12-2010

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

Jaime Nava
The University of Texas at El Paso, jenava@miners.utep.edu

Comments:
Technical Report: UTEP-CS-10-27a
Published in *Proceedings of the 30th Annual Conference of the North American Fuzzy Information Processing Society NAFIPS'2011*, El Paso, Texas, March 18-20, 2011.

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, Texas 79968
Email: jenava@miners.utep.edu

***Abstract*—In many practical situations, molecules can be obtained from a "template" molecule like benzene by replacing some of its hydrogen atoms with ligands (other atoms or atom groups). There can be many possible replacements of this type. To avoid time-consuming testing of all possible replacements, it is desirable to test some of the replacements and then extrapolate to others – so that only the promising molecules, for which the extrapolated values are desirable, will have to be synthesized and tested.**

For this extrapolation, D. J. Klein and co-authors proposed to use a Dempster-Shafer-type poset extrapolation technique developed by G.-C. Rota from MIT. One of the limitations of this approach is that this technique has been originally proposed on a heuristic basis, with no convincing justification of its applicability to chemical (or other) problems. In our previous paper, we showed that for the case when all the ligands are of the same type, the poset technique is actually equivalent to a more familiar (and much more justified) Taylor series extrapolation. In this paper, we show that this equivalence can be extended to the case when we have variant ligands.

I. Formulation of the Problem: Extrapolation Is Needed

In many practical situations, molecules can be obtained from a "template" molecule like benzene C_6H_6 by replacing some of its hydrogen atoms with *ligands* (other atoms or atom groups). There can be many possible replacements of this type. To avoid time-consuming testing of all possible replacements, it is desirable to test some of the replacements and then extrapolate to others – so that only the promising molecules, for which the extrapolated values are desirable, will have to be synthesized and tested.

For this extrapolation, D. J. Klein and co-authors proposed to use a poset extrapolation technique developed by G.-C. Rota from MIT; see, e.g., [10]. They showed that in many practical situations, this technique indeed leads to accurate predictions of many important quantities; see, e.g., [1], [3], [4], [5], [6], [7], [8].

One of the limitations of this approach is that this technique has been originally proposed on a heuristic basis, with no convincing justification of its applicability to chemical (or other) problems. In our previous paper [9], we showed that for the case when all the ligands are of the same type, the poset technique is actually equivalent to a more familiar (and much more justified) Taylor series extrapolation.

In this paper, we show that this equivalence can be extended to the case when we have variant ligands.

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

Main idea. In [10], Gian-Carlo Rota, a renowned mathematician from MIT, considered the situation in which there is a natural partial order relation $\le$ on some set of objects, and there is a numerical value $v(a)$ associated to each object a from this partially ordered set (poset).

Rota's technique is based on the fact that we can represent an arbitrary dependence $v(a)$ as

$$v(a) = \sum_{b:\, b \le a} V(b) \tag{1}$$

for some values $V(b)$. The possibility to find such values $V(b)$ is easy to understand: the above formula (1) is a system of linear equations in which we have as many unknowns $V(b)$ as the number of objects – and as many equations as the number of objects. Thus, we have a system of linear equations with as many equations as there are unknowns. It is known that in general, such a system always has a solution. (In principle, there are degenerate cases when a system of n linear equations with n unknowns does not have a solution, but in [10] it was proven that the poset-related system (1) always has a solution.)

Relation to the Dempster-Shafer approach. From the purely mathematical viewpoint, formula (1) is identical to one of the main formulas of the Dempster-Shafer approach (see, e.g., [12]). Specifically, in this approach,

- in contrast to a probability distribution on a set X when probabilities $p(x) \ge 0$, $\sum\limits_{x \in X} p(x) = 1$, are assigned to different elements $x \in X$ of the set X,
- we have "masses" (in effect, probabilities) $m(A) \ge 0$, $\sum\limits_{A} m(A) = 1$, assigned to *subsets* $A \subseteq X$ of the set X.

The usual meaning of the values $m(B)$ is, e.g., that we have several experts who have different opinions on which

alternatives are possible and which are not. 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 (estimated, e.g., based on his or her previous performance).

For every set $A \subseteq X$ and for every expert, if the expert's set B of possible alternatives is contained in A ($B \subseteq A$), this means that this expert is sure that all possible alternatives are contained in the set A. Thus, our overall belief $\mathrm{bel}(A)$ that the actual alternative is contained in A can be computed as the sum of the masses corresponding to all such experts, i.e., as

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

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

Comment. It should be mentioned that in spite of the above similarity, Rota's poset approach is somewhat different from the Dempster-Shafer approach:

- first, in the Dempster-Shafer approach, we require that all the masses are non-negative, while in the poset approach, the corresponding values $V(b)$ can be negative as well;
- second, in the Dempster-Shafer approach, we require that the sum of all the masses is 1, while in the poset approach, the sum of all the values $V(b)$ can be any real number.

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 $V(b_1), \ldots, V(b_m)$ are non-zeros, we can then:

- measure the value $v(a_1), \ldots, v(a_p)$ of the desired quantity v for $p \ll n$ different objects $a_1, \ldots, a_p$;
- use the Least Squares techniques (see, e.g. [11]) to estimate the values $V(b_j)$ from the system

$$v(a_i) = \sum_{j:\, b_j \le a_i} V(b_j),\ i = 1, \ldots, p; \tag{2}$$

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

$$v(a) = \sum_{j:\, b_j \le a} V(b_j). \tag{3}$$

Comment. The problem of estimating the values $V(b)$ based on the known values $v(a)$ is similar to the problem of determining masses from the belief values in the Dempster-Shafer approach. Thus, to estimate the values $V(b)$, we can use algorithms developed within the Dempster-Shafer approach.

Application to chemistry. In chemistry, objects are molecules, and a natural relation $a \le b$ means that the molecule b either coincides with a, or can be obtained from the molecule a if we replace one or several of its H atoms with ligands.

III. Traditional (Continuous) and Discrete Taylor Series

Traditional (continuous) Taylor series: a brief reminder. Traditionally, in physical and engineering applications, most parameters $x_1, \ldots, x_n$ (such as coordinates, velocity, etc.) are *continuous* – in the sense that their values can continuously change from one value to another. The dependence $y = f(x_1, \ldots, x_n)$ of a quantity y on the parameters x_i is also usually continuous (with the exception of phase transitions); moreover, this dependence is usually smooth (differentiable). It is known that smooth functions can be usually expanded into Taylor series around some point $\widetilde{x} = (\widetilde{x}_1, \ldots, \widetilde{x}_n)$ (e.g., around the point $\widetilde{x} = 0$), i.e., as a sum of constant terms, linear terms, quadratic terms, and terms of higher order.

$$f(x_1, \ldots, x_n) = f(\widetilde{x}_1, \ldots, \widetilde{x}_n) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \cdot \Delta x_i + \frac{1}{2} \cdot \sum_{i=1}^{n} \sum_{i'=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_{i'}} \cdot \Delta x_i \cdot \Delta x_{i'} + \ldots,$$

where $\Delta x_i \stackrel{\mathrm{def}}{=} x_i - \widetilde{x}_i$.

The values of different order terms in the Taylor expansion usually decrease when the order increases – after all, the Taylor series usually converges, which implies that the terms should tend to 0. So, in practice, we can ignore higher-order terms and consider only the first few terms in the Taylor expansion. (This is, for example, how most elementary functions like $\sin(x)$, $\cos(x)$, $\exp(x)$ are computed inside the computers.).

In the simplest case, it is sufficient to preserve linear terms, i.e. to use the approximation

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

When the linear approximation is not accurate enough, we can use the quadratic approximation

$$f(x_1, \ldots, x_n) \approx f(\widetilde{x}_1, \ldots, \widetilde{x}_n) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \cdot \Delta x_i + \frac{1}{2} \cdot \sum_{i=1}^{n} \sum_{i'=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_{i'}} \cdot \Delta x_i \cdot \Delta x_{i'},$$

etc.

Since we do not know the exact expression for the function $f(x_1, \ldots, x_n)$, we thus do not know the actual values of its derivatives $\dfrac{\partial f}{\partial x_i}$ and $\dfrac{\partial^2 f}{\partial x_i \partial x_{i'}}$. Hence, when we actually use this approximation, all we know is that we approximate a general function by a general linear or quadratic formula

$$f(x_1, \ldots, x_n) \approx c_0 + \sum_{i=1}^{n} c_i \cdot \Delta x_i,$$

$$f(x_1, \ldots, x_n) \approx c_0 + \sum_{i=1}^{n} c_i \cdot \Delta x_i + \sum_{i=1}^{n} \sum_{i'=1}^{n} c_{ii'} \cdot \Delta x_i \cdot \Delta x_{i'}, \tag{4}$$

where $c_0 = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$, $c_i = \dfrac{\partial f}{\partial x_i}$, and $c_{ii'} = \dfrac{1}{2} \cdot \dfrac{\partial^2 f}{\partial x_i \partial x_{i'}}$.

In the traditional physical and engineering applications, the values of the coefficients c_0, c_i, and (if needed) $c_{ii'}$ can then be determined experimentally. Namely, in several (E) different experiments $e = 1, 2, \ldots, E$, we measure the values $x_i^{(e)}$ of the parameters and the resulting value $y^{(e)}$, and then determine the desired coefficients by applying the Least Squares method to the corresponding approximate equations. In the case of linear dependence, we use approximate equations

$$y^{(e)} \approx c_0 + \sum_{i=1}^{n} c_i \cdot \Delta x_i^{(e)}; \quad e = 1, 2, \ldots, E. \tag{5}$$

In the case of quadratic dependence, we use approximate equations

$$y^{(e)} \approx c_0 + \sum_{i=1}^{n} c_i \cdot \Delta x_i^{(e)} + \sum_{i=1}^{n} \sum_{i'=1}^{n} c_{ii'} \cdot \Delta x_i^{(e)} \cdot \Delta x_{i'}^{(e)}. \tag{6}$$

From continuous to discrete Taylor series. As we have mentioned in [9], we can extend the Taylor series approach to the discrete case.

In our chemical problem, the discrete case means that for each location, we are only interested in the values of the desired physical quantity in the following situations:

- a situation when there is a ligand at this location, and
- a situation when there is no ligand at this location.

From the macroscopic viewpoint, there are only these two options. However, on the microscopic level, the situation is more complex. Chemical interactions are, in effect, interaction of electrons. A proper description of an electron requires quantum physics; see, e.g., [2].

In classical (pre-quantum) physics, to describe the state of a particle at any given moment of time, it is sufficient to describe its spatial location and momentum. We may not know the exact values of these quantities, but in principle, we can determine them with an arbitrarily high accuracy.

In contrast, in quantum physics, it is impossible to uniquely determine both spatial location and momentum, we can only predict probabilities of different spatial locations (and different momentum values). A quantum description of the state of a particle is a *wave function* $\psi(x)$, a complex-valued function for which, for a small neighborhood of volume ΔV around a point x, the probability to find the electron in this neighborhood is approximately equal to $|\psi(x)|^2 \cdot \Delta V$. In other words, $|\psi(x)|^2$ is the probability density – electronic density in case of electrons.

In principle, electrons can be in many different states, with different electronic density functions $|\psi(x)|^2$. In chemistry, we usually consider only the stable (lowest energy) states. From this viewpoint, we have one of the two situations:

- a situation in which there is a ligand at this location; in this case, we consider the lowest-energy state of a molecule with a ligand at this location; and
- a situation in which there is no ligand at this location; in this case, we consider the lowest-energy state of a molecule with no ligand at this location.

However, from the physical viewpoint, it also makes sense to consider "excited" (higher-energy) states as well, states with arbitrary (not necessarily lowest-energy) electron density functions. Many such states occur as intermediate states in chemical reactions, when a molecule or a group of molecules continuously moves from the original stable state (before the reaction) to a new stable state (after the reaction).

The general physical laws and dependencies are not limited to the discrete (lowest-energy) situations, they work for other (not so stable) situations as well.

So, while we are interested in the values of the desired physical quantity (such as energy) corresponding to the selected stable situations, in principle, we can consider this dependence for other (not so stable) situations as well. The value of, e.g. energy, depends on the values of the electronic density at different points near the ligand locations, etc. For each possible placement of a ligand of type k ($1 \le k \le m$) at a location i ($1 \le i \le n$), let $x_{ik1}, \ldots, x_{ikj}, \ldots, x_{ikN}$ be parameters describing the distribution in the vicinity of this location (e.g., the density at a certain point, the distance to a certain atom, the angle between this atom and the given direction, the angle describing the direction of the spin, etc.). In general, the value of the desired quantity depends on the values of these parameters:

$$y = f(x_{111}, \ldots, x_{11N}, \ldots, x_{nm1}, \ldots, x_{nmN}). \tag{7}$$

We are interested in the situations in which, at each location, there is either a ligand, or there is no ligand. For each location i and for each parameter x_{ij}:

- let d_{i0j} denote the value of the j-th parameter in the situation with no ligand at the location i, and
- let d_{ikj} denote the value of the j-th parameter in the situation with a ligand of type k at the location i.

The default situation with which we start is the situation in which there are no ligands at all, i.e. in which $x_{ij} = d_{i0j}$ for all i and j. Other situations of interest are reasonably close to this one. Thus, we can expand the dependence (7) in Taylor series in the vicinity of the values d_{i0j}. As a result, we obtain the following expression:

$$y = y_0 + \sum_{i=1}^{n} \sum_{j=1}^{N} y_{ij} \cdot \Delta x_{ij} + \sum_{i=1}^{n} \sum_{j=1}^{N} \sum_{i'=1}^{n} \sum_{j'=1}^{N} y_{ij,i'j'} \cdot \Delta x_{ij} \cdot \Delta x_{i'j'}, \tag{8}$$

where $\Delta x_{ij} \stackrel{\text{def}}{=} x_{ij} - d_{i0j}$, and y_0, y_{ij}, and $y_{ij,i'j'}$ are appropriate coefficients.

These formulas can be applied to all possible situations, in which at each location i, different parameters $x_{i1}, \ldots, x_{iN}$ can change independently. Situations in which we are interested are characterized by describing, for each location, whether there is a ligand or not, and if yes, which exactly ligand. 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 $\varepsilon_{ik} = 1$.

By definition, at each location, there can be only one ligand, i.e., if $\varepsilon_{ik} = 1$ for some k, then $\varepsilon_{ik'} = 0$ for all $k' \neq k$.

According to the formula (8), the value y of the desired physical quantity depends on the differences Δx_{ij} corresponding to different i and j. Let us describe the values of these differences in terms of the discrete variables ε_{ik}.

- In the absence of a ligand, when $\varepsilon_i = 0$, the value of the quantity x_{ij} is equal to d_{i0j} and thus, the difference Δx_{ij} is equal to
$$\Delta x_{ij} = d_{i0j} - d_{i0j} = 0.$$
- In the presence of a ligand of type k, when $\varepsilon_{ik} = 1$, the value of the quantity x_{ij} is equal to d_{ikj} and thus, the difference $\Delta x_{ij} = d_{ikj} - d_{i0j}$ is equal to
$$\Delta_{ikj} \stackrel{\text{def}}{=} d_{ikj} - d_{i0j}.$$

Taking into account that for each location i, only one value ε_{ik} can be equal to 1, we can combine the above two cases into a single expression

$$\Delta x_{ij} = \sum_{k=1}^{m} \varepsilon_{ik} \cdot \Delta_{ikj}. \tag{9}$$

Substituting the expression (9) into the expression (8), we obtain an expression which is quadratic in ε_{ik}:

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

$$\sum_{i=1}^{n}\sum_{k=1}^{m}\sum_{j=1}^{N}\sum_{i'=1}^{n}\sum_{k'=1}^{m}\sum_{j'=1}^{N} y_{ij,i'j'} \cdot \varepsilon_{ik} \cdot \varepsilon_{i'k'} \cdot \Delta_{ikj} \cdot \Delta_{i'k'j'}, \tag{10}$$

i.e., equivalently,

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

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

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

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

$$\sum_{i=1}^{n}\sum_{k=1}^{m}\sum_{i'=1}^{n}\sum_{k'=1}^{m} a_{ik,i'k'} \cdot \varepsilon_{ik} \cdot \varepsilon_{i'k'}, \tag{11}$$

where

$$a_{ik} = \sum_{j=1}^{N} y_{ij} \cdot \Delta_{ikj}, \tag{12}$$

and

$$a_{ik,i'k'} = \sum_{j=1}^{N}\sum_{j'=1}^{N} y_{ij,i'j'} \cdot \Delta_{ikj} \cdot \Delta_{i'k'j'}. \tag{13}$$

The expression (11) is similar to the continuous Taylor expression (4), but with the discrete variables $\varepsilon_{ik} \in \{0, 1\}$ instead of the continuous variables Δx_i.

Similar "discrete Taylor series" can be derived if we take into account cubic, quartic, etc., terms in the original Taylor expansion of the dependence (7).

Discrete Taylor expansions can be further simplified. In the following text, we will use the fact that the expression (11) can be further simplified.

First, we can simplify the terms corresponding to $i = i'$. Indeed, 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 $a_{ik,ik} \cdot \varepsilon_{ik}$ and can, therefore, be simply added to the corresponding linear term $a_{ik} \cdot \varepsilon_{ik}$.

Similarly, for every location i and for every two ligand types $k \neq k'$, only one of the terms ε_{ik} and $\varepsilon_{ik'}$ can be different from 0. Thus, the product $\varepsilon_{ik} \cdot \varepsilon_{ik'}$ is always equal to 0. Therefore, we can safely assume that the coefficient $a_{ik,ik'}$ at this product is 0.

Thus, we have no terms $a_{ik,i'k'}$ corresponding to $i = i'$ in our formula, we only have terms with $i \neq i'$. For each two pairs ik and $i'k'$, we can combine terms proportional to $\varepsilon_{ik} \cdot \varepsilon_{i'k'}$ and to $\varepsilon_{i'k'} \cdot \varepsilon_{ik}$. As a result, we obtain a simplified expression

$$y = v_0 + \sum_{i=1}^{n}\sum_{k=1}^{m} v_{ik} \cdot \varepsilon_{ik} + \sum_{i<i'}\sum_{k=1}^{m}\sum_{k'=1}^{m} v_{ik,i'k'} \cdot \varepsilon_{ik} \cdot \varepsilon_{i'k'}, \tag{14}$$

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

This expression (14) – and the corresponding similar cubic and higher order expressions – is what we will understand by a discrete Taylor series.

What we will do in the following text. As we have mentioned earlier, we will show that the poset-related approaches are, in effect, equivalent to the use of a much simpler (and much more familiar) tool of (discrete) Taylor series.

IV. Equivalence Between the Poset-Related Approaches and the Discrete Taylor Series Approach

Discrete Taylor series: reminder. In many practical situations, we have a physical variable y that depends on the discrete parameters ε_{ik} which take two possible values: 0 and 1, and for which, for every i, at most one value ε_{ik} can be equal to 1. Then, in the first approximation, the dependence of y on ε_{ik} can be described by the following linear formula

$$y = v_0 + \sum_{i=1}^{n}\sum_{k=1}^{m} v_{ik} \cdot \varepsilon_{ik}. \tag{15}$$

In the second approximation, this dependence can be described by the following quadratic formula

$$y = v_0 + \sum_{i=1}^{n}\sum_{k=1}^{m} v_{ik}\cdot\varepsilon_{ik} + \sum_{i<i'}\sum_{k=1}^{m}\sum_{k'=1}^{m} v_{ik,i'k'}\cdot\varepsilon_{ik}\cdot\varepsilon_{i'k'}. \quad (16)$$

etc.

Chemical substances. For chemical substances, we have discrete variables ε_{ik} that describe whether there is a ligand of type k at the i-th location:

- the value $\varepsilon_{ik} = 0$ means that there is no ligand of type k at the i-th location, and
- the value $\varepsilon_{ik} = 1$ means that there is a ligand of type k at the i-th location.

Each chemical substance a from the corresponding family can be characterized by the corresponding tuple

$$(\varepsilon_{11}, \ldots, \varepsilon_{1m}, \ldots, \varepsilon_{n1}, \ldots, \varepsilon_{nm}).$$

Poset-related approaches: reminder. We approximate the actual dependence of the desired quantity y on the substance $a = (\varepsilon_{11}, \ldots, \varepsilon_{nm})$ by a formula

$$v(a) = \sum_{b:\, b\le a} V(b), \quad (17)$$

where, in the second order approximation, b runs over all substances with at most two ligands.

Poset-related approaches reformulated in terms of the discrete variables. The discrete Taylor series formula (16) is formulated in terms of the discrete variables ε_{ik}. Thus, to show the equivalence of these two approaches, let us first describe the poset-related formula (17) in terms of these discrete variables.

In chemical terms, the relation $b \le a$ means that a can be obtained from b by adding some ligands. In other words, the corresponding value ε_{ik} can only increase when we move from the substance b to the substance a. So, if $b = (\varepsilon'_{11}, \ldots, \varepsilon'_{nm})$ and $a = (\varepsilon_{11}, \ldots, \varepsilon_{nm})$, then $b \le a$ means that for every i and k, we have $\varepsilon'_{ik} \le \varepsilon_{ik}$.

Thus, the formula (17) means that for every substance $a = (\varepsilon_{11}, \ldots, \varepsilon_{nm})$, the substances $b \le a$ are:

- the original substance $a_0 = (0, \ldots, 0)$;
- substances $a_{ik} \stackrel{\text{def}}{=} (0, \ldots, 0, 1, 0, \ldots, 0)$ with a single ligand of type k at the location i – corresponding to all the places i and types k for which $\varepsilon_{ik} = 1$; and
- substances $a_{ik,i'k'} \stackrel{\text{def}}{=} (0, \ldots, 0, 1, 0, \ldots, 0, 1, 0, \ldots, 0)$ with a ligand of type k at the locations i and a ligand of type k' at a location i' – corresponding to all possible pairs (i, k) and (i', k'), $i < i'$, for which $\varepsilon_{ik} = \varepsilon_{i'k'} = 1$.

Thus, in terms of the discrete variables, the poset formula (17) takes the form

$$y = V(a_0) + \sum_{(i,k):\, \varepsilon_{ik}=1} V(a_{ik}) + \sum_{i<i',k,k':\, \varepsilon_{ik}=\varepsilon_{i'k'}=1} V(a_{ik,i'k'}). \quad (18)$$

Proof that the discrete Taylor series are indeed equivalent to the poset formula. The formulas (16) and (18) are now very similar, so we are ready to prove that they actually coincide.

To show that these formulas are equal, let us take into account that, e.g. the linear part of the sum (18) can be represented as

$$\sum_{(i,k):\, \varepsilon_{ik}=1} V(a_{ik}) = \sum_{(i,k):\, \varepsilon_{ik}=1} V(a_{ik})\cdot\varepsilon_{ik}. \quad (19)$$

Indeed, for all the corresponding pairs (i, k), we have $\varepsilon_{ik} = 1$, and multiplying by 1 does not change a number.

This new representation (19) allows us to simplify this formula by adding similar terms $V(a_{ik})\cdot\varepsilon_{ik}$ corresponding to pairs (i, k) for which $\varepsilon_{ik} = 0$. Indeed, when $\varepsilon_{ik} = 0$, then the product $V(a_{ik})\cdot\varepsilon_{ik}$ is equal to 0, and thus, adding this product will not change the value of the sum. So, in the right-hand side of the formula (19), we can safely replace the sum over all pairs (i, k) for which $\varepsilon_{ik} = 1$ by the sum over all pairs (i, k):

$$\sum_{(i,k):\, \varepsilon_{ik}=1} V(a_{ik}) = \sum_{i=1}^{n}\sum_{k=1}^{m} V(a_{ik})\cdot\varepsilon_{ik}. \quad (20)$$

Similarly, the quadratic part $\sum_{i<i',k,k':\, \varepsilon_{ik}=\varepsilon_{i'k'}=1} V(a_{ik,i'k'})$ of the sum (18) can be first replaced with the sum

$$\sum_{i<i',k,k':\, \varepsilon_{ik}=\varepsilon_{i'k'}=1} V(a_{ik,i'k'}) = \sum_{i<i',k,k':\, \varepsilon_{ik}=\varepsilon_{i'k'}=1} V(a_{ik,i'k'})\cdot\varepsilon_{ik}\cdot\varepsilon_{i'k'}, \quad (21)$$

and then, by the sum

$$\sum_{i<i',k,k':\, \varepsilon_{ik}=\varepsilon_{i'k'}=1} V(a_{ik,i'k'}) = \sum_{i<i'}\sum_{k=1}^{m}\sum_{k'=1}^{m} V(a_{ik,i'k'})\cdot\varepsilon_{ik}\cdot\varepsilon_{i'k'}. \quad (22)$$

Substituting expressions (19) and (22) into the formula (18), we obtain the following expression

$$y = V(a_0) + \sum_{i=1}^{n} V(a_{ik})\cdot\varepsilon_{ik} + \sum_{i<i'}\sum_{k=1}^{m}\sum_{k'=1}^{m} V(a_{ik,i'k'})\cdot\varepsilon_{ik}\cdot\varepsilon_{i'k'}. \quad (23)$$

This expression is identical to the discrete Taylor formula (16), the only difference is the names of the corresponding parameters:

- the parameter v_0 in the formula (16) corresponds to the parameter $V(a_0)$ in the formula (23);

- each parameter v_{ik} in the formula (16) corresponds to the parameter $V(a_{ik})$ in the formula (23); and
- each parameter $v_{ik,i'k'}$ in the formula (16) corresponds to the parameter $V(a_{ik,i'k'})$ in the formula (23).

The equivalence is proven.

V. Conclusion

Several practically useful chemical substances can be obtained by adding ligands to different locations of a "template" molecule like benzene C_6H_6 or cubane C_8H_8. There is a large number of such substances, and it is difficult to synthesize all of them and experimentally determine their properties. It is desirable to be able to synthesize and test only a few of these substances and to use appropriate interpolation to predict the properties of others.

It is known that such an interpolation can be obtained by using Rota's ideas related to partially ordered sets. In our previous paper, we have shown that when we only allow one type of ligand, then the exact same interpolation algorithm can be obtained from a more familiar mathematical technique such as Taylor expansion series. In this paper, we show that the similar equivalence holds in the general case, when we have ligands of different type.

Acknowledgments

This work was supported in part by NSF grant HRD-0734825 and grant 1 T36 GM078000-01 from the National Institutes of Health.

The author would like to thank Dr. Vladik Kreinovich, for his encouragement, to all the participants of the 65th Southwest Regional Meeting of the American Chemical Society, El Paso, Texas, November 4–7, 2009, especially to James Salvador, for valuable discussions, and to the anonymous referees for important suggestions.

References

[1] T. Došlić and D. J. Klein, Splinoid interpolation on finite posets, *Journal of Computational and Applied Mathematics*, vol.177, pp.175–185, 2005.

[2] R. Feynman, R. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Addison Wesley, Boston, Massachusetts, 2005.

[3] T. Ivanciuc, O. Ivanciuc, and D. J. Klein, Posetic quantitative super-structure/activity relationships (QSSARs) for chlorobenzenes *Journal of Chemical Information and Modeling*, vol.45, pp.870–879, 2005.

[4] T. Ivanciuc, O. Ivanciuc, and D. J. Klein, Modeling the bioconcentration factors and bioaccumulation factors of polychlorinated biphenyls with posetic quantitative super-structure/activity relationships (QSSAR), *Molecular Diversity*, vol.10, pp.133–145, 2006.

[5] T. Ivanciuc and D. J. Klein, Parameter-free structure-property correlation via progressive reaction posets for substituted benzenes, *Journal of Chemical Information and Computer Sciences*, vol.44, no.2, pp.610–617, 2004.

[6] T. Ivanciuc, D. J. Klein, and O. Ivanciuc, Posetic cluster expansion for substitution-reaction networks and application to methylated cyclobutanes, *Journal of Mathematical Chemistry*, vol.41, no.4, pp.355–379, 2007.

[7] D. J. Klein, Chemical graph-theoretic cluster expansions, *International Journal of Quantum Chemistry, Quantum Chemistry Symposium*, vol.20, pp.153–171, 1986.

[8] D. J. Klein and L. Bytautas, Directed reaction graphs as posets, *MATCH Communications in Mathematical and in Computer Chemistry (MATCH)*, vol.42, pp.261–290, 2000.

[9] J. Nava, V. Kreinovich, G. Restrepo, and D. J. Klein, Discrete Taylor Series as a Simple Way to Predict Properties of Chemical Substances like Benzenes and Cubanes, *Journal of Uncertain Systems*, 2010, Vol. 4, No. 4, pp. 270–290.

[10] G.-C. Rota, On the foundations of combinatorial theory I. Theory of Möbius functions, *Zeit. Wahrscheinlichkeitstheorie*, vol.2, pp.340–368, 1964.

[11] D. Sheskin, *Handbook of parametric and nonparametric statistical procedures*, Chapman & Hall/CRC, Boca Raton, Florida, 2004.

[12] R. R. Yager and L. Liu, Eds., *Classic Works of the DempsterShafer Theory of Belief Functions*, Springer, Berlin, 2008 (Studies in Fuzziness and Soft Computing, Vol. 219).