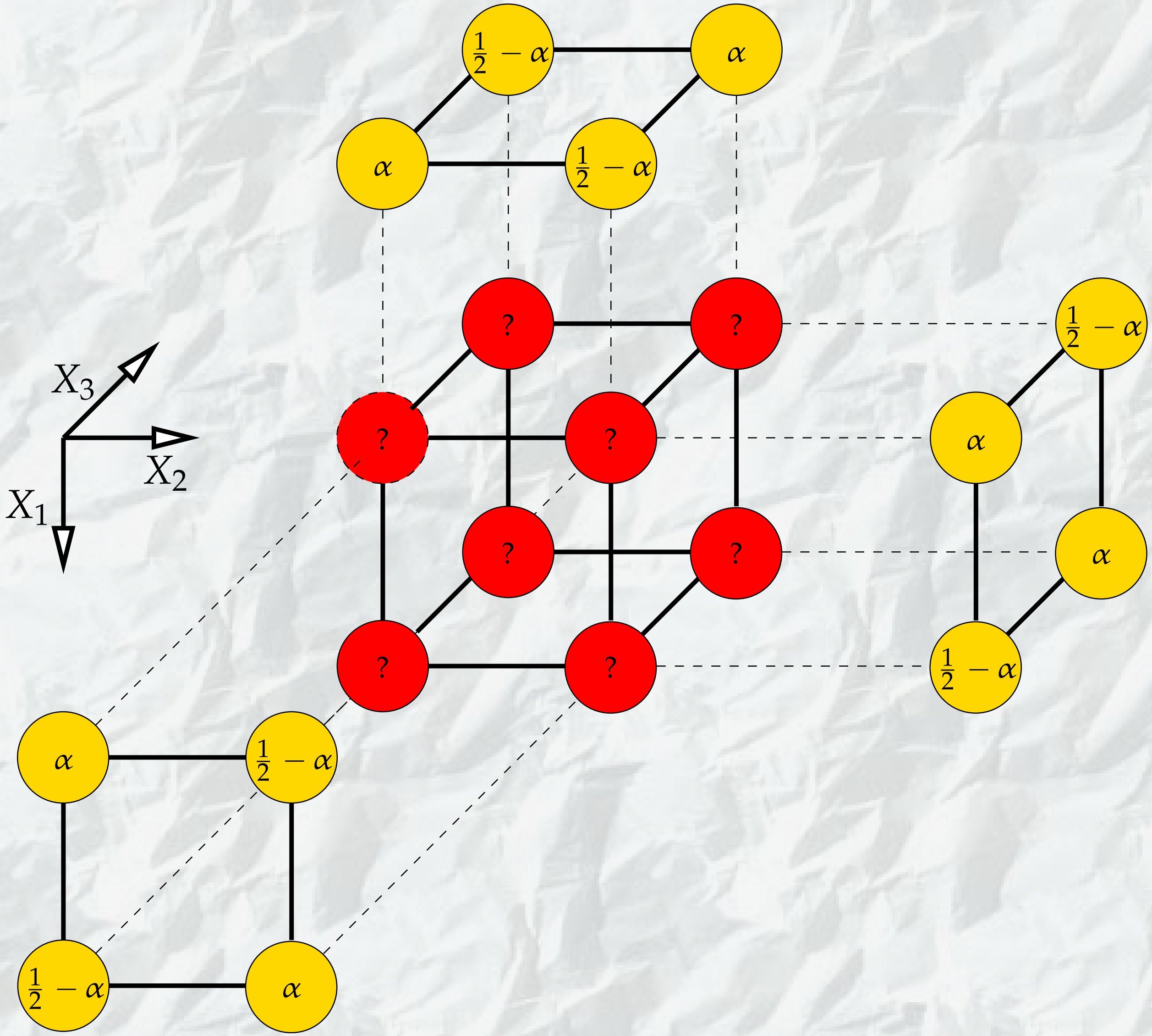
Integrating inconsistent data in a probabilistic model

Jiří Vomlel

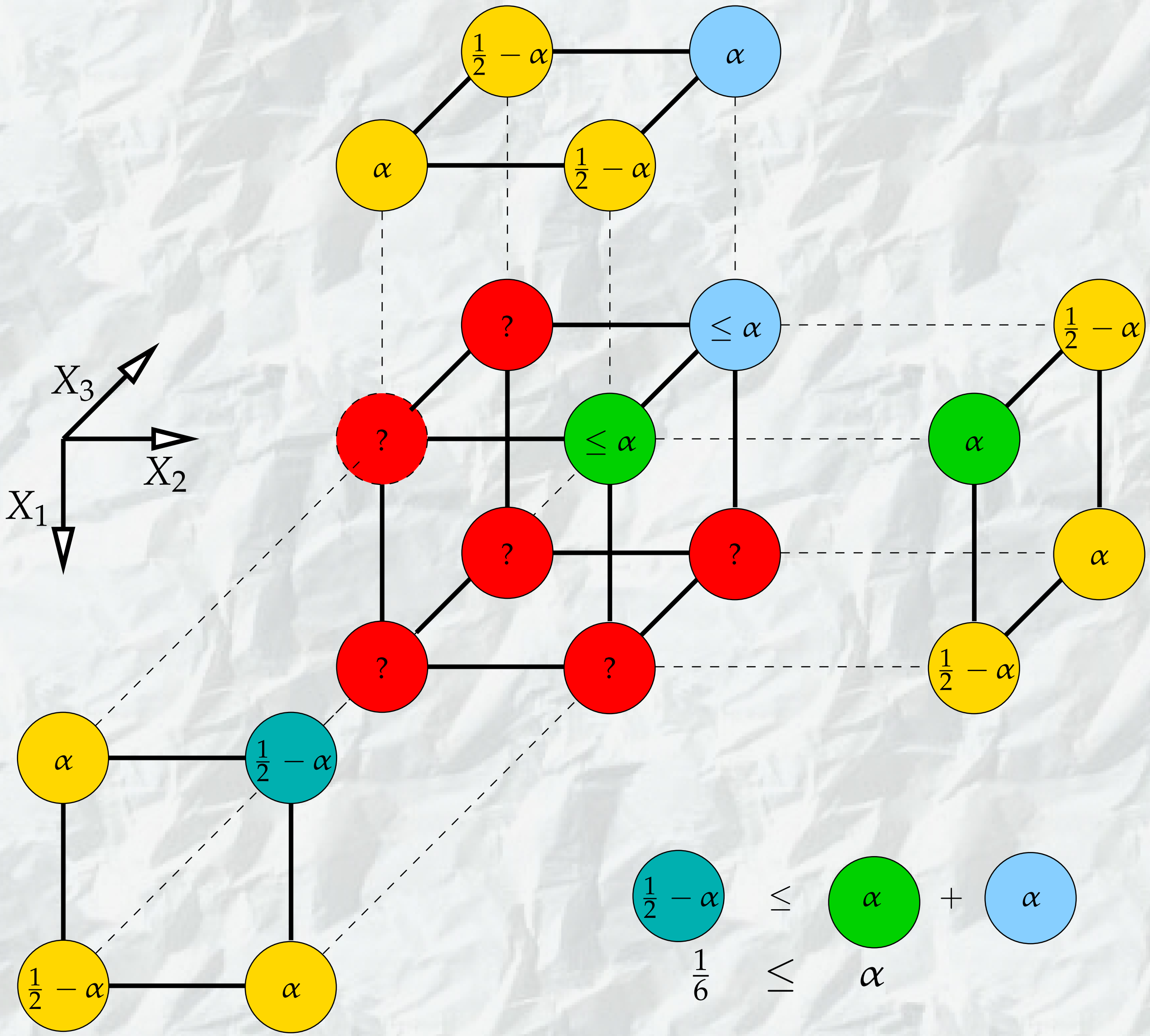
This presentation is available at
http://www.utia.cas.cz/vomlel/

Knowledge integration

- Discrete random variables X_i indexed by natural numbers from $V = \{1, \ldots, n\} \subset \mathbb{N}$.
- Low-dimensional probability distributions P_j, $j = 1, \ldots, k$ defined on variables $\{X_\ell\}_{\ell \in E_j}, E_j \subseteq V$.
- *Knowledge integration* is the process of building a joint probability distribution $Q(X_1, \ldots, X_n)$ from a set of low-dimensional probability distributions $\mathcal{P} = \{P_1, \ldots, P_k\}$.



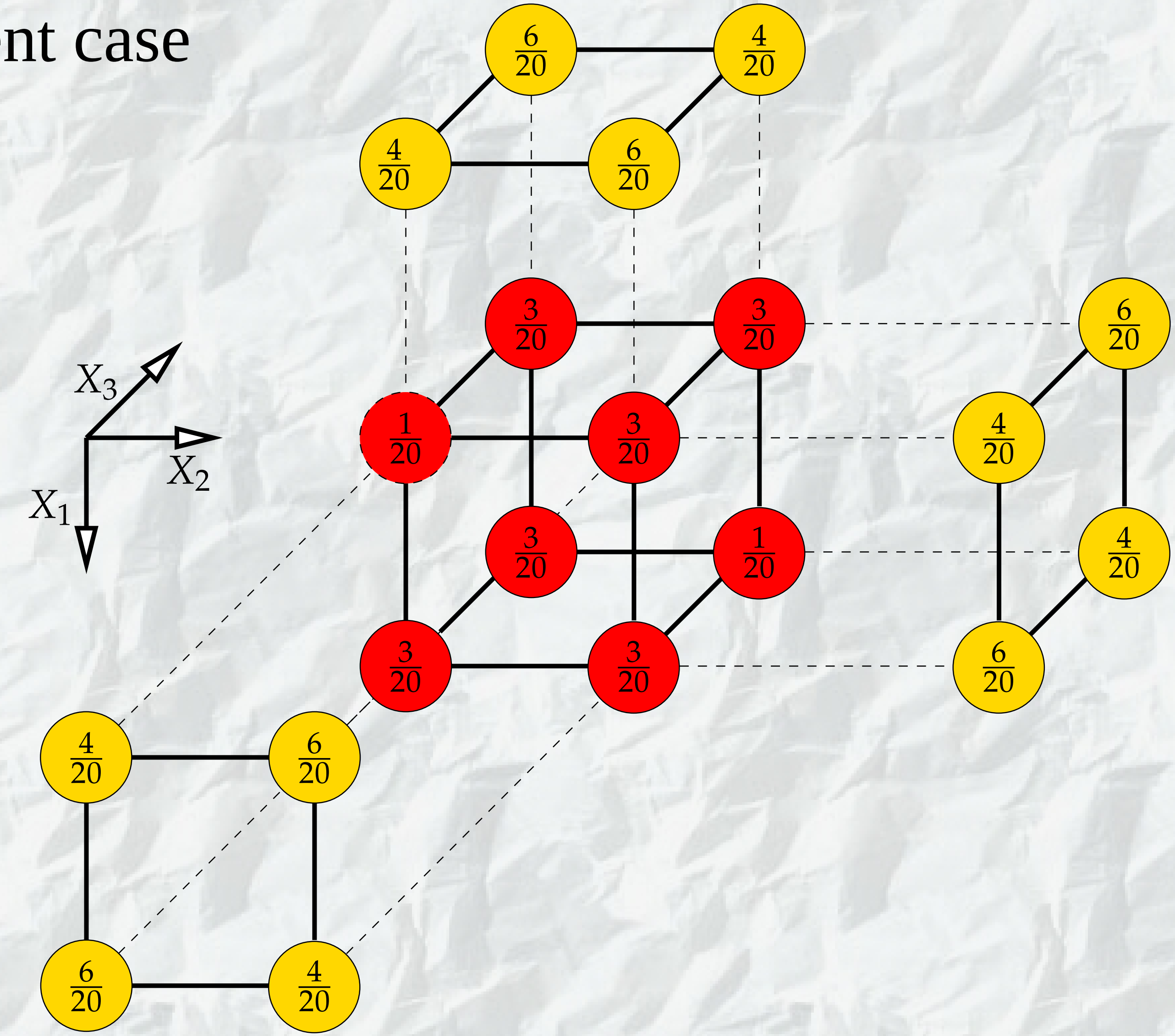
$\frac{1}{2}-\alpha$
α
α
$\frac{1}{2}-\alpha$
?
?
?
?
?
?
?
?
$\frac{1}{2}-\alpha$
α
α
$\frac{1}{2}-\alpha$
α
$\frac{1}{2}-\alpha$
$\frac{1}{2}-\alpha$
α
X_3
X_2
X_1



$\frac{1}{2}-\alpha$
α
α
$\frac{1}{2}-\alpha$
?
$\leq \alpha$
$\frac{1}{2}-\alpha$
X_3
X_2
X_1
?
$\leq \alpha$
α
?
?
α
?
?
$\frac{1}{2}-\alpha$
α
$\frac{1}{2}-\alpha$
$\frac{1}{2}-\alpha$
α
$\frac{1}{2}-\alpha \leq \alpha + \alpha$
$\frac{1}{6} \leq \alpha$

Consistent case

$\alpha = \frac{4}{20}$

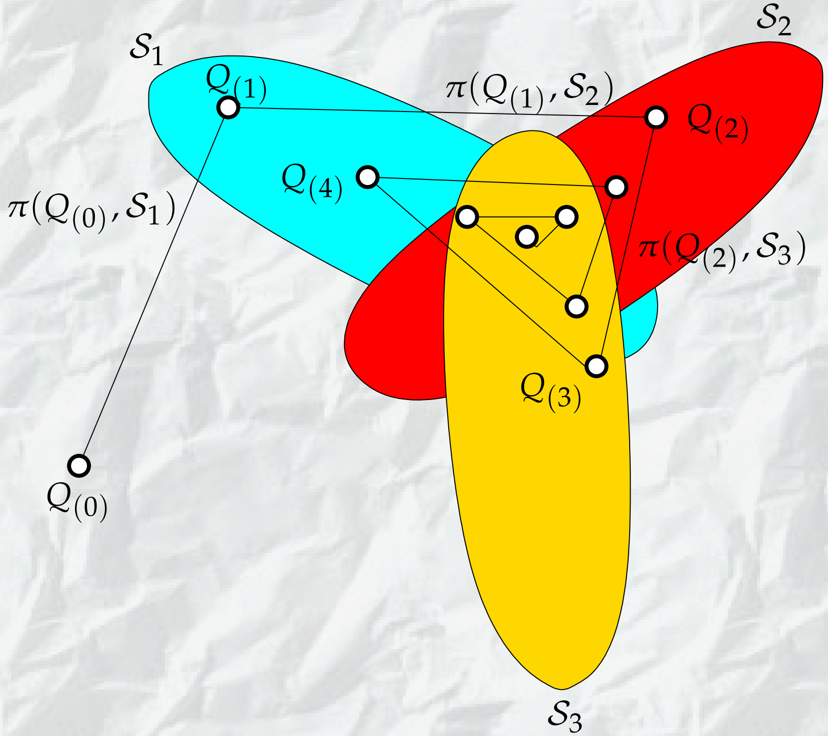


- input set $\mathcal{P} = \{P_1, \ldots, P_k\}$
- set of all distributions having P_j as its marginal
  $\mathcal{S}_j = \{Q : Q^{E_j} = P_j\}$
- set of all distributions having $\{P_1, \ldots, P_k\}$ as its marginals
  $\mathcal{S} = \cap_{j=1}^{k} \mathcal{S}_j$
- I-projection of Q_0 to $\mathcal{S}$
  $\pi(Q_0, \mathcal{S}) = \arg\min_{Q \in \mathcal{S}} I(Q \parallel Q_0)$
- Kullback-Leibler divergence

$$I(P \parallel Q) = \sum_x P(X = x) \log \frac{P(X = x)}{Q(X = x)}$$

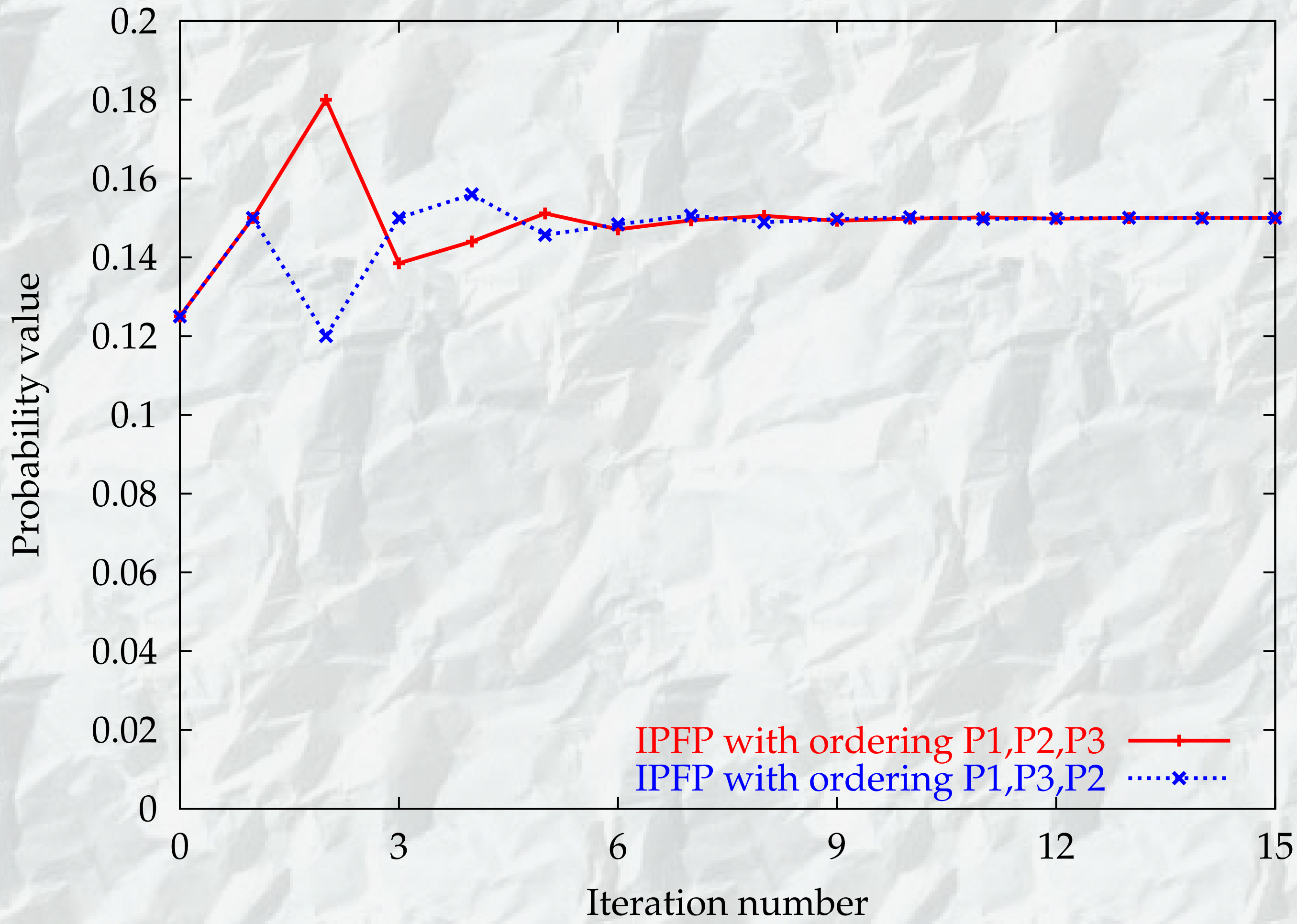
Iterative Proportional Fitting Procedure (IPFP)

Deming & Stephan, 1940



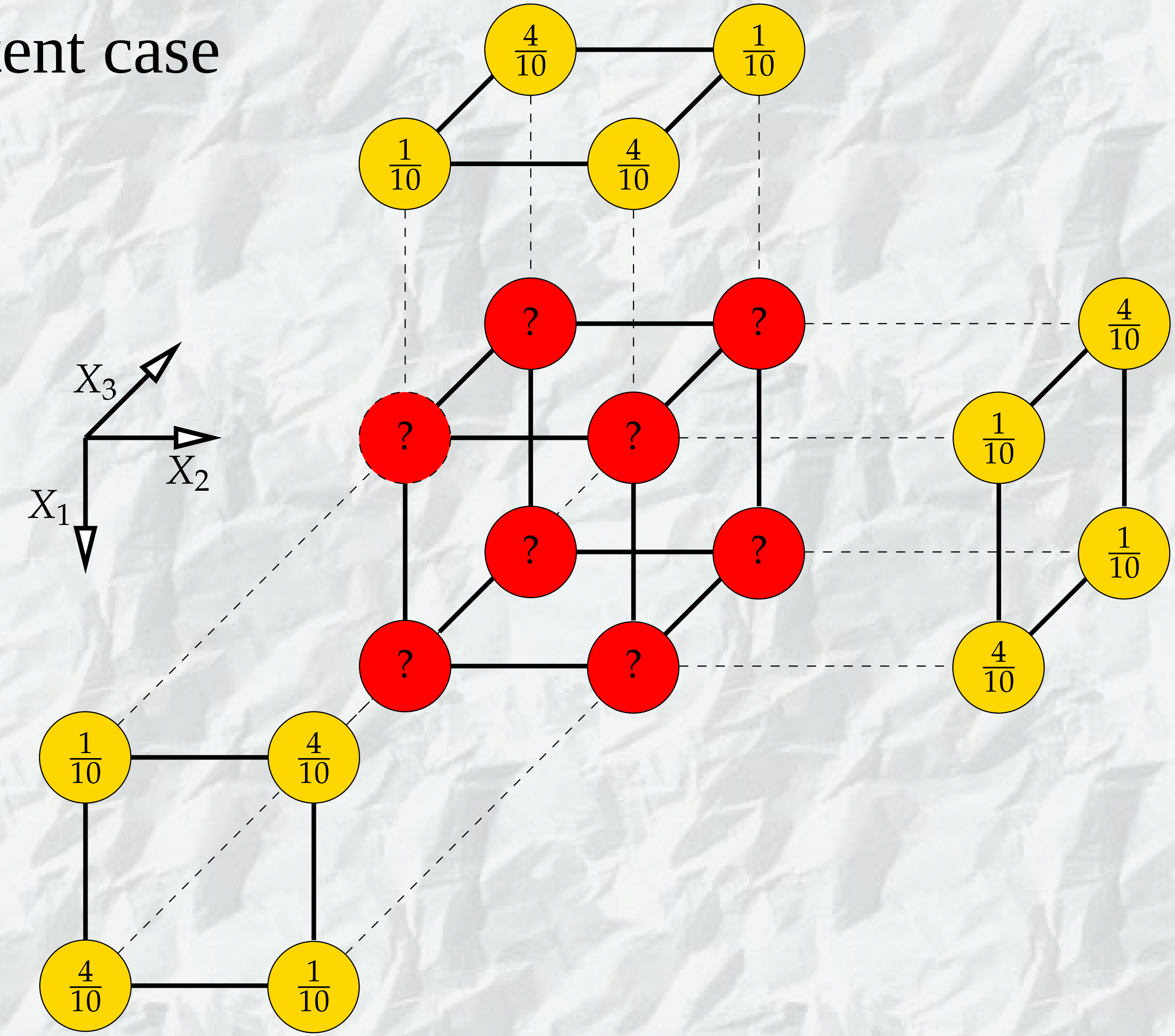
$$\pi(Q, \mathcal{S}_j) = Q \frac{P_j}{Q^{E_j}}$$

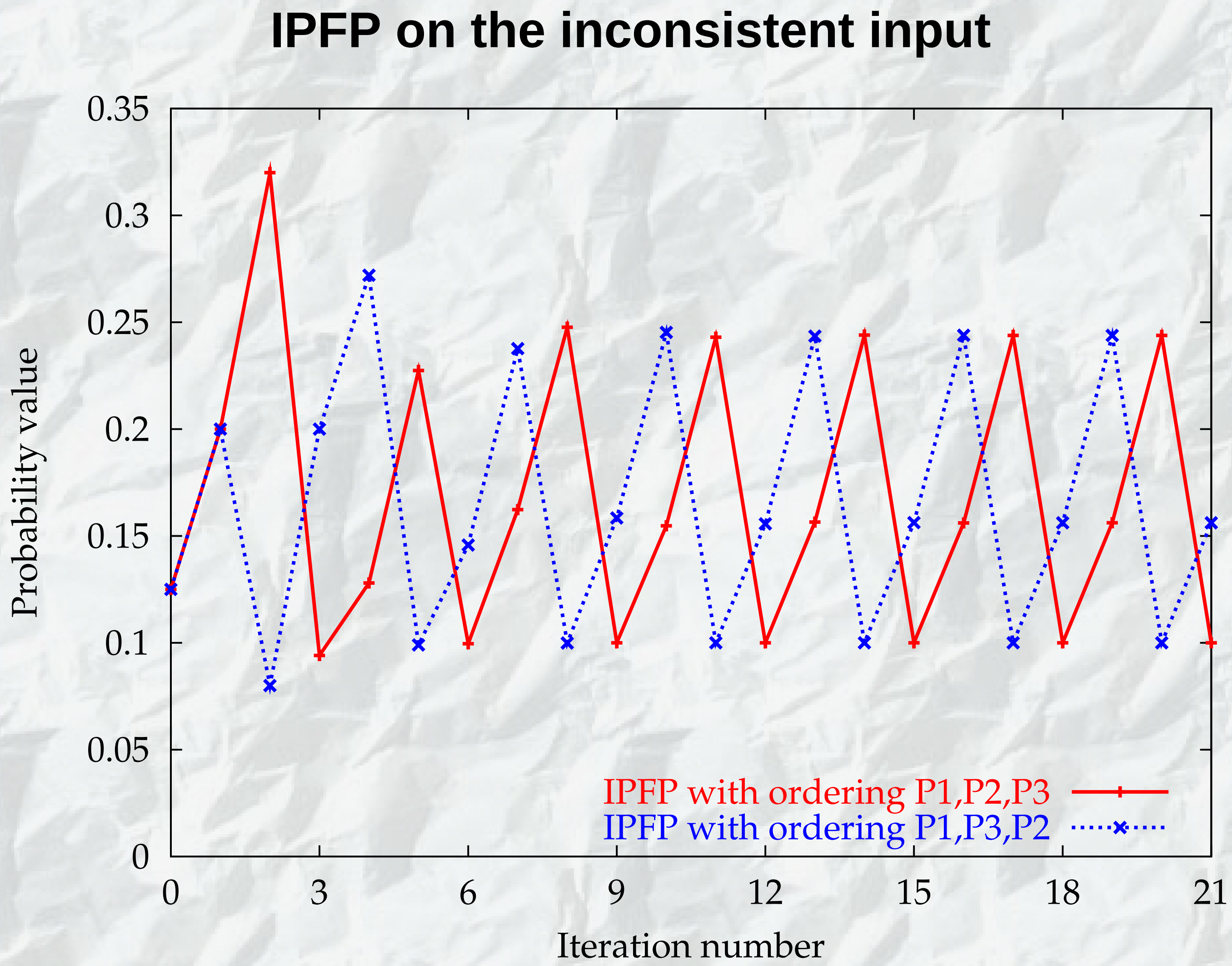
IPFP on the consistent input



Inconsistent case

$\alpha = \frac{1}{10}$




IPFP on the inconsistent input
Probability value
Iteration number
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
3
6
9
12
15
18
21
IPFP with ordering P1,P2,P3
IPFP with ordering P1,P3,P2

Let $r = \sqrt{-3\alpha^2 + 2\alpha}$, $\beta = 0.5(1 - \alpha - r)$, and $\gamma = 0.5(-\alpha + r)$.

The limit cycle for the ordering P_1, P_2, P_3

x	000	001	010	011	100	101	110	111
$lim_{n\to\infty} Q_{3n+1}(x)$	0	α	γ	β	β	γ	α	0
$lim_{n\to\infty} Q_{3n+2}(x)$	0	γ	β	α	α	β	γ	0
$lim_{n\to\infty} Q_{3n+3}(x)$	0	β	α	γ	γ	α	β	0
arithm. average	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0

The limit cycle for the ordering P_1, P_3, P_2

x	000	001	010	011	100	101	110	111
$lim_{n\to\infty} Q_{3n+1}(x)$	0	α	β	γ	γ	β	α	0
$lim_{n\to\infty} Q_{3n+2}(x)$	0	γ	α	β	β	α	γ	0
$lim_{n\to\infty} Q_{3n+3}(x)$	0	β	γ	α	α	γ	β	0
arithm. average	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0

For $\alpha = 0.1$ we get $\beta \doteq 0.244$ and $\gamma \doteq 0.156$.

Inconsistent input set $\mathcal{P} = \{P_1, \ldots, P_k\}$

It means that $\mathcal{S} = \cap_{j=1}^{k} \mathcal{S}_j = \emptyset$.

$Q(X_1, \ldots, X_n)$ is required to:

- **minimize a distance aggregate** with respect to $\mathcal{P}$:

$$\sum_{P_j \in \mathcal{P}} w_j \cdot d(P_j \parallel Q^{E_j})$$

- **factorize** with respect to $\mathcal{E} = \{E_1, \ldots, E_k\}$:
  there exist potentials $\psi_{E_i} : \mathbb{X}^{E_i} \mapsto \mathbb{R}, i = 1, 2, \ldots, k$ such that for all $x \in \mathbb{X}$

$$Q(x) = \prod_{E_i \in \mathcal{E}} \psi_{E_i}(x^{E_i}) \ .$$

Distance

- measured by the Kullback-Leibler divergence

$$d(P_j \parallel Q^{E_j}) \quad = \quad (P_j \parallel Q^{E_j}) \; = \; \sum_{x^{E_j}} P_j(x^{E_j}) \log \frac{P_j(x^{E_j})}{Q^{E_j}(x^{E_j})}$$

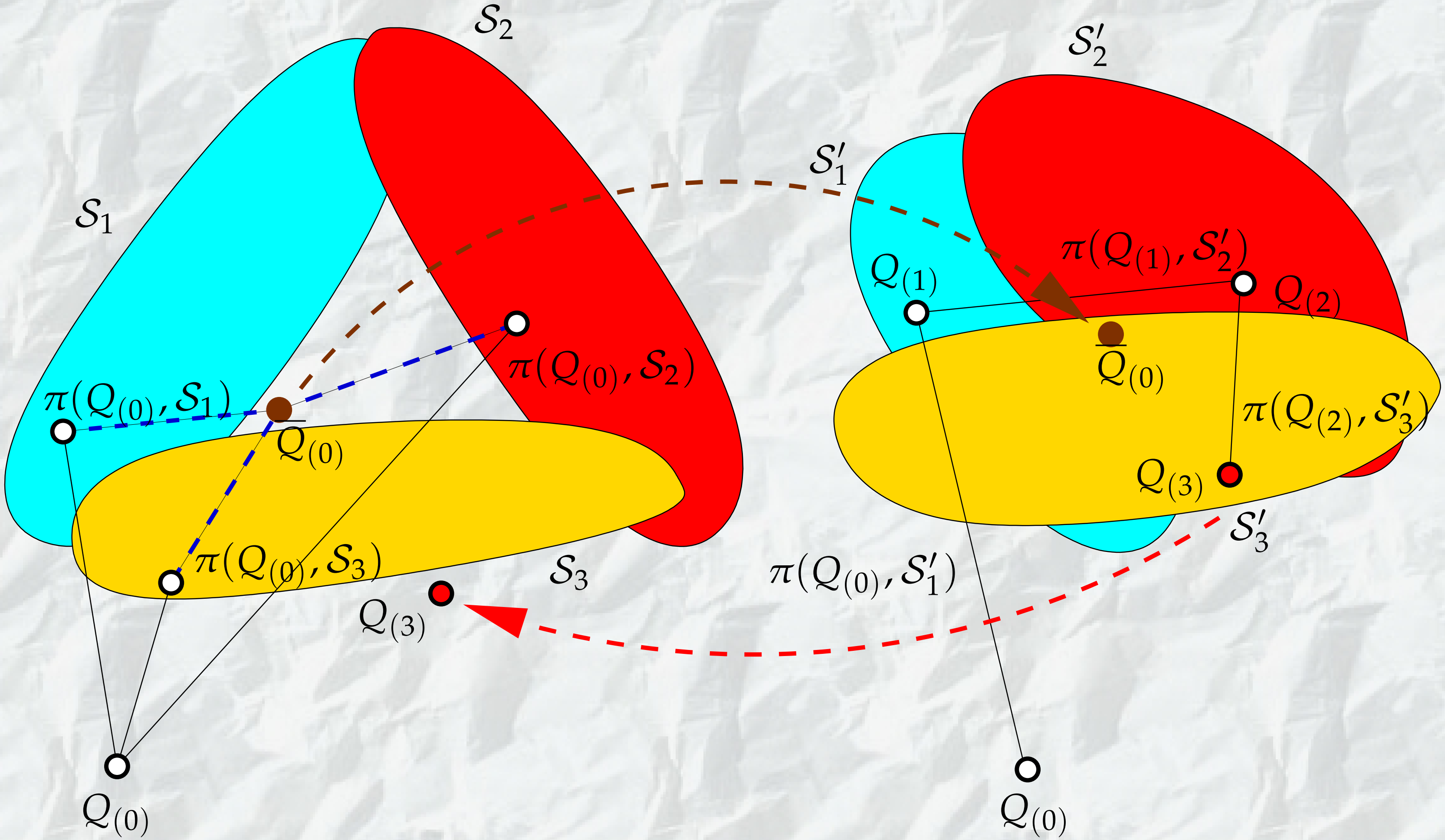
- measured by the total variance

$$d(P_j \parallel Q^{E_j}) \quad = \quad |P_j - Q^{E_j}| \; = \; \sum_{x^{E_j}} |P_j(x^{E_j}) - Q^{E_j}(x^{E_j})|$$

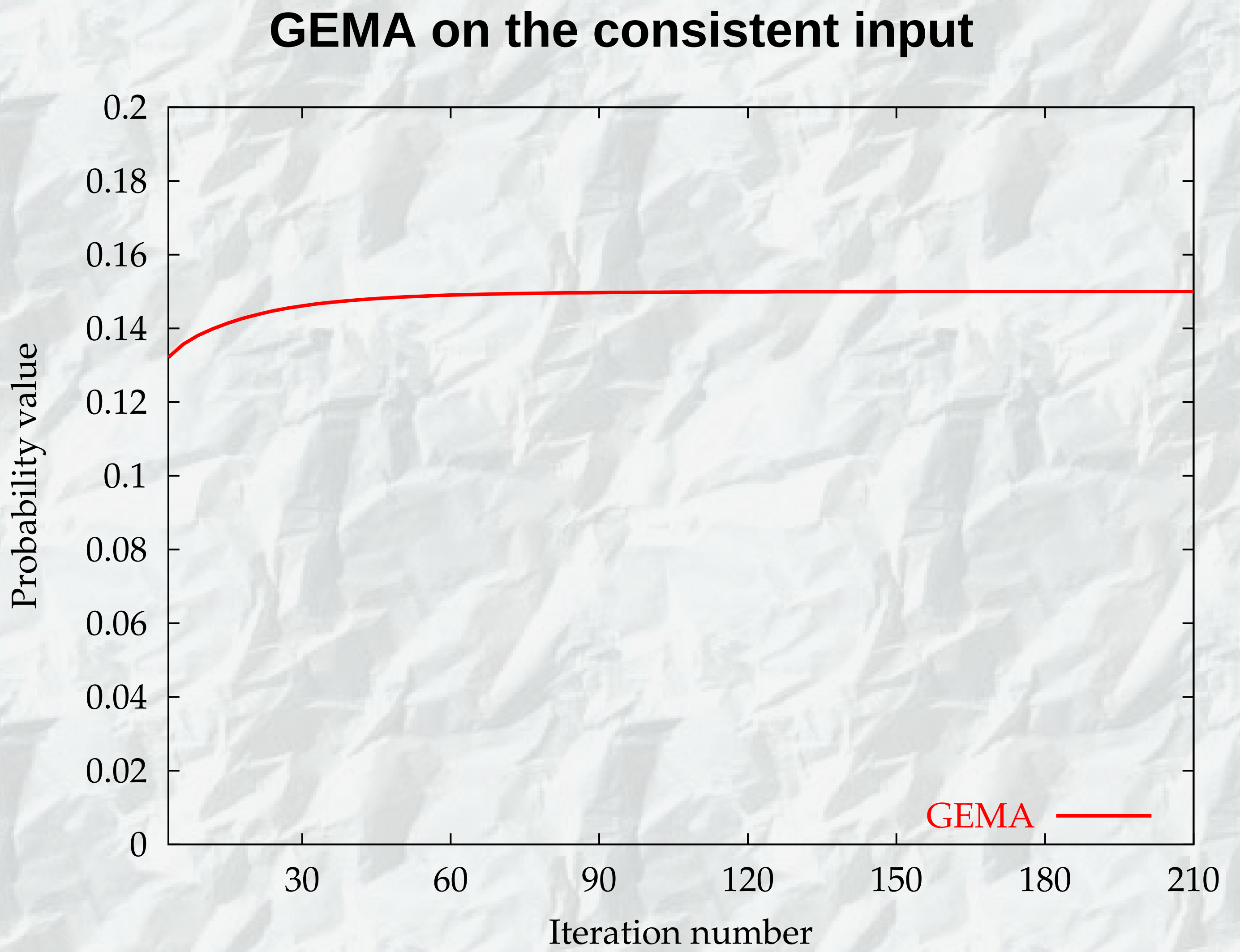
IPFP properties in the inconsistent case

- “converges” to a limit cycle
- distributions in the limit cycle are different for different orderings of the input set
- in the example, the average of the distributions in the limit cycle does not depend on the ordering – but generally, it is not true
- in the example, the distributions in the limit cycles minimized the aggregate of the total variance – but generally it is not known
- there are also other distributions that minimize the aggregate of the total variance that are not computed with IPFP
- generally, the distributions in the limit cycles **do not** minimize the aggregate of the Kullback-Leibler divergence
- distributions computed within a finite number of iterations factorize with respect to $\mathcal{E} = \{E_1, \ldots, E_k\}$

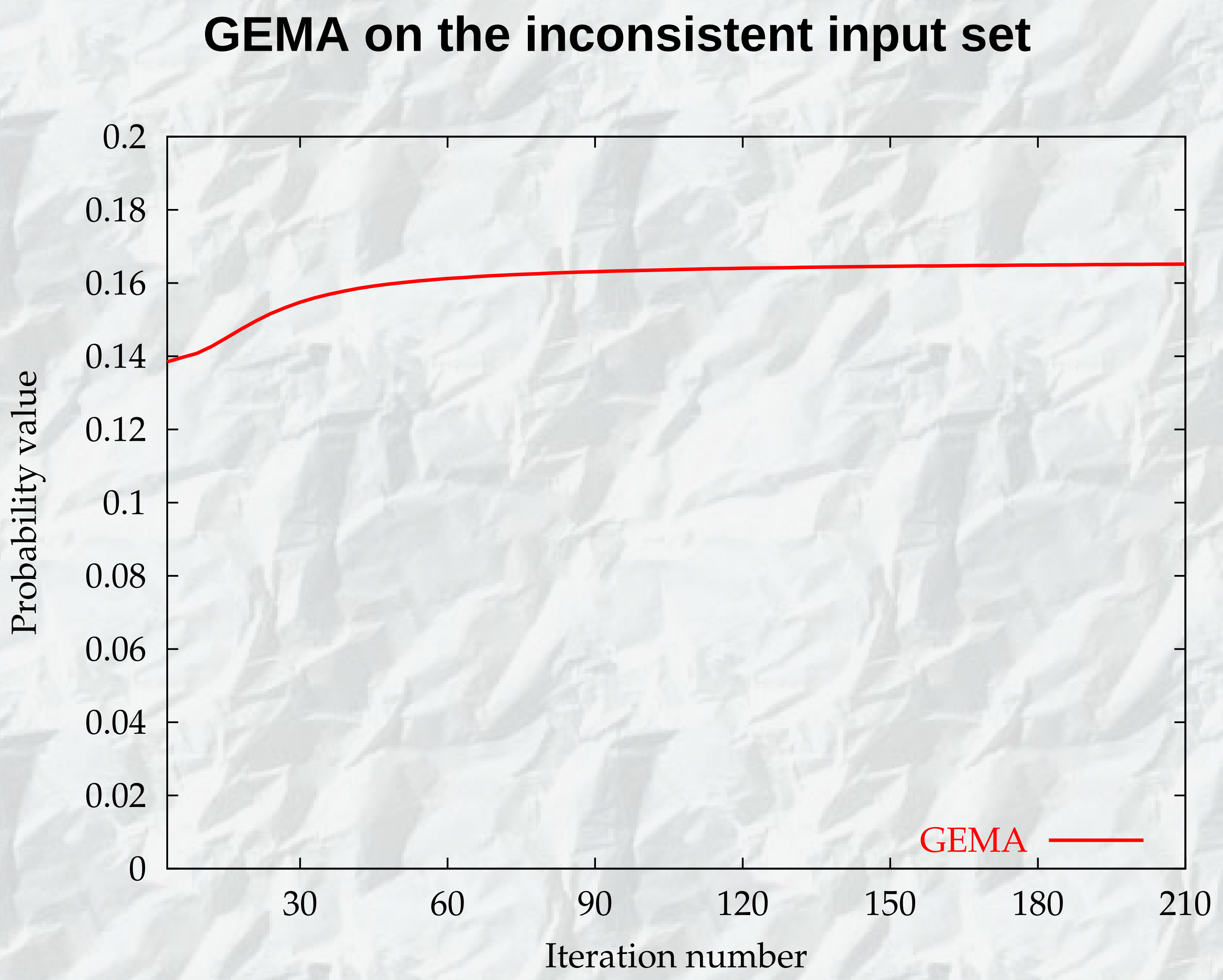
GEMA



$\mathcal{S}_2$
$\mathcal{S}_2'$
$\mathcal{S}_1'$
$\mathcal{S}_1$
$\pi(Q_{(1)}, \mathcal{S}_2')$
$Q_{(1)}$
$Q_{(2)}$
$\bar{Q}_{(0)}$
$\pi(Q_{(0)}, \mathcal{S}_2)$
$\pi(Q_{(0)}, \mathcal{S}_1)$
$\bar{Q}_{(0)}$
$\pi(Q_{(2)}, \mathcal{S}_3')$
$Q_{(3)}$
$\mathcal{S}_3'$
$\pi(Q_{(0)}, \mathcal{S}_3)$
$\mathcal{S}_3$
$\pi(Q_{(0)}, \mathcal{S}_1')$
$Q_{(3)}$
$Q_{(0)}$
$Q_{(0)}$



GEMA on the consistent input
Probability value
Iteration number
GEMA
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
30
60
90
120
150
180
210



GEMA on the inconsistent input set
Probability value
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
30
60
90
120
150
180
210
Iteration number
GEMA

GEMA properties

- converges also in the inconsistent case
- the limit distribution satisfies the necessary condition for the local minima of the aggregate of the Kullback-Leibler divergence
- the distributions computed within a finite number of iterations factorize with respect to $\mathcal{E} = \{E_1, \ldots, E_k\}$