IsoDittwald
Law of Localised Mass Spec Fine Structure

Mateusz Łącki* and Anna Gambin

Faculty of Mathematics, Informatics and Mechanics
University of Warsaw, Banacha 2, 02-097 Warszawa, Poland
mateusz.lacki@biol.uw.edu.pl
aniag@mimuw.edu.pl
http://bioputer.mimuw.edu.pl

Abstract. Approximative distributions theory is used to obtain more tractable formulas describing the localised fine structure of isotopic peaks. We present a new method for calculating localised fine structure isotopic peaks based on the above-mentioned approximations. *abstract* environment.

Keywords: Isotopic Fine Structure, Poisson Approximation, Little Sexy Fox

1 Introduction

There are many reasons why mass spectrometry analysis is hard. It is hard in that there are potentially many many sources of interferences that can distort the information about the actual composition of a sample. The study of the nature of these interferences is needed to achieve the goal of making out of mass spectrometers yet more reliable an identification tool.

Part of the noise in the mass to charge domain is innately related to the elements themeselves and stems from the existence of isotopes. It is because of them that a given analyte is represented as a series of peaks, a spectrum, rather that only one peak. The theoretical underpinnings of how to mathematically model the impact of isotopes are already well established, see [18]. The main idea behind the model is to abstract from the exact positionings of the extra neutrons on a particular chemical compound and thus concentrate only on their relative amounts among all atoms of a given element. Assuming that the isotopic configurations are independent and follow the element dependent distribution, one arrives to the conclusion that the correct law describing occurence of different isotopes in a chemical compound is the product of multinomial distributions.

There is one huge problem with that law: together with the growth of molecule one observes an exponential growth in the number of possibile isotope configurations, which precludes their direct enumeration. To solve this problem, different

* Special thanks to Santa Claus.

semplifications were proposed, amounting to different ways of binning configurations together explicitly [4], by hiding them under the guise of Fourier Transform [13], or by ...

However, it is considered of paramount importance in Mass Spectrometry to develop machines with still higher resolution powers and it is very likely that this trend will continue. Even today there are machines that already can distinguish peaks attributed to different configurations with the same number of extra neutrons.

Add Olson and others but Olson above all.

Here we propose to approach the problem of fine structure so that it overcomes the shortcomings of the aggregate model, as used in [4]. That particular model bins together configurations having the same number of extra neutrons distributed on different atoms. For instance, if one considers water molecule H_2O, the model would glue together configuration with one extra neutron only on the first hydrogen togeter with that having it on the second together with that on oxygen atom. We devise an algorithm to deaggregate these probability clusters. We call the peaks obtained via that algorithm a *localised fine structure*.

What motivates the solution to this problem is a search for better molecule fingerprints. The development of new mass spectrometers capable of distinguishing differences in masses of neutrons is proceeding at a vigorous pace. Soon, scientists will face the need of more detailed models than those abstracting from mass defects. It is also common for chemists to search for presence of specific substance in the sample. Usually this is done by looking at some highly specific range in the mass domain of the gathered spectra. Our model provides deeper insight to what might happen while focussing on that particular bit of collected data: conditioning on configurations with the same number of extra neutrons translates directly into focussing in a specific region of the mass-to-charge domain.

The algorithm assumes that one can easily find a peak not far from the most probable one and that the distribution is close to what one would call unimodal[1] and that the most of distributions probability lies in a rather small neighbourhood of the mode. Both the guess about the starting point and the way the neighbourhood gets explored depend on the Poisson approximation to the distribution under study. To our best knowledge this type of approximation have not yet been used for algorithmic purposes. It has been used however in the context of proteomic and peptide research: in [3] it is being used for high throughput protein identification; its use was revalidated in [17] in case of peptides.

We also observe that the use of Poisson approximation gives a theoretical explanation for the equatransneutronic binning used in [12] and actually helps deaggregating results obtained using that approach as well.

Diophantine equations.

[1] We provide a precise definition of unimodality for discrete probability distributions in Section ...

2 Approximations

By an isotopic configuration we understand information about the number of different isotopes in the sample. For the purpose of simplicity, we focus here on chemical compounds composed of carbon, hydrogen, nitrogen, oxygen, and sulfur; still, the results of this section generalize to any compounds whatsoever. Thus, we concentrate on compounds like $\mathrm{C_cH_hO_oN_nS_s}$, where the low case letters describe the numbers of atoms of particular element type. Among such compounds one can already find peptides and proteins. An isotopic configuration could be represented by an extended empirical formula,

$$ {}^{12}\mathrm{C}_{\mathrm{c}_0}\ {}^{13}\mathrm{C}_{\mathrm{c}_1}\ {}^{1}\mathrm{H}_{\mathrm{h}_0}\ {}^{2}\mathrm{H}_{\mathrm{h}_1}\ {}^{14}\mathrm{N}_{\mathrm{n}_0}\ {}^{15}\mathrm{N}_{\mathrm{n}_1}\ {}^{16}\mathrm{O}_{\mathrm{o}_0}\ {}^{17}\mathrm{O}_{\mathrm{o}_1}\ {}^{18}\mathrm{O}_{\mathrm{o}_2}\ {}^{32}\mathrm{S}_{\mathrm{s}_0}\ {}^{33}\mathrm{S}_{\mathrm{s}_1}\ {}^{34}\mathrm{S}_{\mathrm{s}_2}\ {}^{36}\mathrm{S}_{\mathrm{s}_4}. \quad (1) $$

In the above representation, small letters with indices represent counts of different atoms with indices displaying the number of additional neutrons an isotope has with respect to the lighest possible isotopic variant.

Rather than (1), we shall be using an equivalent probabilistic notation, treating upper case letters, like ${}^{12}\mathrm{C}$, as random variables and considering small case letters, c_0, to be their realisations. An expression like $A = \{{}^{13}\mathrm{C} = \mathrm{c}_1,\ {}^{2}\mathrm{H} = \mathrm{h}_1\}$ is shorthand for saying: let us focus on all configurations (1) that have c_1 heavy carbons and h_1 deuters in total.

Following [10], one assumes that the law of vector

$$ \left({}^{12}\mathrm{C},\ {}^{13}\mathrm{C},\ {}^{1}\mathrm{H},\ {}^{2}\mathrm{H},\ {}^{14}\mathrm{N},\ {}^{15}\mathrm{N},\ {}^{16}\mathrm{O},\ {}^{17}\mathrm{O},\ {}^{18}\mathrm{O},\ {}^{32}\mathrm{S},\ {}^{33}\mathrm{S},\ {}^{34}\mathrm{S},\ {}^{36}\mathrm{S}\right), \quad (2) $$

given $\mathrm{C_cH_hO_oN_nS_s}$, is a product of independent multinomial distributions,

$$ \mathbb{M} = \mathrm{Multi}\Big(\mathbb{P}({}^{12}\mathrm{C}), \mathbb{P}({}^{13}\mathrm{C}); c\Big) \otimes \cdots \otimes \mathrm{Multi}\Big(\mathbb{P}({}^{32}\mathrm{S}), \mathbb{P}({}^{33}\mathrm{S}), \mathbb{P}({}^{34}\mathrm{S}), \mathbb{P}({}^{36}\mathrm{S}); s\Big), \quad (3) $$

where the probabilities of observing particular isotopes, $\mathbb{P}({}^{12}\mathrm{C}), \ldots, \mathbb{P}({}^{36}\mathrm{S})$, are established in independent experiments[2]. For instance, the probability of a given carbons configuration $(\mathrm{c}_0, \mathrm{c}_1)$ equals

$$ \mathrm{Multi}\left(\mathbb{P}({}^{12}\mathrm{C}), \mathbb{P}({}^{13}\mathrm{C}); c\right)\Big((\mathrm{c}_0, \mathrm{c}_1)\Big) = \binom{\mathrm{c}}{\mathrm{c}_0, \mathrm{c}_1} \mathbb{P}({}^{12}\mathrm{C})^{\mathrm{c}_0} \mathbb{P}({}^{13}\mathrm{C})^{\mathrm{c}_1} $$

and it should be multiplied by similar expression for hydrogen, nitrogen, oxygen and sulfur to obtain probability for expression like (1).

Observe, that given $\mathrm{C_cH_hO_oN_nS_s}$, part of the information in (2) is redundant and can be shortened by neglecting counts of the lightest isotope variants, leaving us with

$$ \left({}^{13}\mathrm{C},\ {}^{2}\mathrm{H},\ {}^{15}\mathrm{N},\ {}^{17}\mathrm{O},\ {}^{18}\mathrm{O},\ {}^{33}\mathrm{S},\ {}^{34}\mathrm{S},\ {}^{36}\mathrm{S}\right). \quad (4) $$

Missing therms can be retrieved from relationships ${}^{12}\mathrm{C} + {}^{13}\mathrm{C} = \mathrm{c}$, ${}^{1}\mathrm{H} + {}^{2}\mathrm{H} = \mathrm{h}$, and so on, that occur with probability one.

[2] Consult Table 1 for details.

Definition 1 *We call the set of configurations*

$$LFS_K = \left\{ {}^{13}C + {}^{2}H + {}^{15}N + {}^{17}O + 2\times{}^{18}O + {}^{33}S + 2\times{}^{34}S + 4\times{}^{36}S = K \right\} \tag{5}$$

a localised fine structure with K extra neutrons.

The reason for numbers 2 and 4 appearing above is that ^{18}O and ^{34}S have two additional neutrons, and ^{36}S – four; confront Table 1.

The problem of finding the cardinal number of LFS_K is also known as the money exchange problem. In general, enumeration of all elements of LFS_K corresponds to finding all integer solutions $(x_1, \ldots, x_k)$ to a *Linear Diophantine Equation*

$$a_1 x_1 + \cdots + a_k x_k = K, \tag{6}$$

where $(a_1, \ldots, a_k)$ are integer coefficients. According to [1], if the greatest common divisor of $(a_1, \ldots, a_k)$ is equal to one, then the number of solutions to (6) is approximately $\frac{K^{k-1}}{(k-1)!a_1 \ldots a_k}$. Since carbon has only one addtional isotope, then the above estimate encompasses all organic chemistry.

The problem is therefore polynomial in K, making it big. However, configurations in LFS_K are naturally prioritized by probability (3) and, rather than enumerating all of them, one would be satisfied by only the most probable ones.

Problem 1 *For a given K, find a small set $B \subset LFS_K$ of configurations s.t.*

$$\mathbb{M}_K(B) := \frac{\mathbb{M}(B)}{\mathbb{M}(LFS_K)} \approx 1\,, \tag{7}$$

where $\mathbb{M}_K$ is the product of multinomial laws (3) *conditional on the set of configurations in LFS_K and is refered to as* The Law of Localised Fine Structure.

In statistical terms, we are interested in approximating some critical set of large probability, as measured by the *Law of Localised Fine Structure*.

Why should one study law described by (7) in the first place? Simply because the masses of different configurations in LFS_K concentrate around the compound's monoisotopic mass shifted to the right by K Daltons[3]. For medium sized compounds, LFS_K's for different K should in principle form disjoint clusters in the mass to charge domain, as argued in the yet unpublished [5]; for bigger compounds some interference would be expected, but in that case one would simply study three or more consecutive sets of configurations, e.g. LFS_{K-1}, LFS_K, and LFS_{K+1}. All in all, by studying LFS_K we get a guarantee to explore thoroughly a precised place in the mass to charge domain. More reasons behind this idea are exposed in the **Conclusions and Discussion** section.

How to quote that? It is yet unpublished.

The to solve Problem 1 we approximate measure $\mathbb{M}_K$ by a more analytically tractable measure $\mathbb{Q}_K$ defined on the LFS_K. We then devise an algorithm to find a possibly small set of configurations $B^* \subset LFS_K$, s.t. $\mathbb{Q}_K(B^*) \approx 1$. Since $\mathbb{Q}_K \approx \mathbb{M}_K$, so $\mathbb{M}_K(B^*) \approx 1$ and B^* solves Problem 1, possibly suboptimally.

[3] For underlying physical principles consult [8].

A natural way to define proper $\mathbb{Q}_K$ is to first approximate $\mathbb{M}$ by some $\mathbb{Q}$ and then pose $\mathbb{Q}_K(\circ) := \frac{\mathbb{Q}(\circ \cap LFS_K)}{\mathbb{Q}(LFS_K)}$, i.e. condition $\mathbb{Q}$ on the occurence of configurations from LFS_K. To prove it works, we have to first mention, that by approximation we understand convergence in distribution, as described in [9]. Then, we make use of the following lemma:

Lemma 1. *Let $\mu^{[n]}, \mu$ be discrete measures. If $\mu^{[n]}$ converges in distribution to μ, $\mu^{[n]} \rightharpoonup \mu$, and an event A has nonzero probability under any of that measures, $\underset{n}{\forall} \mu^{[n]}(A)\,,\ \mu(A) > 0$, then measures conditioned by A, $\mu_A^{[n]}(\circ) := \frac{\mu^{[n]}(\circ \cap A)}{\mu^{[n]}(A)}$ converge in distribution to $\mu_A(\circ) := \frac{\mu(\circ \cap A)}{\mu(A)}$; or $\mu_A^{[n]} \rightharpoonup \mu_A$ for short.*

The proof is exposed in the **Appendix**.

Let us now unveil the usefulness of Lemma 1. There is an entire family of measures mentioned in it, $\mu^{[n]}$. We assume, that one of them is simply our initial meausure: there exists n^* s.t. $\mathbb{M} = \mu^{[n^*]}$. Also, we assume the approximation of $\mu^{[n^*]}$ by measure μ is already o good one. Our choice for μ is to be the product of independent Poisson measures, which is stimulated by the following, well known lemma[4].

Lemma 2. *If all $\lim_{n\to\infty} np_{k,n} = \lambda_k$ exist for $k \in \{1, \ldots, w\}$, then*

$$\mathrm{Multi}\left(p_0^{[n]}, p_1^{[n]}, \ldots, p_w^{[n]}; n\right) \rightharpoonup \mathrm{Poiss}(\lambda_1) \otimes \cdots \otimes \mathrm{Poiss}(\lambda_w), \tag{8}$$

where Poiss *stands for the Poisson distribution,* $\mathrm{Poiss}(\lambda)(k) = \frac{\lambda^k}{k!} e^{-\lambda}$.

In Lemma 2 one assumes that the number of trials n goes to infinity. In our model this corresponds to an infinite enlargement of the compound. The existence of limits assumes that this enlargement is done so that on such an idealized compound only the lightest isotopes would appear infinitely often. Moreover, since the support of any Poisson distibution is equal to the set of all integer numbers, the state space of configurations gets significantly enlarged and contains configurations that are nonphysical for any real chemical compound. For instance, positive probabilities would be prescribed to configurations with numbers of isotopes greater then the number of possible places for them on any finite compound[5]. Observe also, that the probabilities $p_k^{[n]}$ are pending towards zero: for good approximation one would expect therefore the probabilities of observing heavier isotopes, e.g. quantities like $\mathbb{P}(^{13}\mathrm{C}), \mathbb{P}(^{2}\mathrm{H}), \ldots, \mathbb{P}(^{36}\mathrm{S})$, to be relatively small. Confront Table 1 to check, that this is really the case.

Observe, that Lemma 2 defines a proper limit for just one multinomial distribution, whereas $\mathbb{M}$ is a product thereof. The problem is other than what to do

[4] The proof is common knowledge and we omit it.

[5] There is no mathematical incongruence here, however, since the approximation assumes that the limiting compound is of infinite size. For mathematical correctness we also note, that we can transfer virtually any initial measure $\mathbb{M}$ on the enlarged state space, i.e. where the approximation is defined, simply by assuming, that $\mathbb{M}$ measure on any nonphysical configuration equals zero.

with products: one can approximate independently each multinomial. However, the quality of such approximation would depend on all the counts of different elements in a molecule. For instance, in case of $\mathrm{C_cH_hO_oN_nS_s}$ the better the approximation[6] the bigger the smallest among numbers $(\mathrm{c}, \mathrm{h}, \mathrm{n}, \mathrm{o}, \mathrm{s})$. Due to the polymer structure, one would expect some more information could be reveiled on that matter for proteins and peptides. Indeed, for such chain structures Senko et al. [16], basing on empirical research, introduced the concept of avergine, i.e. an averaged protein: any protein composed of m amino acids should have its mass approximately equal to the mass of the idealised compound

$$\mathrm{C}_{\lfloor m\times 4.9384\rfloor}\mathrm{H}_{\lfloor m\times 7.7583\rfloor}\mathrm{O}_{\lfloor m\times 1.4773\rfloor}\mathrm{N}_{\lfloor m\times 1.3577\rfloor}\mathrm{S}_{\lfloor m\times 0.0417\rfloor}.$$

The weakest link in the approximation might result from small numbers of sulfur. This is an acknowledged problem in empirical studies, as exposed in [17]. The longer the polymers however, the smaller the differences should be.

The final questions is: what values should be used as λ's in Lemma 2? We *calibrate* those values by equating them to the averages of the multinomial distributions from (3)[7]: in case of carbon we set $\lambda_{{}^{13}\mathrm{C}} \approx \mathrm{c} \times \mathbb{P}({}^{13}\mathrm{C})$. In contrast to our method, λ's in [3, 17] are chosen to be the minimisers in a free parameter optimisation scheme with χ^2 penalty[8].

At days end, the probability assigned to event

$$\left\{{}^{13}\mathrm{C} = \mathrm{c}_1,\ {}^{2}\mathrm{H} = \mathrm{h}_1,\ {}^{15}\mathrm{N} = \mathrm{n}_1,\ {}^{17}\mathrm{O} = \mathrm{o}_1,\ {}^{18}\mathrm{O} = \mathrm{o}_2,\ {}^{33}\mathrm{S} = \mathrm{s}_1,\ {}^{34}\mathrm{S} = \mathrm{s}_2,\ {}^{36}\mathrm{S} = \mathrm{s}_4\right\}$$

is given by

$$\frac{\lambda_{{}^{13}\mathrm{C}}^{\mathrm{c}_1}}{\mathrm{c}_1!}\frac{\lambda_{{}^{2}\mathrm{H}}^{\mathrm{h}_1}}{\mathrm{h}_1!}\frac{\lambda_{{}^{15}\mathrm{N}}^{\mathrm{n}_1}}{\mathrm{n}_1!}\frac{\lambda_{{}^{17}\mathrm{O}}^{\mathrm{o}_1}}{\mathrm{o}_1!}\frac{\lambda_{{}^{33}\mathrm{S}}^{\mathrm{s}_1}}{\mathrm{s}_1!}e^{-\mu}\frac{\lambda_{{}^{18}\mathrm{O}}^{\mathrm{o}_2}}{\mathrm{o}_2!}\frac{\lambda_{{}^{34}\mathrm{S}}^{\mathrm{s}_2}}{\mathrm{s}_2!}e^{-\eta}\frac{\lambda_{{}^{36}\mathrm{S}}^{\mathrm{s}_1}}{\mathrm{s}_1!}e^{-\gamma}, \tag{9}$$

where

$$\begin{aligned}\mu &= \lambda_{{}^{13}\mathrm{C}} + \lambda_{{}^{2}\mathrm{H}} + \lambda_{{}^{15}\mathrm{N}} + \lambda_{{}^{17}\mathrm{O}} + \lambda_{{}^{33}\mathrm{S}}\\ \eta &= \lambda_{{}^{18}\mathrm{O}} + \lambda_{{}^{34}\mathrm{S}}\\ \gamma &= \lambda_{{}^{36}\mathrm{S}}.\end{aligned}$$

The usefulness of approximation by a product of independent Poisson lies in two important properties, as summarised in the following lemmas.

Lemma 3. *Suppose we have a collection of m independent Poisson-distributed random variables, $X_i \sim \mathrm{Poiss}(\kappa_i)$. Then $X_1 + \cdots + X_m \sim \mathrm{Poiss}(\kappa_1 + \cdots + \kappa_m)$.*

[6] The *goodness* of approximation is expressed in the total variance distance; see [14].

[7] Such *calibration* is common practice in statistics.

[8] Note however, that these two solutions should not differ too much for larger compounds, for it is known that both the Poisson and Multinomial distributions are concentrated near their means, see [2].

Lemma 4. *Suppose we have a collection of m independent Poisson-distributed random variables, $X_i \sim \mathrm{Poiss}(\kappa_i)$. Then $X_1, \ldots, X_m$ given that $X_1 + \cdots + X_m = K$ is multinomially distributed,*

$$\Big(X_1, \ldots, X_m | X_1 + \cdots + X_m = K\Big) \sim \mathrm{Multi}\Big(\frac{\kappa_1}{\sigma}, \ldots, \frac{\kappa_m}{\sigma}; K\Big),$$

where $\sigma = \sum_{i=1}^{m} \kappa_i$.

Both lemmas are proved in [11]. Lemma 3 shows how to semplify calculations for a Diophantine equations with all parameters set to one, $a_i \equiv 1$. Lemma 4 describes the law resulting from conditioning independent Poisson variables by such an expression.

Suppose that we concentrated on molecules composed entirely of elements that can have only one additional neutron, e.g. $\mathrm{C_cH_hN_n}$. By Lemma 4 we get:

Result 1 *For $C_cH_hN_n$, let $\tilde{\mu} := \lambda_{^{13}C} + \lambda_{^{2}H} + \lambda_{^{15}N}$. Then*

$$\mathbb{Q}_K = \mathrm{Multi}\left(\frac{\lambda_{^{13}C}}{\tilde{\mu}}, \frac{\lambda_{^{2}H}}{\tilde{\mu}}, \frac{\lambda_{^{15}N}}{\tilde{\mu}}; K\right).$$

Proof. The corresponding Diophantine equation is $^{13}\mathrm{C} + {}^{2}\mathrm{H} + {}^{15}\mathrm{N}$.

It is valuable to see, how Lemma 4 generalizes while conditioning on a more complex Diophantie equation. Observe, that (5) can be rewritten as

$$LFS_K = \Big\{ \underbrace{{}^{13}\mathrm{C} + {}^{2}\mathrm{H} + {}^{15}\mathrm{N} + {}^{17}\mathrm{O} + {}^{33}\mathrm{S}}_{G_1} + 2 \times \underbrace{({}^{18}\mathrm{O} + {}^{34}\mathrm{S})}_{G_2} + 4 \times \underbrace{{}^{36}\mathrm{S}}_{G_4} = K \Big\},$$

so that in light of Lemma 3, $\mathbb{Q}(A)$ can be calculated in an easier way:

$$\mathbb{Q}(LFS_K) = \sum_{k_1+2k_2+4k_4=K} \mathbb{P}(G_1 = k_1, G_2 = k_2, G_4 = k_4),$$

where $G_1 \sim \mathrm{Poiss}(\mu)$, $G_2 \sim \mathrm{Poiss}(\eta)$, and $G_4 \sim \mathrm{Poiss}(\gamma)$ are mutually independent. There is a strict link between G_i and the concept of *equatransneutronic groups* described in [12]: it is equal to the total number of atoms bearing exactly i additional neutrons.

To calculate $\mathbb{Q}_K$ it remains to divide (9) by $\mathbb{Q}(LFS_K)$. Observe however that a more significant expression is to be obtained, if additionally we multiply both the nominator and the denominator of that expression by $\frac{\mu^{k_1}}{k_1!}\frac{\eta^{k_2}}{k_2!}\frac{\gamma^{k_4}}{k_4!}$:

Result 2 *The approximative* fine structure law *with K additional neutrons for $C_cH_hO_oN_nS_s$ is equal to*

$$\mathrm{Multi}\left(\frac{\lambda_{^{13}C}}{\mu}, \frac{\lambda_{^{2}H}}{\mu}, \frac{\lambda_{^{15}N}}{\mu}, \frac{\lambda_{^{17}O}}{\mu}, \frac{\lambda_{^{33}S}}{\mu}; k_1\right) \otimes \mathrm{Multi}\left(\frac{\lambda_{^{18}O}}{\eta}, \frac{\lambda_{^{34}S}}{\eta}; k_2\right) \otimes \mathbb{L}(k_1, k_2, k_4),$$

where

$$\mathbb{L}(k_1, k_2, k_4) = \frac{\frac{\mu^{k_1}}{k_1!}\frac{\eta^{k_2}}{k_2!}\frac{\gamma^{k_4}}{k_4!}}{\sum_{k_1'+2k_2'+4k_4'=K} \frac{\mu^{k_1'}}{k_1'!}\frac{\eta^{k_2'}}{k_2'!}\frac{\gamma^{k_4'}}{k_4'!}}. \tag{10}$$

Otherwise stated, the approximative distribution is a mixture of independent multinomial distributions weighted by the $\mathbb{L}$ distribution, which, for lack of name, we shall call the *lucky law*. Under the Poisson approximation, the *lucky law* is the resulting law on the *equatransneutronic configurations*.

If the compound contains elements with their *additional neutron acceptances* in set $I = \{1, 2, 4\}$, formula (10) generalizes to

$$\mathbb{L}(\boldsymbol{k}) = \frac{\prod_{i\in I} \frac{\mu_i^{k_i}}{k_i!}}{\sum_{\{\boldsymbol{k}^*:\sum_{i\in I} i k_i^* = K\}} \prod_{i\in I} \frac{\mu_i^{k_i^*}}{k_i^*!}},$$

where $\boldsymbol{k}$ is an ordered tuple indexed by I. Nature poses a natural limit on the complexity of the *lucky law*, as at most $\#I \leq 10$.

Ascertain that asking Frederik.

3 Algorithms

4 Discussion and Conclusions

Novelties:

- Greedy approach based on concentration of measure considerations.
- Top-down identification compatibility.

Advantage: Say that it suits extremely well a top-down identification strategy.

One could say that Problem 1 is not the one worth studying. Assuming that the machines are already advanced enough to resolve peaks that come from different configurations with the same number of neutrons, why should one not pose the following:

Problem 2 *Find a small set C among all possible configurations, s.t. $\mathbb{M}(C) \approx 1$.*

Write about the gready approach we use to get the peaks.

Observe, that such a problem could be solved in a gready manner[9], similar to one described in the previous section, with a naturally arrising candidate for the biggest peak: the direct product of modes of each multinomial model described by (3). However, at least in case of Time of Flight analyzers, there is one potential advantage of studying the LFS_K configurations over the ones gathered in the critical set C: it is known, that in these instruments the resolution depends on the

[9] The development of such an algorithm is already proceeding

mass of analyte, see [7]. It is more difficult to differentiate correctly molecules with similar masses, when both of them are big. There is no guarantee that while proceeding with calculations of elements of C one would not obtain peaks from different regions in the mass to charge domain. Therefore, some sort of additional error term in the resolution would have to be included and it is not straightforward to model it. By studying LFS_K we neglect that sort of problem, because of the localisation in the mass to charge domain.

Observe also, that one might have though of approaching Problem 1 by using the above mentioned algorithm, or, in general, any algorithm that results in obtaining *fine structure* masses, and subsequently, by discarding all configurations with masses outside a prespecified interval. This might, however, tremendously prolongue the algorithm's operating time and additional considerations to bypass this problem would have to be made.

Note, that after finding a critical set amounting to, say 95% probability of $\mathbb{M}_K$, one might test what would be the optimal binning procedure to match result from a particular mass spectrometer and, in this way, exactly measure the true empirical machine's resolution, with no need to refer to somewhat underdefined *p percent valley* and *peak width* definitions, see [7].

Acknowledgments. We would like to thank Piotr Dittwald for thourough introduction to the problem of fine isotopic structure and many productive brawls over proper definitions. We also thank Dirk Valkenborg for pointing out the existence of the concept of equatransneutronic isotopes. Finally, a huge thanks goes to prof. Alan Rockwood, for the time we spent together discussing the mass spec related issues and also for showing me America.

References

1. Agnarsson, G.: On the sylvester denumerants for general restricted partitions. (154), 49–60 (2002)
2. Bobkov, S., Ledoux, M.: On modified logarithmic sobolev inequalities for bernoulli and poisson measures. Journal of Functional Analysis 156(2), 347 – 365 (1998), http://www.sciencedirect.com/science/article/pii/S0022123697931876
3. Breen, E.J., Hopwood, F.G., Williams, K.L., Wilkins, M.R.: Automatic poisson peak harvesting for high throughput protein identification. Electrophoresis 21(11), 2243–2251 (2000), http://dx.doi.org/10.1002/1522-2683(20000601)21:11<2243::AID-ELPS2243>3.0.CO;2-K
4. Claesen, J., Dittwald, P., Burzykowski, T., Valkenborg, D.: An Efficient Method to Calculate the Aggregated Isotopic Distribution and Exact Center-Masses. Journal of The American Society for Mass Spectrometry 23(4), 753–763 (Apr 2012), http://dx.doi.org/10.1007/s13361-011-0326-2
5. Dittwald, P., Valkenborg, D., Claesen, J., Rockwood, A., Gambin, A.: On the fine isotopic distribution and limits to resolution in mass spectrometry. in preparation (2014)
6. Dugundji, J.: Topology. Allyn and Bacon series in advanced mathematics, Allyn and Bacon, Boston (1966), http://opac.inria.fr/record=b1079460

7. Eidhammer, I., Flikka, K., Martens, L., Mikalsen, S.O.: Computational Methods for Mass Spectrometry Proteomics. Wiley-Interscience (2007)
8. Hughey, C.A., Hendrickson, C.L., Rodgers, R.P., Marshall, A.G., Qian, K.: Kendrick mass defect spectrum: a compact visual analysis for ultrahigh-resolution broadband mass spectra. Anal. Chem. 73(19), 4676–4681 (Oct 2001)
9. Kallenberg, O.: Foundations of modern probability. Probability and its applications, Springer, New York (2002)
10. Kienitz, H.: Mass Spectrometry and its Applications to Organic Chemistry. Angewandte Chemie 73(17-18) (1961)
11. Kingman, J.F.C.: Poisson processes, Oxford Studies in Probability, vol. 3. The Clarendon Press Oxford University Press, New York (1993), Oxford Science Publications
12. Olson, M., Yergey, A.: Calculation of the isotope cluster for polypeptides by probability grouping. Journal of the American Society for Mass Spectrometry 20(2), 295–302 (2009), http://dx.doi.org/10.1016/j.jasms.2008.10.007
13. Rockwood, A.L.: Relationship of Fourier transforms to isotope distribution calculations. Rapid Commun. Mass Spectrom. 9(1), 103–105 (jan 1995), http://dx.doi.org/10.1002/rcm.1290090122
14. Roos, B.: On the rate of multivariate poisson convergence. Journal of Multivariate Analysis 69(1), 120 – 134 (1999), http://www.sciencedirect.com/science/article/pii/S0047259X98917894
15. Rosman, K., Taylor, P.: Isotopic Compositions of the Elements 1997. J. Phys. Chem. Ref. Data 27(6) (1998)
16. Senko, M.W., Beu, S.C., McLaffertycor, F.W.: Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. J. Am. Soc. Mass Spectrom. 6(4), 229–233 (Apr 1995)
17. Valkenborg, D., Assam, P., Thomas, G., Krols, L., Kas, K., Burzykowski, T.: Using a poisson approximation to predict the isotopic distribution of sulphur-containing peptides in a peptide-centric proteomic approach. Rapid Communications in Mass Spectrometry 21(20), 3387–3391 (2007), http://www.scopus.com/inward/record.url?eid=2-s2.0-35349002401&partnerID=40&md5=d68ac3d300587b84f183dcd1c7f508cd, cited By (since 1996)15
18. Valkenborg, D., Mertens, I., Lemière, F., Witters, E., Burzykowski, T.: The isotopic distribution conundrum. Mass spectrometry reviews 31(1), 96–109 (2012), http://dx.doi.org/10.1002/mas.20339

Tables

Table 1. Basic Information on Stable Isotopes, as found in [15].

Element	Isotope	Extra Neutrons	Mass [Da]	Probability
Carbon	^{12}C	0	12	0.9893
	^{13}C	1	13.0033	0.0107
Hydrogen	^{1}H	0	1.0078	0.999885
	^{2}H	1	2.0141	0.000115
Nitrogen	^{14}N	0	14.0031	0.99632
	^{15}N	1	15.0001	0.00368
Oxygen	^{16}O	0	15.9949	0.99757
	^{17}O	1	16.9991	0.00038
	^{18}O	2	17.9992	0.00205
Sulfur	^{32}S	0	31.9721	0.9493
	^{33}S	1	32.9714	0.0076
	^{34}S	2	33.9679	0.0429
	^{36}S	4	35.9671	0.0002

Appendix

Proof of Lemma 1

We want to prove that if $\mu^{[n]} \rightharpoonup \mu$ and $\mu^{[n]}(A), \mu(A) > 0$, then also $\mu_A^{[n]} \rightharpoonup \mu_A$. We do this under the assumption that both $\mu^{[n]}$ and μ are discrete measures on probability space E.

By the *Portmanteau Lemma*, see [9], $\mu^{[n]} \rightharpoonup \mu$ implies that for any set A with boundry ∂A subject to $\mu(\partial A) = 0$, one should observe

$$\lim_{n\to\infty} \mu^{[n]}(A) = \mu(A). \tag{11}$$

The notion of boundry requires the notion of topology: thus, we decide on the discrete topology, which is natural in this context [10]. In this topology however, $\partial A = \emptyset$, for it is a set theoretical difference of the closure and the interior, both of which are equal to A. Hence, $\mu(\partial A) = 0$. Thus, (11) always holds.

Ex definitione, $\mu^{[n]} \rightharpoonup \mu$ means, that for any bounded function $f : E \to \mathbb{R}$ one observes

$$\int f \mathrm{d}\mu^{[n]} \xrightarrow[n\to\infty]{} \int f \mathrm{d}\mu\,. \tag{12}$$

A simple calculation using both (11) and (12) completes the proof:

$$\int f \mathrm{d}\mu_A^{[n]} = \frac{\int f \mathrm{d}\mu^{[n]}}{\mu^{[n]}(A)} \xrightarrow[n\to\infty]{} \frac{\int f \mathrm{d}\mu}{\mu(A)} = \int f \mathrm{d}\mu\,.$$

[10] For appropriate topological notions consult [6].