Bayesian analysis of systems with random chemical composition: Renormalization-group approach to Dirichlet distributions and the statistical theory of dilution

Marcel Ovidiu Vlad,[1,2] Masa Tsuchiya,[3] Peter Oefner,[3] and John Ross[1]
[1]*Department of Chemistry, Stanford University, Stanford, California 94305-5080*
[2]*Center of Mathematical Statistics, Casa Academiei Romane, Calea Septembrie 13, 76100 Bucharest, Romania*
[3]*Stanford Genome Technology Center, Stanford University School of Medicine, 855 California Avenue, Palo Alto, California 94304*

(Received 31 August 2001; published 20 December 2001)

We investigate the statistical properties of systems with random chemical composition and try to obtain a theoretical derivation of the self-similar Dirichlet distribution, which is used empirically in molecular biology, environmental chemistry, and geochemistry. We consider a system made up of many chemical species and assume that the statistical distribution of the abundance of each chemical species in the system is the result of a succession of a variable number of random dilution events, which can be described by using the renormalization-group theory. A Bayesian approach is used for evaluating the probability density of the chemical composition of the system in terms of the probability densities of the abundances of the different chemical species. We show that for large cascades of dilution events, the probability density of the composition vector of the system is given by a self-similar probability density of the Dirichlet type. We also give an alternative formal derivation for the Dirichlet law based on the maximum entropy approach, by assuming that the average values of the chemical potentials of different species, expressed in terms of molar fractions, are constant. Although the maximum entropy approach leads formally to the Dirichlet distribution, it does not clarify the physical origin of the Dirichlet statistics and has serious limitations. The random theory of dilution provides a physical picture for the emergence of Dirichlet statistics and makes it possible to investigate its validity range. We discuss the implications of our theory in molecular biology, geochemistry, and environmental science.

DOI: 10.1103/PhysRevE.65.011112 PACS number(s): 05.40.−a, 05.10.Cc, 87.15.Cc, 91.65.−n

I. INTRODUCTION

The statistical analysis of various problems of physics, chemistry, and biology involves the consideration of systems with random chemical compositions. Typical examples include statistical studies of the abundances of different chemical species in geochemistry [1], the distribution of pollutants in the environment [2], or the nucleotide frequencies in genomes [3]. For many systems with random composition, the statistics of the fluctuations in composition can be satisfactorily described by means of the Dirichlet probability density [4]

$$\mathcal{P}_N(\boldsymbol{\theta};\boldsymbol{\alpha})d\boldsymbol{\theta}=[Z(\boldsymbol{\alpha})]^{-1}\prod_{u=1}^{N}(\theta_u)^{\alpha_a-1}\delta\left(\sum_{v=1}^{N}\theta_\nu-1\right)d\boldsymbol{\theta}, \tag{1}$$

where the composition vector $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_N)$ is expressed by the mass, volume, or mole fractions $\theta_1,\ldots,\theta_N$ of the different species present in the system, $\alpha_1>0,\ldots,\alpha_N>0$ are positive integers and

$$Z(\boldsymbol{\alpha})=\int\cdots\int\prod_{u=1}^{N}(\theta_\nu)^{\alpha_u-1}\delta\left(\sum_{v=1}^{N}\theta_\nu-1\right)d\boldsymbol{\theta}$$
$$=\frac{\prod_{u=1}^{N}\Gamma(\alpha_u)}{\Gamma\left(\sum_{u=1}^{N}\alpha_u\right)} \tag{2}$$

is a partition function. The standard method used in mathematical statistics for the generation of the Dirichlet probability density is to express the fractions $\theta_1,\ldots,\theta_N$ in terms of N random variables $X_1\ldots,X_N$, as $\theta_u=X_u/\Sigma_{u=1}^{N}X_u$, where each random variable X_u is selected from a different Gamma (or χ^2) probability density. Under these circumstances, it is easy to show that the vector $\boldsymbol{\theta}=(\theta_1\ldots\theta_N)$ obeys a probability law of the type (1). Unfortunately, this is only a formal statistical derivation that does not clarify the meaning of the probability density (1).

Recently, the empirical use of the Dirichlet distribution has become popular, especially in molecular biology where it provides a satisfactory description of nucleotide statistics in DNA strands or amino acid statistics in proteins [4]. Other applications include the description of pollutant distribution in the environment [2], its use in material science for describing the chemical composition of disordered systems [5], as well as its use in geochemistry [1]. In all of these cases, the Dirichlet distribution is employed merely as an empirical law, which manages to give a satisfactory description of the observed data. No simple physical explanation for the occurrence of the Dirichlet law has been given. The purpose of this paper is the presentation of a simple physical explanation for the Dirichlet law (1) for the composition fluctuations. Our main assumption is that the random variations in composition are due to the occurrence of a succession of a random number of dilution events. Such a mechanism seems reasonable not only in environmental chemistry and geochemistry but also in molecular biology, where the process of nucleotide substitution can act as a dilution factor, which tends to destroy the correlations among the different nucleotides in a DNA strand. In our theoretical description, the possible chemical reactions among the various species are not taken into account explicitly. Such reactions may be supplementary dilution factors and thus, at least in principle, the reactions may be taken into account by a suitable description of the statistics of the dilution process.

The structure of this paper is the following. In Sec. II we formulate the problem of evaluating the probability density of composition fluctuations by using a Bayesian approach involving inverse probabilities. In Sec. III we use a stochastic renormalization-group approach for computing the probability densities of the abundances of the different chemical species present in the system and derive the Dirichlet probability law. In Sec. IV we present an alternative, formal derivation of the Dirichlet probability density based on the maximum information entropy approach. In Sec. V, we discuss the implications of our approach in molecular biology, environmental chemistry, and geochemistry.

II. BAYESIAN ANALYSIS OF SYSTEMS WITH RANDOM CHEMICAL COMPOSITION

We denote by

$$\tilde{p}_u(F_u)dF_u \quad \text{with} \quad \int_0^\infty \tilde{p}_u(F_u)dF_u=1, \tag{3}$$

the probability that the (extensive) amount of a chemical species u in a space region of dimension Ω is between F_u and F_u+dF_u. We also introduce the notation

$$\mathcal{B}_N(\mathbf{F}|F)d\mathbf{F} \quad \text{with} \quad \int\cdots\int \mathcal{B}_N(\mathbf{F}|F)d\mathbf{F}=1, \quad \sum_{u=1}^{N} F_u=F, \tag{4}$$

for the conditional probability that in a space region of dimension Ω, the extensive amounts of species $u=1,...,N$ are between F_u and F_u+dF_u, $u=1,\ldots,N$, respectively, provided that the total amount of chemicals is constant and equal to F. Here, $\mathbf{F}=(F_1,\ldots F_N)$ is the extensive composition vector of the system. Since the fractions $\theta_1,\ldots,\theta_N$ of different chemicals are given by

$$\theta_u=F_u \Big/ \sum_{u=1}^{N} F_u; \;\; F_u=\theta_u F; \;\; u=1,\ldots,N, \tag{5}$$

it follows that

$$\mathcal{P}_N(\boldsymbol{\theta})d\boldsymbol{\theta}=\mathcal{B}_N(\mathbf{F}|F)d\mathbf{F}=\mathcal{B}_N(\mathbf{F}=F\boldsymbol{\theta}|F)F^N d\theta. \tag{6}$$

Thus, the evaluation of the probability density $\mathcal{P}_N(\theta)$ of the intensive composition vector $\boldsymbol{\theta}$ reduces to the evaluation of the conditional probability density $\mathcal{B}_N(\mathbf{F}|F)$.

The unconditional probability $P_N\ (F)\,dF$ of the total amount of chemicals may be computed by evaluating the average of a delta function, which is a standard procedure in statistical physics. We have

$$\begin{aligned} P_N(F)d\mathbf{F} &= \left\langle \delta\left(F-\sum_{u=1}^{N} f_u\right)dF\right\rangle \\ &= \int_0^\infty\cdots\int_0^\infty \delta\left(F-\sum_{u=1}^{N} f_u\right)dF \\ &\quad \times\prod_{u=1}^{N}[\tilde{p}_u(f_u)]df_1,\ldots,df_N \\ &= \tilde{p}(F)\otimes\tilde{p}_2(F)\otimes,\ldots,\otimes\tilde{p}_N(F), \end{aligned} \tag{7}$$

where $\otimes$ denotes the additive convolution product. Since the amounts of chemicals $F_1\ldots,F_N$ and F are non-negative random variables, the characteristic functions of the probability densities $\tilde{p}_u(F_u)$, $u=1,\ldots,N$ and $P_N(F)$ may be expressed as Laplace transforms

$$g_u(s)=\int_0^\infty \exp(-sF_u)\tilde{p}_u(F_u)dF_u, \quad u=1,\ldots,N, \tag{8}$$

$$G_N(s)=\int_0^\infty \exp(-sF)P_N(F)dF. \tag{9}$$

From Eqs. (7)–(9), it follows that

$$G_N(s)=\prod_{u=1}^{N}[g_u(s)], \;\; P_N(F)=\mathcal{L}_{F,s}^{-1}\left\{\prod_{u=1}^{N}[g_u(s)]\right\}, \tag{10}$$

where $\mathcal{L}_{F,s}^{-1}$ denotes the inverse Laplace transformation

Now we introduce the joint probability $\mathcal{R}_N(\mathbf{F},F)d\mathbf{F}\,dF$, that the extensive composition vector of the system is between $\mathbf{F}$ and $\mathbf{F}+d\mathbf{F}$ and that the total amount of chemicals is between F and $F+dF$. This probability can be computed by averaging a product of delta functions

$$\begin{aligned} \mathcal{R}_N(\mathbf{F},F)d\mathbf{F}\,dF &= \left\langle \prod_{u=1}^{N}[\delta(F_u-f_u)F_u]\delta\left(F-\sum_{u=1}^{N} f_u\right)dF\right\rangle \\ &= \int_0^\infty\cdots\int_0^\infty \prod_{u=1}^{N}[\delta(F_u-f_u)F_u] \\ &\quad \times\delta\left(F-\sum_{u=1}^{N} f_u\right)dF\prod_{u=1}^{N}[\tilde{p}_u(f_u)]df_u,\ldots,df_N. \end{aligned} \tag{11}$$

We define the characteristic function

$$\mathfrak{G}_N(\mathbf{x},s)=\int_0^\infty \exp\left(-sF-\sum_{u=1}^{N} x_u F_u\right)\mathcal{R}_N(\mathbf{F},F)d\mathbf{F}\,dF, \tag{12}$$

and take the multiple Laplace transform of Eq. (11), with respect to $F_1,\ldots,F_N$ and F. We come to

$$\mathfrak{G}_N(\mathbf{x},s)=\prod_{u=1}^{N}[g_u(s+x_u)],$$

$$\mathcal{R}_N(\mathbf{F},F)=\mathcal{L}^{-1}_{F,s}\mathcal{L}^{-1}_{F_1,x_1}\ldots,\mathcal{L}^{-1}_{F_N,x_N}\left\{\prod_{u=1}^{N}[g_u(s+x_u)]\right\}. \tag{13}$$

By using the Bayes method, [6], $\mathcal{B}_N$ $(\mathbf{F}|F)$ may be expressed as an inverse probability density, which is the ratio between the joint probability density of the extensive composition vector and the total amount of chemicals, $\mathcal{R}_N(\mathbf{F},F)\,d\mathbf{F}\,dF$, and the probability for the total amount of chemicals $P_N(F)\,dF$:

$$\mathcal{B}_N(\mathbf{F}|F)d\mathbf{F}=\frac{\mathcal{R}_N(\mathbf{F},F)d\mathbf{F}\,dF}{P_N(F)dF}$$
$$=\frac{\mathcal{L}^{-1}_{F,s}\mathcal{L}^{-1}_{F_N,x_N}\ldots\mathcal{L}^{-1}_{F_1,x_1}\{\prod_{u=1}^{N}[g_u(s+x_u)]\}}{\mathcal{L}^{-1}_{F,s}\{\prod_{u=1}^{N}[g_u(s)]\}}d\mathbf{F}. \tag{14}$$

In conclusion, in this section we have shown that the probability density of the relative composition vector $\boldsymbol{\theta}$ may be evaluated from the probability densities of the abundances of the different species by using a Bayesian approach involving inverse probabilities. Equation (14) derived in this section is used in Sec. III for the derivation of the Dirichlet probability density.

III. RANDOM DILUTION AND RENORMALIZATION-GROUP THEORY

In this section, we analyze the dilution of various components in a complex system. We notice that there are at least two different sources of stochasticity. In the first place, random sampling concentration fluctuations may emerge, due to the fact that matter is made up of molecules. A second source of stochasticity is due to the fact that the size of the region within which the dilution may take place may be randomly varying in size; this second type of fluctuation may be described as multiplicative noise. In our following analysis, we assume that the molecular fluctuations may be neglected in comparison with the contribution of the multiplicative noise. We also assume that the different components in the system may be diluted in different ways; we shall show that this last assumption results in different scaling exponents $\alpha_1>0,\ldots,\alpha_N>0$ for the different species in the Dirichlet probability density (1).

In order to apply Eq. (14), we need to know the probability densities $\tilde{p}_u(F_u)$ of the extensive amounts of various chemicals present in the domain of extension Ω. In order to compute $\tilde{p}_u(F_u)$, we develop a random theory of dilution. We assume that the extension Ω is small enough so that the concentration $C_u=F_u/\Omega$ is uniform throughout the region. Each dilution event leads to the spreading of a given chemical initially present in a region of size Ω to a larger region, characterized by a size $\Omega'>\Omega$. After a dilution event takes place, the amount of chemical F_u is spread out in a larger region of size Ω' and thus, the concentration of the species becomes smaller $C'_u=F_u/\Omega'<C_u$. Out of the amount F_u initially present in the region of size Ω, the amount

$$F'_u=C'_u\Omega=F_u\Omega/\Omega'=bF_u \quad \text{with} \quad 0<b=\Omega/\Omega'<1, \tag{15}$$

remains in the region and the difference

$$\Delta F_u=F'_u-F_u=(1-b)F_u, \tag{16}$$

moves away. The ratio between two successive sizes, $b=\Omega/\Omega'$ is a dilution factor between zero and one. A succession of dilution events may be characterized by two different sets of random parameters: the dilution factors b_q, $q=1,2\ldots$, and the probabilities λ_q, $q=1,2,\ldots$, that the various dilution events take place. We consider that for each dilution events $q=1,2,\ldots$, the parameters b_q,λ_q are randomly selected from a constant probability density

$$\Phi_u(b,\lambda)db\,d\lambda, \quad \text{with} \quad \int_0^1\int_0^1\Phi_u(b,\lambda)\,db\,d\lambda=1. \tag{17}$$

We assume that before the occurrence of any dilution events the concentrations of the species are rather large and the concentration fluctuations are rather small. Under these circumstances, the initial probability density for the amount of species u, $p_u^{(0)}(F_u)$, is rather narrow with a sharp maximum corresponding to the most probable value of F_u. After each dilution event, the probability density of F_u becomes flatter and flatter, the concentration fluctuations increase, and the typical values of F_u become smaller and smaller.

The succession of dilution events can be described in terms of the joint probabilities

$$\Psi_u^{(q)\pm}(\lambda_u^{(q)},F_u^{(q)})d\lambda_u^{(q)}dF_u^{(q)}, \quad q=0,1,2,\ldots \tag{18}$$

with

$$\sum_{q=0}^{\infty}\int_0^1\int_0^{\infty}\Psi_u^{-}(\lambda_u^{(q)},F_u^{(q)})d\lambda_u^{(q)}dF_u^{(q)}=1. \tag{19}$$

$\Psi_u^{(q)+}(\lambda_u^{(q)},F_u^{(q)})\,d\lambda_u^{(q)}dF_u^{(q)}$ is the probability that q dilution events have occurred and that the dilution factor has a value between $\lambda_u^{(q)}$ and $\lambda_u^{(q)}+d\lambda_u^{(q)}$ and that the amount of the u species is between $F_u^{(q)}$ and $F_u^{(q)}+dF_a^{(q)}$: the superscript + means that the dilution process has not been terminated after q steps. The probability $\Psi_u^{(q)-}(\lambda_u^{(q)},F_u^{(q)})\,d\lambda_u^{(q)}dF_u^{(q)}$ has a similar significance with the difference that the minus sign means that the succession of dilution events finishes after q steps. In terms of these probabilities, we may write down the following evolution equations:

$$\Psi_u^{(q)+}(\lambda_u^{(q)},F_u^{(q)})=\lambda_u^{(q)}\int_0^1\int_0^\infty\int_0^1\Phi_u(b_u^{(q)},\lambda_u^{(q)})\Psi_u^{(q-1)+}(\lambda_u^{(q-1)},F_u^{(q-1)})\,\delta(F_u^{(q)}-b_u^{(q)}F_u^{(q-1)})d\lambda_u^{(q-1)}dF_u^{(q-1)}db_u^{(q)},\tag{20}$$

$$\Psi_u^{(q)-}(\lambda_u^{(q)},F_u^{(q)})=(1-\lambda_u^{(q)})\int_0^1\int_0^\infty\int_0^1\Phi_u(b_u^{(q)},\lambda_u^{(q)})\Psi_u^{(q-1)+}(\lambda_u^{(q-1)},F_u^{(q-1)})\times\delta(F_u^{(q)}-b_u^{(q)}F_u^{(q-1)})d\lambda_u^{(q-1)}dF_u^{(q-1)}db_u^{(q)},\tag{21}$$

with the initial conditions

$$\Psi_u^{(0)+}(\lambda_u^{(0)},F_u^{(0)})=\lambda_u^{(0)}p_u^{(0)}(F_u^{(0)})\int_0^1\Phi_u(b_u^{(0)},\lambda_u^{(0)})\,db_u^{(0)},\tag{22}$$

$$\Psi_u^{(0)-}(\lambda_u^{(0)},F_u^{(0)})=(1-\lambda_u^{(0)})p_u^{(0)}(F_u^{(0)})\int_0^1\Phi_u(b_u^{(0)},\lambda_u^{(0)})db_u^{(0)}.\tag{23}$$

The probability density $\tilde{p}_u(F_u)$ of the total amount of species u after the occurrence of a random number of dilution events may be expressed in terms of $\Psi_u^{(q)-}(\lambda_u^{(q)},F_u^{(q)})$. We have

$$\tilde{p}_u(F_u)=\sum_{q=1}^{\infty}\int_0^1\Psi_a^{(q)-}(\lambda_u^{(q)},F_u)d\lambda_u^{(q)}.\tag{24}$$

Equations (20)–(24) may be solved step by step. After some calculations, we may express the probability density $\tilde{p}_u(F_u)$ by an expansion of the Lippmann-Schwinger type

$$\tilde{p}_u(F_u)=P_u^{(0)}(F_u)\int_0^1\int_0^1(1-\lambda_u^{(0)})\Phi_u(b_u^{(0)},\lambda_u^{(0)})db_u^{(0)}d\lambda_u^{(0)}+\sum_{q=1}^{\infty}\int_0^1\int_0^1 db_u^{(q)}d\lambda_u^{(q)}\cdots\int_0^1\int_0^1 db_u^{(0)}d\lambda_u^{(0)}\times(1-\lambda_u^{(q)})\frac{\lambda_u^{(q-1)},\dots,\lambda_u^{(0)}}{b_u^{(q)},\dots,b_u^{(1)}}p_u^{(0)}\times\left(\frac{F_u}{b_u^{(q)},\dots,b_u^{(1)}}\right)\prod_{w=0}^{q}[\Phi_u(b_u^{(w)},\lambda_u^{(w)})].\tag{25}$$

We notice that the Lippmann-Schwinger series (25) has a self-similar structure that makes it possible to derive a stochastic renormalization-group equation [7] for $\tilde{p}_u(F_u)$. By using a summation label $q'=q-1$, Eq. (25) leads to

$$\tilde{p}_u(F_u)=p_u^{(0)}(F_u)\int_0^1\int_0^1(1-\lambda_u^{(0)})\Phi_u(b_u^{(0)},\lambda_u^{(0)})db_u^{(0)}d\lambda_u^{(0)}+\int_0^1\int_0^1 db_u d\lambda_u\Phi_u(b_u,\lambda_u)\frac{\lambda_u}{b_u}p_u^{(0)}\left(\frac{F_u}{b_u}\right)\times\int_0^1\int_0^1(1-\lambda_u^{(0)})\Phi_u(b_u^{(0)},\lambda_u^{(0)})db_u^{(0)}d\lambda_u^{(0)}+\int_0^1\int_0^1 db_u d\lambda_u\Phi_u(b_u,\lambda_u)\frac{\lambda_u}{b_u}\sum_{q'=1}^{\infty}\int_0^1\int_0^1 db_u^{(q')}d\lambda_u^{(q')}\cdots\times\int_0^1\int_0^1 db_u^{(0)}d\lambda_u^{(0)}(1-\lambda_u^{(q')})\frac{\lambda_u^{(q'-1)},\dots,\lambda_u^{(0)}}{b_u^{(q')},\dots,b_u^{(1)}}\times p_u^{(0)}\left(\frac{F_u}{b_u^{(q')},\dots,b_u^{(1)}b_u^{(1)}}\right)\prod_{w=0}^{q}[\Phi_u(b_u^{(w)},\lambda_u^{(w)})].\tag{26}$$

By comparing Eqs. (25)–(26), we notice that the probability density $\tilde{p}_u(F_u)$ is self similar: its self-similar properties are expressed by the fact that in Eq. (26), the terms of order bigger than one may be grouped together in an integral expression that contains a scaled form of the probability density $\tilde{p}_u$. By using this property, from Eqs. (25)–(26) we may derive a self-similar integral equation for $\tilde{p}_u(F_u)$,

$$\tilde{p}_u(F_u)=p_u^{(0)}(F_u)\int_0^1\int_0^1(1-\lambda_u)\Phi_u(b_u,\lambda_u)db_u d\lambda_u + \int_0^1\int_0^1 db_u d\lambda_u \Phi_u(b_u,\lambda_u)\frac{\lambda_u}{b_u}\tilde{p}_u\left(\frac{F_u}{b_u}\right). \tag{27}$$

Equation (27) is a renormalization-group equation, which expresses the self-similar features of the cascade of dilution events that generate the probability density $\tilde{p}_u(F_u)$. We notice that in Eq. (27) the integral term in $\tilde{p}_u(F_u)$ has the structure of a multiplicative convolution product, which suggests that the equation may be solved by using the Mellin transform. The solution of the equation is made up of the sum of analytic and a nonanalytic components. The nonanalytic part of the solution, which has the dominant contribution for small F_u, has the following structure:

$$\tilde{p}_u(F_u)\sim(F_u)^{\alpha_0-1}A_u^{(0)}+\sum_{\rho=1}^{\infty}(F_u)^{\zeta_\rho-1}A_u^{(\rho)}[\ln F_u], \tag{28}$$

where $\alpha=\alpha_0$ is the unique real root of a secular transcendental equation

$$I(\alpha)=\int_0^1\int_0^1\lambda b^{-\alpha}\Phi_u(b,\lambda)db\, d\lambda=1, \tag{29}$$

$\alpha_\rho=\zeta_\rho\pm i\sigma_\rho$ are the complex roots of the same equation, $A_u^{(0)}$ is a constant, and $A_u^{(\rho)}[\ln F_u]$ are periodic functions of $\ln F_u$ with periods $2\pi/\sigma_\rho$. From Eq. (29), it follows that $I(0)=\langle\lambda\rangle\leqslant 1$ and $dI(\alpha)/d\alpha>0$ and thus, there is a single real root $\alpha=\alpha_0$, which is non-negative $\alpha_0\geqslant 0$. Since Eq. (29) has real coefficients, the complex roots, if they exist, must occur in conjugated pairs $\alpha_\rho=\zeta_\rho\pm i\sigma_\rho$. It is easy to show that

$$\begin{aligned}\operatorname{Re} I(\alpha_\rho=\zeta_\rho\pm i\sigma_\rho)&=\int_0^1\int_0^1\lambda\exp[\zeta_\rho\ln(1/b)]\\ &\quad\times[\cos\sigma\ln(1/b)]\Phi_u(b,\lambda)db\, d\lambda=1\\ &=\int_0^1\int_0^1\lambda\exp[\alpha_0\ln(1/b)]\Phi_u(b,\lambda)db\, d\lambda\end{aligned} \tag{30}$$

and since $\cos[\sigma_\rho\ln(1/b)]\leqslant 1$, it turns out that the real parts of complex roots fulfill the inequality

$$\operatorname{Re}\alpha_\rho=\zeta_\rho\geqslant\alpha_0. \tag{31}$$

From Eqs. (28) and (31) it follows that, if $\zeta_\rho>\alpha_0$, the real root α_0 dominates the asymptotic behavior of $\tilde{p}_u(F_u)$ as $F_u\to 0$ and in this limit the logarithmic oscillations may be neglected. However, if at least one complex root $\alpha_{\rho *}=\zeta_{\rho *}\pm i\sigma_{\rho *}$ has a real part equal to the real root, $\alpha_0=\xi_{\rho *}$ then the logarithmic oscillations may influence the statistics of the dilution process.

The occurrence of logarithmic oscillations is a known issue in renormalization-group theory [8]. In most cases, they are assumed to be artifacts generated by the discrete nature of the renormalization-group equations and special limits are considered in order to get rid of them. In a few cases, however, the logarithmic oscillations are real and may be observed experimentally [9,10]. In the case of our random dilution theory, the logarithmic oscillations are not compatible with the Dirichlet probability density (1) and for this reason, we prefer to eliminate them from the evolution equations. Since the solution of our renormalization-group Eq. (27) is a probability density we cannot just neglect the oscillatory terms from the solution, because such a crude approach would result in a violation of the normalization condition for the probability density $\tilde{p}_u(F_u)$. Instead, we consider the physical and mathematical circumstances under which the logarithmic oscillations vanish. For our model, the logarithmic oscillations are generated by the dilution factor b, which changes discretely from dilution event to dilution event. The logarithmic oscillations vanish for processes involving very large numbers of dilution events and for which the variation of the dilution factor from event to event is very small. In order to identify this type of process, we compute the probability χ_q that q dilution events takes place. We have

$$\begin{aligned}\chi_q&=\int_0^1\int_0^\infty\Psi_u^-(\lambda_u^{(q)},F_u^{(q)})\,d\lambda_u^{(q)}dF_u^{(q)}\\ &=\int_0^\infty dF_u^{(q)}\int_0^1\int_0^1 db_u^{(q)}d\lambda_u^{(q)}\cdots\int_0^1\int_0^1 db_u^{(0)}d\lambda_u^{(0)}\\ &\quad\times(1-\lambda_u^{(q)})\frac{\lambda_u^{(q-1)},..,\lambda_u^{(0)}}{b_u^{(q)},\ldots,b_u^{(1)}}p_u^{(0)}\left(\frac{F_u^{(q)}}{b_u^{(q)},\ldots,b_u^{(1)}}\right)\\ &\quad\times\prod_{w=0}^{q}[\Phi_u(b_u^{(w)},\lambda_u^{(w)})]=(1-\langle\lambda_u\rangle)\langle\lambda_u\rangle^q,\end{aligned} \tag{32}$$

where

$$\langle\lambda_u\rangle=\int_0^1\int_0^1\lambda_u\Phi_u(b_u,\lambda_u)db_u d\lambda_u, \tag{33}$$

is the average probability of occurrence of a dilution event. According to Eq. (33), the average number of dilution events

$$\langle q_u\rangle=\sum_{q=0}^{\infty}q\chi_q=1/(1-\langle\lambda_u\rangle), \tag{34}$$

tends to infinity, $\langle q_u\rangle\to\infty$, as the average probability of occurrence of a dilution event tends to unity $\langle\lambda_u\rangle\to 1$. In this limit, λ_u is not random anymore. For $\lambda_u\to 1$, we should also consider that $b_u\to 1$ because otherwise the b_u-dependent factors in Eq. (25) may lead to the violation of the normalization condition for $\tilde{p}_u(F_u)$. A straightforward analysis shows that the limit

$$\lambda_u\to 1,\quad b_u\to 1 \tag{35}$$

leads to an indetermination. We should supplement this limit with a constraint, which preserves the scaling features of the renormalization-group Eq. (27). In order to determine this constraint, we consider values of λ_u, and b_u close to unity

$$\lambda_u = 1 - \varepsilon_{u\lambda}, \quad b_u = 1 - \varepsilon_{ub}, \quad \text{with} \quad \varepsilon_{u\lambda}, \varepsilon_{ub} \quad \text{close to} \quad 0. \tag{36}$$

In this case, the probability density $\mathbf{\Phi}_u(b,\lambda)$ may be approximated by

$$\mathbf{\Phi}_u(b,\lambda) = \delta(\lambda - 1 + \varepsilon_{u\lambda})\,\delta(b - 1 + \varepsilon_{\rm ub}), \tag{37}$$

and the secular Eq. (29) may be solved analytically. We have

$$\alpha_0 = [\ln(1 - \varepsilon_{u\lambda})]/[\ln(1 - \varepsilon_{\rm ub})], \tag{38}$$

$$\begin{aligned} \alpha_\rho = \zeta_\rho \pm i\sigma_\rho = [\ln(1 - \varepsilon_{u\lambda})]/[\ln(1 - \varepsilon_{\rm ub})] \\ \pm i2\pi\rho/[\ln(1 - \varepsilon_{\rm ub})], \quad \rho = 1,2, \ldots . \end{aligned} \tag{39}$$

We require that in the limit (35), the fractal exponent α_0 given by Eq. (38), which expresses the scaling properties of the renormalization-group equation, remains constant. This condition leads to the following constraint:

$$\alpha_{0u} = \alpha_u = \ln \lambda_u / \ln b_u \quad \text{constant as} \quad \lambda_u \to 1, \quad b_u \to 1, \tag{40}$$

where we have taken into account that different scaling exponents exist for different chemical species and, for simplicity, we dropped the superscript 0.

We apply the limit (35) with the constraint (40) to the renormalization-group Eq. (27). We come to a differential equation in $\tilde{p}_u(F_u)$,

$$\alpha_u p_u^{(0)}(F_u) + F_u \frac{\partial}{\partial F_u} \tilde{p}_u(F_u) = (\alpha_u - 1)\tilde{p}_u(F_u). \tag{41}$$

The normalized solution of Eq. (41) is

$$\tilde{p}_u(F_u) = \alpha_u (F_u)^{\alpha_u - 1} \int_{F_u}^{\infty} \frac{p_u^{(0)}(y)}{y^{\alpha_u}}\, dy. \tag{42}$$

From Eq. (42), we may compute the characteristic functions $g_u(s)$, $u = 1,...,N$ of the probability densities $\tilde{p}_u(F_u)$, $u = 1,...,N$. We obtain

$$g_u(s) = \alpha_u s^{-\alpha_u} \int_0^{\infty} \frac{p_u^{(0)}(y)}{y^{\alpha_a}}\, \gamma(\alpha_u, sy)\, dy, \tag{43}$$

where

$$y(a,x) = \int_0^x x^{a-1} \exp(-x)\, dx, \quad x \geqslant 0, \quad a > 0, \tag{44}$$

is the incomplete gamma function. In our computations, we have assumed that the initial probability densities $p_u^{(0)}(y)$ are rather narrow with very sharp maxima corresponding to the most probable values of $F_u^{(p)}$. It follows that in Eq. (43), the main contribution to the integral in Eq. (43) comes from values of y close to the most probable values of the initial amount of species u. As x increases, the incomplete gamma function $\gamma(a,x)$ tends fast towards a threshold value given by the corresponding complete gamma function $\Gamma(a) = \int_0^\infty x^{a-1} \exp(-x) dx$. If the initial most probable value $F_u^{(p)}$ is large enough, then the integral in Eq. (43) may be approximated by replacing the incomplete gamma function by the corresponding complete gamma function, resulting in

$$g_u(s) \sim s^{-\alpha_u} \langle [F_u]^{-\alpha_u} \rangle^{(0)} \Gamma(1 + \alpha_u), \tag{45}$$

where

$$\langle [F_u]^{\alpha_u} \rangle^{(0)} = \int_0^\infty [F_u]^{-a_u} p_u^{(0)}(F_u)\, dF_u, \tag{46}$$

is a negative moment of the initial amount of species u, evaluated in terms of the initial probability density $p_u^{(0)}(F_u)$.

Now we have all information necessary for computing the conditional probability density $\mathcal{B}_N(\mathbf{F}|F)$. By inserting Eq. (45) into Eq. (14), we come to

$$\begin{aligned} \mathcal{B}_N(\mathbf{F}|F) &= \frac{\mathcal{L}_{F,s}^{-1} \mathcal{L}_{F_1,x_1}^{-1}, \ldots, \mathcal{L}_{F_N,x_N}^{-1} \left\{ \prod_{u=1}^{N} [(s + x_u)^{-\alpha_u}] \right\}}{\mathcal{L}_{F,s}^{-1}[s^{-\Sigma_u \alpha_u}]} \\ &= \frac{\mathcal{L}_{F,s}^{-1} \left\{ e^{-s\Sigma_u F_u} \prod_{u=1}^{N} \left[\frac{(F_u)^{\alpha_u - 1}}{\Gamma(\alpha_u)} \right] \right\}}{\dfrac{F^{\Sigma_u \alpha_u - 1}}{\Gamma\left(\sum_u \alpha_u \right)}} \\ &= \frac{\Gamma\left(\sum_u \alpha_u \right)}{\prod_{u=1}^{N} \Gamma(\alpha_u)} \frac{\prod_{u=1}^{N} (F_u)^{\alpha_u - 1}}{F^{\Sigma_u \alpha_u - 1}} \delta\left(F - \sum_u F_u \right), \end{aligned} \tag{47}$$

from which, by using the transformation of variables (5), we may compute the probability density $\mathcal{P}_N(\boldsymbol{\theta})$ of the chemical composition of the system, expressed in terms of the intensive vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$. By combining Eqs. (5), (6), and (47), we obtain the Dirichlet law (1) and the expression (2) for the partition function $Z(\boldsymbol{\alpha})$.

In this section, we have derived a system of stochastic evolution equations for the probability densities of the abundances of the different species present in the system. We have shown that these evolution equations have a self-similar structure that makes it possible to evaluate their solution by using the renormalization-group theory. We have shown that the self-similar structure of the evolution equations leads to the Dirichlet distribution.

IV. MAXIMUM INFORMATION ENTROPY APPROACH TO DIRICHLET DISTRIBUTIONS

In the particular case where the chemical mixture may be assumed to be an ideal solution, an alternative derivation of Eqs. (1)–(2) may be given by using the maximum entropy approach. For an ideal solution, the chemical potential $\mu_u(\theta_u)$ is a linear function of the logarithm of the molar (molecular) fraction θ_u of the species u in the mixture

$$\mu_u(\theta_u)=\mu_u^0+k_BT\ln\theta_u\,,\quad u=1,\ldots,N. \tag{48}$$

We consider the Kullback information measure (information gain)

$$\mathfrak{K}[\mathcal{P}_N(\boldsymbol{\theta});\mathcal{P}_N^{\mathrm{Prior}}(\boldsymbol{\theta})]=\int\cdots\int\mathcal{P}_N(\boldsymbol{\theta})\ln\left[\frac{\mathcal{P}_N(\boldsymbol{\theta})}{\mathcal{P}_N^{\mathrm{Prior}}(\boldsymbol{\theta})}\right]d\boldsymbol{\theta}, \tag{49}$$

where $\mathcal{P}_N^{\mathrm{Prior}}(\boldsymbol{\theta})$ is a suitable prior probability density of the intensive composition vector $\boldsymbol{\theta}$. We search for an extremum of the information gain $\mathfrak{K}[\mathcal{P}_N(\boldsymbol{\theta});\mathcal{P}_N^{\mathrm{Prior}}(\boldsymbol{\theta})]$ with respect to the probability density $\mathcal{P}_N(\boldsymbol{\theta})$, which is compatible with the following constraints: (i) The normalization condition for $\mathcal{P}_N(\theta)$

$$\int\cdots\int\mathcal{P}_N(\boldsymbol{\theta})d\boldsymbol{\theta}=1, \tag{50}$$

(ii) The average values of the chemical potentials of the different species are constant

$$\langle\mu_u\rangle=\int\cdots\int(\mu_u^0+k_BT\ln\theta_u)\mathcal{P}_N(\boldsymbol{\theta})\,d\boldsymbol{\theta}=\text{constant},$$
$$u=1,\ldots,N. \tag{51}$$

We must also implement the conservation condition $\Sigma_{u=1}^N\theta_u=1$, by requiring that both $\mathcal{P}_N(\boldsymbol{\theta})$ and $\mathcal{P}_N^{\mathrm{Prior}}(\boldsymbol{\theta})$ include a delta function

$$\mathcal{P}_N(\boldsymbol{\theta})=\mathcal{A}_N(\boldsymbol{\theta})\,\delta\left(\sum_{u=1}^N\theta_u-1\right),$$
$$\mathcal{P}_N^{\mathrm{Prior}}(\boldsymbol{\theta})=\mathcal{A}_N^{\mathrm{Prior}}(\boldsymbol{\theta})\,\delta\left(\sum_{u=1}^N\theta_u-1\right). \tag{52}$$

By carrying out the computations, we get the following expressions for the probability density $\mathcal{P}_N(\boldsymbol{\theta})$ of the composition vector and for the extremal value $\mathfrak{K}_{\mathrm{extr.}}$ of the information gain:

$$\mathcal{P}_N(\boldsymbol{\theta})=[(N-1)!Z(\boldsymbol{\alpha})]^{-1}\mathcal{A}_N^{\mathrm{Prior}}(\theta)\prod_{u=1}^N(\theta_u)^{\alpha_u-1}$$
$$\times\,\delta\left(\sum_{v=1}^N\theta_u-1\right), \tag{53}$$

and

$$\mathfrak{K}_{\mathrm{extr.}}=[(N-1)!Z(\boldsymbol{\alpha})]^{-1}\int\cdots\int\mathcal{A}_N^{\mathrm{Prior}}(\boldsymbol{\theta})\prod_{u=1}^N(\theta_u)^{\alpha_u-1}$$
$$\times\,\delta\left(\sum_{u=1}^N\theta_u-1\right)\ln\left[\frac{1}{(N-1)!Z(\boldsymbol{\alpha})}\prod_{u=1}^N(\theta_u)^{\alpha_u-1}\right]d\boldsymbol{\theta}, \tag{54}$$

where the partition function $Z(\boldsymbol{\alpha})$ is given by

$$Z(\boldsymbol{\alpha})=[(N-1)!]^{-1}\int\cdots\int\mathcal{A}_N^{\mathrm{Prior}}(\theta)\prod_{u=1}^N(\theta_N)^{\alpha_u-1}$$
$$\times\,\delta\left(\sum_{u=1}^N\theta_u-1\right)d\boldsymbol{\theta}, \tag{55}$$

and the scaling exponents α_u, $u=1,\ldots,N$ are the solutions of the equations

$$\langle\mu_u\rangle=\mu_u^0+k_BT\frac{\partial}{\partial\alpha_u}\ln Z(\boldsymbol{\alpha}),\quad u=1,\ldots,N. \tag{56}$$

In particular, if the prior probability is constant,

$$\mathcal{A}_N^{\mathrm{Prior}}(\theta)=\mathcal{A}_N^{\mathrm{Prior}}\ \text{ independent of }\ \boldsymbol{\theta}, \tag{57}$$

we have

$$\mathcal{A}_N^{\mathrm{Prior}}=\left[\int\cdots\int\delta\left(\sum_{u=1}^N\theta_u-1\right)d\boldsymbol{\theta}\right]^{-1}=(N-1)! \tag{58}$$

and Eqs. (53) and (55) reduce to the Dirichlet Eqs. (1)–(2).

The maximum entropy derivation of Eqs. (1)–(2) introduced in this section is much simpler than the renormalization-group approach based on the random theory of dilution. Unfortunately, the maximum entropy approach is no more illuminating than the standard statistical derivation of Eqs. (1)–(2) starting from the beta distribution, and does not clarify the physical origins of the Dirichlet statistics. There is no simple physical justification for the assumption that the average values of the chemical potentials of the various chemical species are constant.

V. DISCUSSION

Although more complicated than the maximum information entropy approach, the random theory of dilution provides a simple physical explanation for the emergence of Dirichlet statistics, which is the result of a cascade of large numbers of successive dilution events that lead to broad distributions for the amounts of the different chemicals present in the system. From the point of view of statistical physics, this cascade of dilution events may be viewed as a renormalization-group transformation. The renormalization-group approach provides an explanation for the self-similar features of the Dirichlet law, expressed by the scaling exponents $\alpha_1,\ldots\alpha_N$ attached to the different species present in the system. The cascades of dilution events tend to increase the contributions of small concentrations and the probability densities $\tilde{p}_u(F_u)$ have a self-similar scaling behavior as $F_u \sim 0$. For long cascades, this scaling behavior is not only valid as $F_u \sim 0$ but also for relatively large F_u. In the limit of an infinite number of dilution events, the self-similar scaling behavior controls completely the concentration fluctuations, resulting in Dirichlet statistics.

The random theory of dilution makes it possible to analyze the limitations of Dirichlet statistics. It is reasonable to assume that the Dirichlet must provide a satisfactory data fit whenever long cascades of dilution events are likely to occur. Long times necessary for the development evolutionary transformations in molecular biology or in geology may justify the use of Dirichlet distribution in nucleotide statistics or in geochemistry. Concerning the applications in environmental science, we expect that the Dirichlet distribution may be applied for describing the distribution of pollutants in the limit of large times. For the Dirichlet statistics to hold, it is necessary that a long enough time interval has elapsed since the release of the pollutant, so that a large number of dilution events have taken place.

In our derivation of the Dirichlet law, we considered a limited region of extension Ω, which is assumed to be homogeneous. For large systems, it is likely that the homogeneity assumption does not hold, and thus, we expect that there are deviations from the Dirichlet statistics. Such size effects have been already observed in molecular biology. Although in most cases the nucleotide statistics corresponding to a limited DNA strand may be described by the Dirichlet law, the description fails for large sets of data [3,11]. For such large systems, the nucleotide statistics may still be described by a linear combination of Dirichlet distributions. According to our theory, each Dirichlet function from the linear combination corresponds to a subregion of the system, which is small enough so that the homogeneity holds.

In this paper, we have aimed at a simple derivation of the Dirichlet law, which is applicable both to geochemistry, environmental chemistry, as well as molecular biology. In order to achieve this, we have described the dilution process in terms of continuous random variables. For a deeper understanding of the DNA and protein composition statistics in molecular biology, further research must focus on describing the dilution process in terms of discrete random variables.

ACKNOWLEDGMENTS

This research has been supported in part by the National Science Foundation, NIH Grant No. HG00205 and by the Department of Energy, BES/Engineering Research Program.

[1] R. P. Wayne, *Chemistry of Atmospheres* (Clarendon, Oxford, 1991); B. J. Finlayson-Pitts and J. N. Pitts, *Atmospheric Chemistry* (Wiley, New York, 1986); H. D. Holland, *The Chemistry of the Atmosphere and Oceans* (Wiley, New York, 1978); J. H. Seinfeld, *Atmospheric Chemistry and Physics of Air Pollution* (Wiley, New York, 1986); P. L. Brezonik, *Chemical Kinetics and Process Dynamics in Aquatic Systems* (Lewis Publishers, Boca Raton, FL, 1994).

[2] S. S. Isukapalli, A. Roy, and P. G. Georgopoulos, Risk Anal **18(3)**, 351 (1988); *Chemometrics In Environmental Chemistry Statistical Methods*, edited by J. Einax (Springer, Berlin, 1995).

[3] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press, Cambridge, 1998).

[4] N. L. Johnson and S. Kotz, *Distributions in Statistics: Continuous Multivariate Distributions* (Wiley, New York, 1972).

[5] J. Aitchison, *The Statistical Analysis of Compositional Data* (Chapman and Hall, New York, 1986).

[6] E. T. Jaynes, *Bayesian Methods-An Introductory Tutorial*, in *Maximum Entropy and Bayesian Methods in Applied Statistics*, edited by J. H. Justice (Cambridge University Press, Cambridge, 1986).

[7] M. F. Shlesinger and B. D. Hughes, Physica A **109**, 597 (1981); E. W. Montroll and M. F. Shlesinger, Proc. Natl. Acad. Sci. U.S.A. **79**, 338 (1982); J. Stat. Phys. **32**, 209 (1983); E. W. Montroll and M. F. Shlesinger, in *Nonequilibrium Phenomena II. From Stochastics to Hydrodynamics*, edited by J. L. Lebowitz and E. W. Montroll (North-Holland, Amsterdam, 1984), pp. 3–117.

[8] M. Schrekenberg, Z. Phys. B: Condens. Matter **60**, 483 (1985). A. Giacometti, A. Maritan, and A. Stella, Int. J. Mod. Phys. B **5**, 709 (1991).

[9] F. Anselmet, Y. Gagne, E. Hopfinger, and R. Antonia, J. Fluid Mech. **140**, 331 (1984); L. A. Smith, J. D. Fournier, and E. A. Spiegel, Phys. Lett. **114A**, 465 (1986); B. J. West, B. Bhargava, and A. L. Goldberger, J. Appl. Phys. **60**, 1089 (1986); T. R. Nelson, B. J. West, and A. L. Goldberger, Experientia **46**, 251 (1990); M. F. Shlesinger and B. J. West, Phys. Rev. Lett. **67**, 2106 (1991); B. J. West and W. Deering, Phys. Rep. **246**, 100 (1994).

[10] M. O. Vlad, G. F. Cerofolini, and John Ross, J. Phys. Chem. A **103**, 4798 (1999).

[11] K. Sjölander and Collaboration CABIOS, Comput. Appl. Biosci. **12**, 327 (1996).