

Integrating inconsistent data in a probabilistic model

Jiří Vomlel

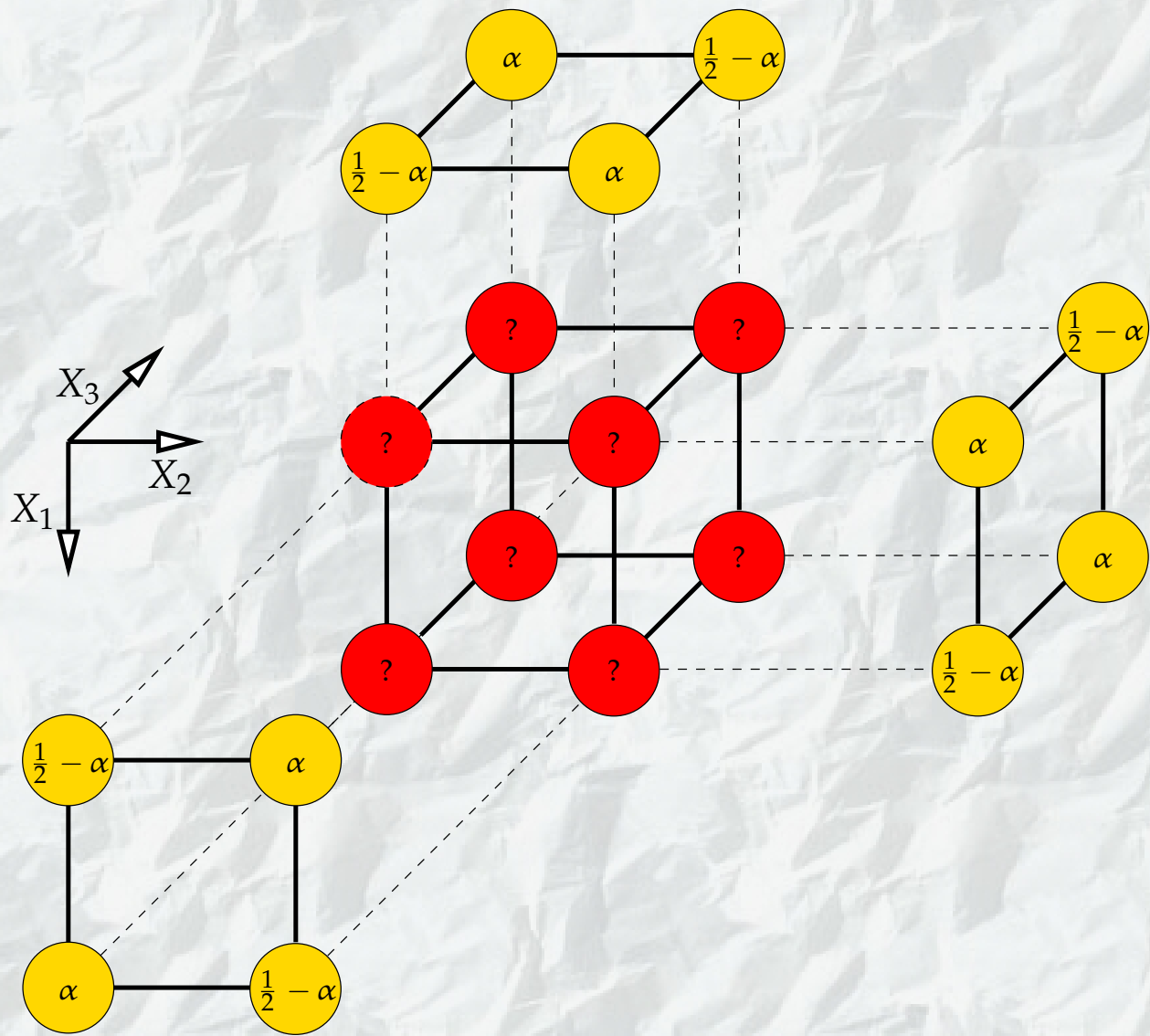
University of Economics

Academy of Science

} Prague, Czech Republic

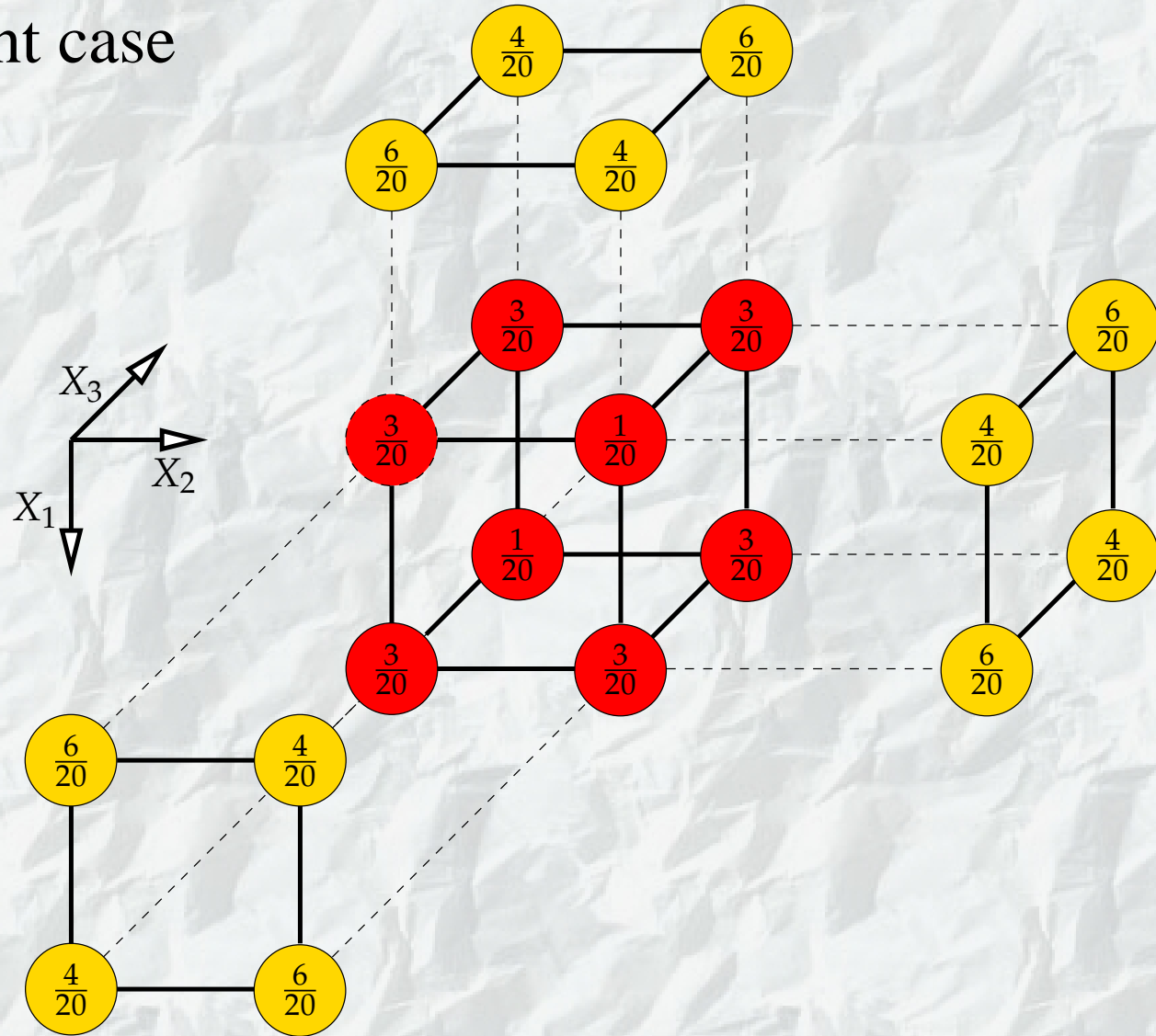
Knowledge integration

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



Consistent case

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

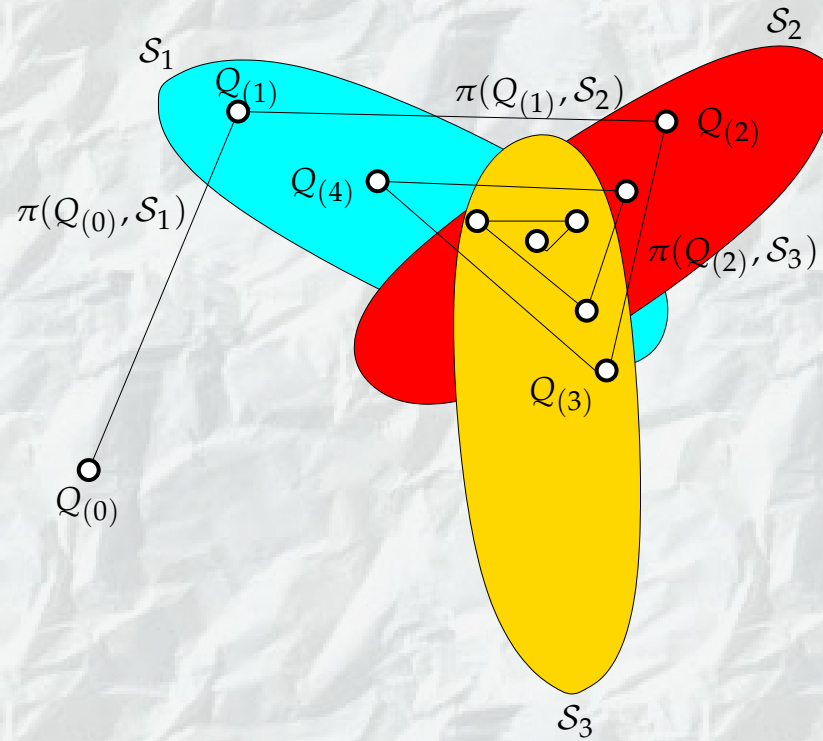


- **input set** $\mathcal{P} = \{P_1, \dots, P_k\}$
- **set of all distributions having P_j as its marginal**
 $\mathcal{S}_j = \{Q : Q^{E_j} = P_j\}$
- **set of all distributions having $\{P_1, \dots, P_k\}$ as its marginals**
 $\mathcal{S} = \bigcap_{j=1}^k \mathcal{S}_j$
- **I -projection** of Q_0 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)

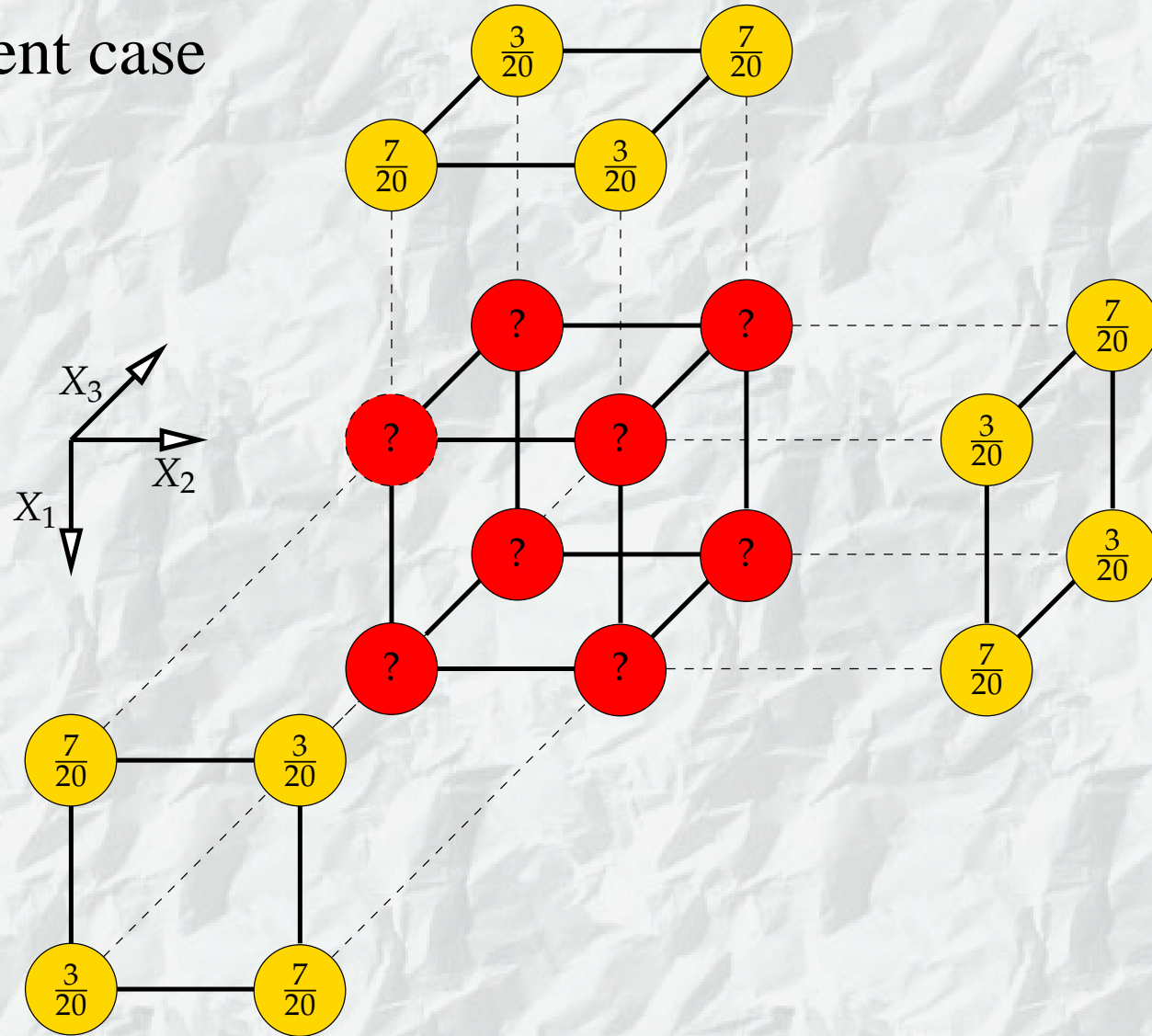
Deming & Stephan, 1940



$$\pi(Q, S_j) = Q \frac{P_j}{Q^{E_j}}$$

Inconsistent case

$$\alpha = \frac{3}{20}$$



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

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

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

- **minimize I -aggregate** with respect to \mathcal{P} :

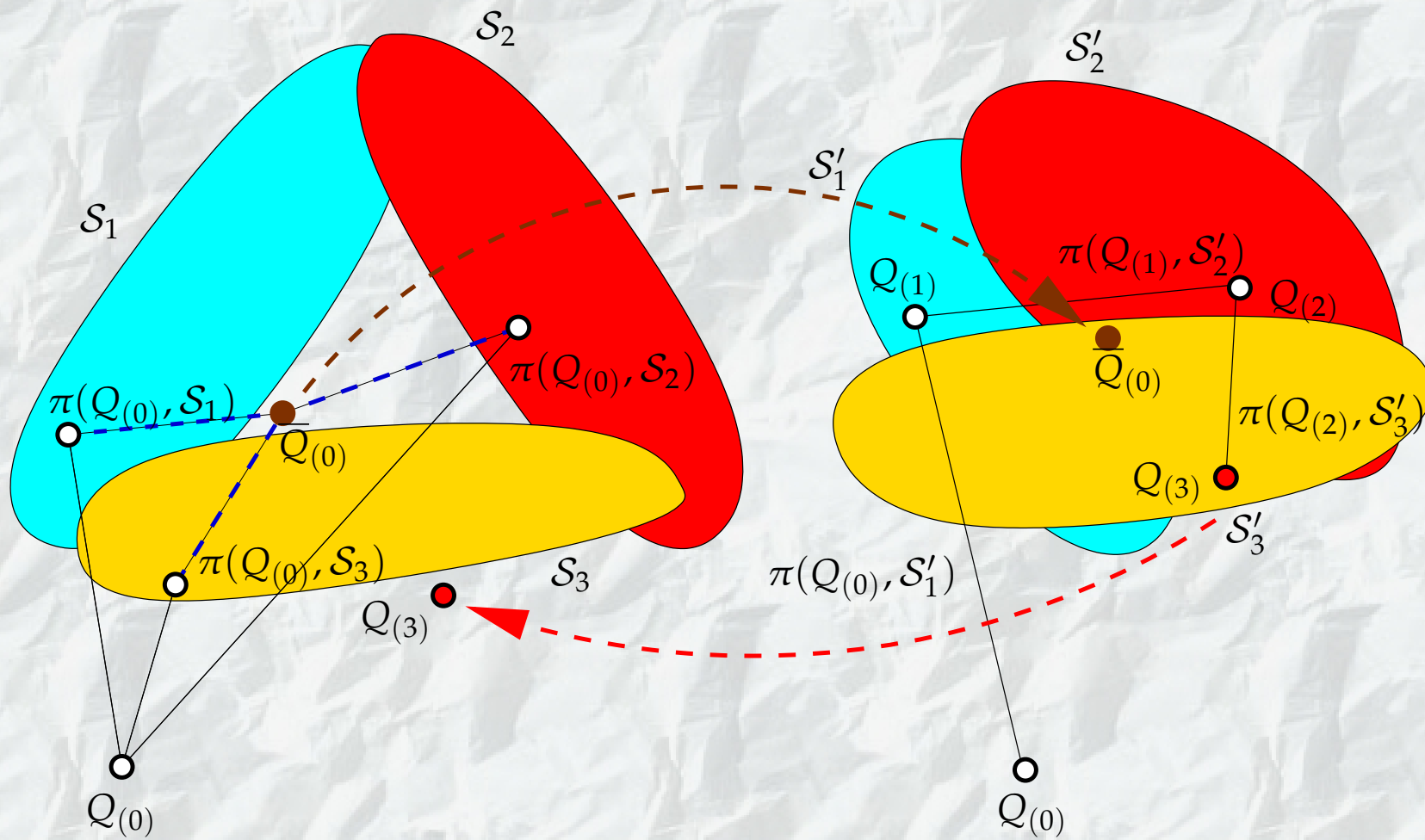
$$\sum_{P_j \in \mathcal{P}} w_j \cdot I(P_j \parallel Q^{E_j})$$

- **factorize** with respect to $\mathcal{E} = \{E_1, \dots, E_k\}$:

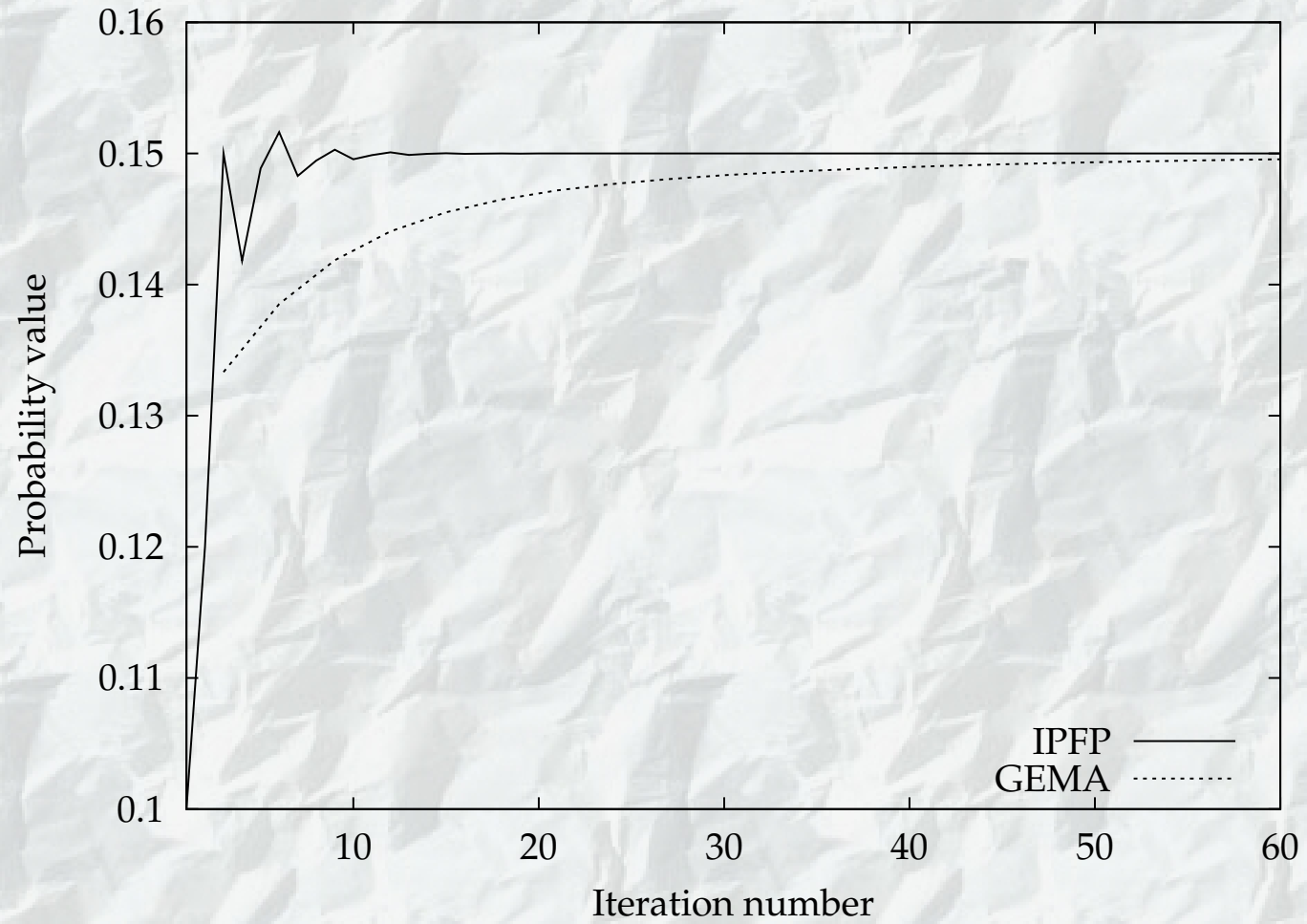
there exist potentials $\psi_{E_i} : \mathbb{X}^{E_i} \mapsto \mathbb{R}, i = 1, 2, \dots, k$ such that for all $x \in \mathbb{X}$

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

GEMA



IPFP and GEMA on the consistent input set



IPFP and GEMA on the inconsistent input set

