

Towards Chemical Applications of Dempster-Shafer-Type Approach: Case of Variant Ligands

Jaime Nava

Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968
jenava@miners.utep.edu

Formulation of the ...

Rota's Dempster-...

Relation to the...

Practical Applications...

Traditional...

From Continuous to...

Discrete Taylor...

Comparing Poset and...

Proof that The...

Home Page

Title Page

⏪

⏩

◀

▶

Page 1 of 17

Go Back

Full Screen

Close

Quit

1. Formulation of the Problem: Extrapolation Is Needed

- In many practical situations, molecules can be obtained from a “template” molecule like benzene C_6H_6 .
- *How:* by replacing some of its hydrogen atoms with *ligands* (other atoms or atom groups).
- *Fact:* there can be many possible replacements of this type.
- Testing of all possible replacements would be time-consuming.
- *It is desirable:* to test some of the replacements and then extrapolate to others.
- *Thus:* only the promising molecules will have to be synthesized and tested.

2. Formulation of the Problem: Extrapolation Is Needed (cont-d)

- D. J. Klein and co-authors proposed to use a poset extrapolation technique developed by G.-C. Rota.
- In many practical situations, this technique indeed leads to accurate predictions of many important quantities.
- *Limitation:* this technique has been originally proposed on a heuristic basis.
- There is no convincing justification of its applicability to chemical (or other) problems.
- *In this presentation:* we show that this equivalence can be extended to the case when we have variant ligands.

3. Rota's Dempster-Shafer-Type Poset Approach to Extrapolation: Reminder

- Rota considered the situation in which there is
 - a natural partial order relation \leq on some set of objects, and
 - a numerical value $v(a)$ associated to each object a from this partially ordered set (poset).
- *Main idea:* is that we can represent an arbitrary dependence $v(a)$ as

$$v(a) = \sum_{b: b \leq a} V(b)$$

for some values $V(b)$.

- To find values $V(b)$, solve a system of linear equations with as many unknowns $V(b)$ as the number of objects.
- It was proven that the poset-related system always has a solution.

Formulation of the ...

Rota's Dempster-...

Relation to the...

Practical Applications...

Traditional...

From Continuous to...

Discrete Taylor...

Comparing Poset and...

Proof that The...

Home Page

Title Page

◀◀

▶▶

◀

▶

Page 4 of 17

Go Back

Full Screen

Close

Quit

4. Relation to the Dempster-Shafer Approach

- The poset formula is identical to one of the main formulas of the Dempster-Shafer approach.
- Specifically, in this approach:
 - in contrast to a probability distribution when probabilities are assigned to different *elements* $x \in X$,
 - we have “masses” (in effect, probabilities) assigned to *subsets* $A \subseteq X$ of the set X .
- For each expert:
 - B is the set of alternatives that is possible according to this expert, and
 - $m(B)$ is the probability that this expert is correct based on his or her previous performance.

5. Relation to the Dempster-Shafer Approach (cont-d)

- For every set $A \subseteq X$ and for every expert, the expert's set B of possible alternatives can be contained in A .
- This means that this expert is sure that all possible alternatives are contained in the set A .
- *Thus:* our overall belief $\text{bel}(A)$ that the actual alternative is contained in A can be computed as

$$\text{bel}(A) = \sum_{B \subseteq A} m(B).$$

- This is the exact analog of the above formula, with
 - $v(a)$ instead of belief,
 - $V(b)$ instead of masses, and
 - $B \subseteq A$ as the ordering relation $b \leq a$.

Formulation of the ...

Rota's Dempster-...

Relation to the...

Practical Applications...

Traditional...

From Continuous to...

Discrete Taylor...

Comparing Poset and...

Proof that The...

Home Page

Title Page



Page 6 of 17

Go Back

Full Screen

Close

Quit

6. Practical Applications of the Poset Approach

- *In practice*: many values $V(b)$ turn out to be negligible and thus, can be safely taken as 0s.
- If we know which values $V(b_1), \dots, V(b_m)$ are non-zeros, we can then:
 - measure the value $v(a_1), \dots, v(a_p)$ of the desired quantity v for $p \ll n$ different objects a_1, \dots, a_p ;
 - use the Least Squares techniques to estimate the values $V(b_j)$ from the system

$$v(a_i) = \sum_{j: b_j \leq a_i} V(b_j), \quad i = 1, \dots, p;$$

- use the resulting estimates $V(b_j)$ to predict all the remaining values $v(a)$ ($a \neq a_1, \dots, a_m$), as

$$v(a) = \sum_{j: b_j \leq a} V(b_j).$$

Formulation of the ...

Rota's Dempster-...

Relation to the...

Practical Applications...

Traditional...

From Continuous to...

Discrete Taylor...

Comparing Poset and...

Proof that The...

Home Page

Title Page



Page 7 of 17

Go Back

Full Screen

Close

Quit

7. Traditional (Continuous) and Discrete Taylor Series

- In physical and engineering applications, most parameters x_1, \dots, x_n are *continuous*.
- The dependence $y = f(x_1, \dots, x_n)$ is also usually continuous and smooth (differentiable).
- Smooth functions can be usually expanded into Taylor series around some point $\tilde{x} = (\tilde{x}_1, \dots, \tilde{x}_n)$:

$$f(x_1, \dots, x_n) = f(\tilde{x}_1, \dots, \tilde{x}_n) + \sum_{i=1}^n \frac{\partial f}{\partial x_i} \cdot \Delta x_i +$$

$$\frac{1}{2} \cdot \sum_{i=1}^n \sum_{i'=1}^n \frac{\partial^2 f}{\partial x_i \partial x_{i'}} \cdot \Delta x_i \cdot \Delta x_{i'} + \dots,$$

where $\Delta x_i \stackrel{\text{def}}{=} x_i - \tilde{x}_i$.

Formulation of the ...

Rota's Dempster-...

Relation to the ...

Practical Applications ...

Traditional ...

From Continuous to ...

Discrete Taylor ...

Comparing Poset and ...

Proof that The ...

Home Page

Title Page

◀ ▶

◀ ▶

Page 8 of 17

Go Back

Full Screen

Close

Quit

8. Traditional (Continuous) and Discrete Taylor Series (cont-d)

- In practice, we can ignore higher-order terms.
- *Example:* if linear approximation is not accurate enough, we can use quadratic approximation.
- If we do not know the exact expression for $f(x_1, \dots, x_n)$, we do not know the values of its derivatives.

- All we know is that we approximate a general function by a general linear or quadratic formula

$$f(x_1, \dots, x_n) \approx c_0 + \sum_{i=1}^n c_i \cdot \Delta x_i + \sum_{i=1}^n \sum_{j=1}^n c_{ii'} \cdot \Delta x_i \cdot \Delta x_{i'}$$

- The values of the coefficients c_0 , c_i , and (if needed) $c_{ii'}$ can then be determined experimentally.

Formulation of the ...

Rota's Dempster-...

Relation to the...

Practical Applications...

Traditional...

From Continuous to...

Discrete Taylor...

Comparing Poset and...

Proof that The...

Home Page

Title Page



Page 9 of 17

Go Back

Full Screen

Close

Quit

9. From Continuous to Discrete Taylor Series

- *General case:* $y = f(x_{11}, \dots, x_{1N}, \dots, x_{n1}, \dots, x_{nN})$,
SO

$$y = y_0 + \sum_{i=1}^n \sum_{j=1}^N y_{ij} \cdot \Delta x_{ij} + \sum_{i=1}^n \sum_{j=1}^N \sum_{i'=1}^n \sum_{j'=1}^N y_{ij,i'j'} \cdot \Delta x_{ij} \cdot \Delta x_{i'j'},$$

where $\Delta x_{ij} \stackrel{\text{def}}{=} x_{ij} - d_{i0j}$.

- Let ε_{ik} denote the discrete variable that describes the presence of a ligand of type k at the location i :
 - when there is no ligand of type k at the location i , we take $\varepsilon_{ik} = 0$, and
 - when there is a ligand of type k at the location i , we take $\varepsilon_{ik} = 1$.
- If no ligand, $x_{ij} = d_{i0j}$. Thus $\Delta x_{ij} = d_{i0j} - d_{i0j} = 0$.
- If ligand, $x_{ij} = d_{ikj}$. Thus $\Delta x_{ij} = d_{ikj} - d_{i0j}$ is equal to $\Delta_{ikj} \stackrel{\text{def}}{=} d_{ikj} - d_{i0j}$.

10. From Continuous to Discrete Taylor Series (cont-d)

- For each location i , only one value ε_{ik} can be equal to 1, we can combine the above two cases into

$$\Delta x_{ij} = \sum_{k=1}^m \varepsilon_{ik} \cdot \Delta_{ikj}.$$

- Substituting into the original expression we obtain

$$y = y_0 + \sum_{i=1}^n \sum_{k=1}^m \sum_{j=1}^N y_{ij} \cdot \varepsilon_{ik} \cdot \Delta_{ikj} +$$

$$\sum_{i=1}^n \sum_{k=1}^m \sum_{j=1}^N \sum_{i'=1}^n \sum_{k'=1}^m \sum_{j'=1}^N y_{ij,i'j'} \cdot \varepsilon_{ik} \cdot \varepsilon_{i'k'} \cdot \Delta_{ikj} \cdot \Delta_{i'k'j'},$$

11. From Continuous to Discrete Taylor Series (cont-d)

- The above formula is equivalent to

$$y = y_0 + \sum_{i=1}^n \left(\sum_{k=1}^m \sum_{j=1}^N y_{ij} \cdot \Delta_{ikj} \right) \cdot \varepsilon_{ik} +$$

$$\sum_{i=1}^n \sum_{i'=1}^n \left(\sum_{j=1}^N \sum_{k=1}^m \sum_{j'=1}^N \sum_{k'=1}^m y_{ij,i'j'} \cdot \Delta_{ikj} \cdot \Delta_{i'k'j'} \right) \cdot \varepsilon_{ik} \cdot \varepsilon_{i'k'}.$$

- Combining terms proportional to each variable ε_{ik} and to each product $\varepsilon_{ik} \cdot \varepsilon_{i'k'}$, we obtain the expression

$$y = a_0 + \sum_{i=1}^n \sum_{k=1}^m a_{ik} \cdot \varepsilon_{ik} + \sum_{i=1}^n \sum_{k=1}^m \sum_{i'=1}^n \sum_{k'=1}^m a_{ik,i'k'} \cdot \varepsilon_{ik} \cdot \varepsilon_{i'k'},$$

where $a_{ik} = \sum_{j=1}^N y_{ij} \cdot \Delta_{ikj}$, and $a_{ik,i'k'} = \sum_{j=1}^N \sum_{j'=1}^N y_{ij,i'j'} \cdot \Delta_{ikj} \cdot \Delta_{i'k'j'}$.

12. Discrete Taylor Expansions Can be Further Simplified

- First, for each discrete variable $\varepsilon_{ik} \in \{0, 1\}$, we have $\varepsilon_{ik}^2 = \varepsilon_{ik}$.
- *Thus:* the term $a_{ik,ik} \cdot \varepsilon_{ik} \cdot \varepsilon_{ik}$ corresponding to $i = i'$ and $k = k'$ is equal to $a_{ik,ik} \cdot \varepsilon_{ik}$.
- *Therefore:* the term can be simply added to the corresponding linear term $a_{ik} \cdot \varepsilon_{ik}$.
- Second, we combine terms proportional to $\varepsilon_{ik} \cdot \varepsilon_{i'k'}$ and to $\varepsilon_{i'k'} \cdot \varepsilon_{ik}$.
- *As a result:* we obtain the following simplified expression

$$y = v_0 + \sum_{i=1}^n \sum_{k=1}^m v_{ik} \cdot \varepsilon_{ik} + \sum_{i < i'}^m \sum_{k=1}^m \sum_{k'=1}^m v_{ik,i'k'} \cdot \varepsilon_{ik} \cdot \varepsilon_{i'k'},$$

where $v_0 = c_0$, $v_{ik} = c_{ik}$, and $v_{ik,i'k'} = c_{ik,i'k'} + c_{i'k',ik}$.

13. Comparing Poset and Discrete Taylor Series Approaches

- *Reminder:* $\varepsilon_{ik} = 0$ means no ligand, $\varepsilon_{ik} = 1$ means ligand

- *Taylor series:*

$$y = v_0 + \sum_{i=1}^n \sum_{k=1}^m v_{ik} \cdot \varepsilon_{ik} + \sum_{i < i'}^m \sum_{k=1}^m \sum_{k'=1}^m v_{ik,i'k'} \cdot \varepsilon_{ik} \cdot \varepsilon_{i'k'}.$$

- *Poset approach:* $v(a) = \sum_{b: b \leq a} V(b)$
- Here, $b \leq a$ means that a can be obtained from b by adding ligands.
- So, if $b = (\varepsilon'_{11}, \dots, \varepsilon'_{nm})$ and $a = (\varepsilon_{11}, \dots, \varepsilon_{nm})$, then $b \leq a$ means that for every i and k , we have $\varepsilon'_{ik} \leq \varepsilon_{ik}$.
- *Resulting formula:*

$$y = V(a_0) + \sum_{(i,k): \varepsilon_{ik}=1} V(a_{ik}) + \sum_{i < i', k, k': \varepsilon_{ik}=\varepsilon_{i'k'}=1} V(a_{ik,i'k'}).$$

Formulation of the ...

Rota's Dempster-...

Relation to the ...

Practical Applications ...

Traditional ...

From Continuous to ...

Discrete Taylor ...

Comparing Poset and ...

Proof that The ...

Home Page

Title Page



Page 14 of 17

Go Back

Full Screen

Close

Quit

14. Proof that The Discrete Taylor Series are Indeed Equivalent to the Poset Formula

- *Taylor series:*

$$y = v_0 + \sum_{i=1}^n \sum_{k=1}^m v_{ik} \cdot \varepsilon_{ik} + \sum_{i < i'}^m \sum_{k=1}^m \sum_{k'=1}^m v_{ik,i'k'} \cdot \varepsilon_{ik} \cdot \varepsilon_{i'k'}.$$

- *Poset:*

$$y = V(a_0) + \sum_{(i,k): \varepsilon_{ik}=1} V(a_{ik}) + \sum_{i < i', k, k': \varepsilon_{ik}=\varepsilon_{i'k'}=1} V(a_{ik,i'k'}).$$

- Proof that these formulas coincide:

$$\sum_{(i,k): \varepsilon_{ik}=1} V(a_{ik}) = \sum_{(i,k): \varepsilon_{ik}=1} V(a_{ik}) \cdot \varepsilon_{ik} = \sum_{i=1}^n \sum_{k=1}^m V(a_{ik}) \cdot \varepsilon_{ik}.$$

15. Proof that The Discrete Taylor Series are Indeed Equivalent to the Poset Formula (cont-d)

- Similarly, the quadratic part of the sum

$$\sum_{i < i', k, k': \varepsilon_{ik} = \varepsilon_{i'k'} = 1} V(a_{ik, i'k'}) = \sum_{i < i', k, k': \varepsilon_{ik} = \varepsilon_{i'k'} = 1} V(a_{ik, i'k'}) \cdot \varepsilon_{ik} \cdot \varepsilon_{i'k'} =$$
$$\sum_{i < i'} \sum_{k=1}^m \sum_{k'=1}^m V(a_{ik, i'k'}) \cdot \varepsilon_{ik} \cdot \varepsilon_{i'k'}.$$

- Substituting, we obtain

$$y = V(a_0) + \sum_{i=1}^n V(a_{ik}) \cdot \varepsilon_{ik} + \sum_{i < i'} \sum_{k=1}^m \sum_{k'=1}^m V(a_{ik, i'k'}) \cdot \varepsilon_{ik} \cdot \varepsilon_{i'k'}.$$

- This expression is identical to the discrete Taylor formula.

16. Acknowledgments

- The author would like to thank Dr. Vladik Kreinovich, for his encouragement.
- Thanks to Dr. James Salvador, for valuable discussions.
- This work was supported in part by:
 - by National Science Foundation grants HRD-0734825 and DUE-0926721,
 - by Grant 1 T36 GM078000-01 from the National Institutes of Health.

Formulation of the...

Rota's Dempster...

Relation to the...

Practical Applications...

Traditional...

From Continuous to...

Discrete Taylor...

Comparing Poset and...

Proof that The...

[Home Page](#)

[Title Page](#)



Page 17 of 17

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)