Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Computational Physics 211 (2006) 198–209
JOURNAL OF COMPUTATIONAL PHYSICS
www.elsevier.com/locate/jcp

A practical computational framework for the multidimensional moment-constrained maximum entropy principle

Rafail Abramov

Courant Institute of Mathematical Sciences, New York University, 251 Mercer St., NY 10012, United States

Received 13 October 2004; received in revised form 17 May 2005; accepted 23 May 2005
Available online 6 July 2005

Abstract

The maximum entropy principle is a versatile tool for evaluating smooth approximations of probability density functions with a least bias beyond given constraints. In particular, the moment-based constraints are often a common prior information about a statistical state in various areas of science, including that of a forecast ensemble or a climate in atmospheric science. With that in mind, here we present a unified computational framework for an arbitrary number of phase space dimensions and moment constraints for both Shannon and relative entropies, together with a practical usable convex optimization algorithm based on the Newton method with an additional preconditioning and robust numerical integration routine. This optimization algorithm has already been used in three studies of predictability, and so far was found to be capable of producing reliable results in one- and two-dimensional phase spaces with moment constraints of up to order 4. The current work extensively references those earlier studies as practical examples of the applicability of the algorithm developed below.

1. Introduction

The idea of applying the rigorous framework of information theory for ensemble forecasting was proposed in [1], utilizing Shannon entropy as a measure of the lack of predictive information in a forecast. It later evolved into the concept of measuring the lack of information in a known average long-term statistical state (climate) relative to a forecast via relative entropy [2,3]. In particular, the latter was formalized in [4] by developing a hierarchy of rigorous lower bounds for the relative entropy, and turned into a practical predictability tool in [5] via applying the relative entropy for information measurements in higher moments of forecast ensembles (non-Gaussianity) and in cross-moments of multidimensional probability densities, represented by mutual information. The framework was successfully used in [6] for the triangular

E-mail address: abramov@cims.nyu.edu.

0021-9991/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2005.05.008

T21 spherical truncation of barotropic flow with realistic Earth-like topography. The same relative entropy framework was employed in [7] for rigorous estimates of the credibility of small-sampled statistical ensembles. Other applications of the maximum entropy principle include, in particular, solid state physics [8] and econometrics [9,10].

The moment-based maximum entropy framework is designed for practical estimates of probability densities via a smooth approximation with the least bias beyond the prior set of measured statistical properties of such an ensemble (usually the mean state, variance, and higher moments). Although a variety of methods to obtain smooth approximations to statistical ensembles is available (such as kernel density estimates [11–13]), we assume instead that the moment data is available. For example, input data can be a result of post-processing observations, which are not directly related to the quantities of interest, and thus are not necessarily represented by an ensemble of phase space states. In addition, unlike kernel estimates, our approach provides a rigorous upper bound on Shannon entropy and lower bound on relative entropy under the given constraints, which is a highly desirable feature. The approximation itself is obtained by either maximizing Shannon entropy or minimizing relative entropy (depending on the definition of the optimization problem considered) subject to taking measured moments as constraints [14] within the framework known as the Hausdorff moment problem (for details see [15] and references therein). The corresponding smooth optimization problems are convex, hence a local optimum is automatically global. Standard methods to resolve such optimization problems include the gradient descent method, the Newton algorithm [16], and the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [17]. Any of these algorithms can be incorporated into the maximum entropy framework presented here. Our algorithm of choice is the Newton method due to its simplicity, affine invariance property, fast local convergence, and the fact that both the gradient and Hessian of our optimization problem can be easily computed. We combine the Newton method with a simple line search technique to ensure its global stability, which is a precautionary measure to avoid numerical instability and blow-up.

Despite its theoretical clarity, a successful implementation of such a technique requires the computational algorithm to have additional properties, such as swiftness and sufficiently wide range of input parameters for which the algorithm converges (dynamic range). While the skeleton of the optimization algorithm consists of the Newton method, additional procedures are implemented to achieve the necessary degree of robustness. Those include the preconditioning of input parameters to improve the dynamic range of convergence and the fast, yet accurate, algorithm for numerical integration which is tailored to the problem and is required to compute both the gradient and Hessian for the Newton iterations.

The manuscript is organized as follows. In Section 2, the two optimization problems for the Shannon and relative entropies are formulated for an arbitrary set of constraints. In Section 3, it is shown that, in the case of moment constraints, the principle of minimum relative entropy can be reduced to the principle of maximum Shannon entropy, thus establishing the common optimization framework for both types of entropy which greatly simplifies the practical implementation of the numerical optimization algorithm. Section 4 contains the description of the preconditioning of input data to achieve better dynamic range of convergence, as well as the fast numerical integration algorithm for both the gradient and Hessian of the optimization problem based on Gauss–Hermite quadrature. Instead of a systematic numerical study of the developed algorithm, detailed references to concurrent work [5,6] in which this algorithm was successfully applied are presented in Section 5. These references should constitute sufficient evidence of this algorithm's practical usability and overall robustness. Section 6 summarizes the results of this work.

2. The formulation of two optimization problems for general constraints

Before providing the detailed explanation of the two moment-constrained entropy problems, a more general optimization framework for an arbitrary set of constraints is formulated in this section for both

the Shannon entropy and relative entropy. Let the partial knowledge of a probability density $p(\vec{\lambda})$, where $\vec{\lambda}$ denotes a vector of phase space coordinates, be available in the general form of integral constraints

$$\mathscr{F}_0[p] = \int_{\mathbb{R}^N} p(\vec{\lambda})\,\mathrm{d}\vec{\lambda} = 1, \tag{1a}$$

$$\mathscr{F}_i[p] = \int_{\mathbb{R}^N} C_i(\vec{\lambda})p(\vec{\lambda})\,\mathrm{d}\vec{\lambda} = f_i, \quad 1 \leqslant i \leqslant L, \tag{1b}$$

where $C_i(\vec{\lambda})$ are such that the integrals (1) are finite, and arbitrary otherwise. Then, the two optimization problems are formulated:

(1) *Optimization problem of the first kind*: find the optimal probability density p_S which maximizes the Shannon entropy

$$S(p) = -\int_{\mathbb{R}^N} p(\vec{\lambda}) \ln p(\vec{\lambda})\mathrm{d}\vec{\lambda}, \tag{2}$$

(i.e., such that $S(p_S) = \max_p S(p)$) and simultaneously satisfies the constraints in (1);

(2) *Optimization problem of the second kind*: find the optimal probability density p_P which minimizes the relative entropy

$$P(p, \Pi) = \int_{\mathbb{R}^N} p(\vec{\lambda}) \ln\left[\frac{p(\vec{\lambda})}{\Pi(\vec{\lambda})}\right]\mathrm{d}\vec{\lambda}, \tag{3}$$

(i.e., such that $P(p_P, \Pi) = \min_p[P(p, \Pi)]$) and simultaneously satisfies the constraints in (1), where $\Pi(\vec{\lambda})$ is some known prior probability density.

A convex optimization approach allows the reformulation of the above two problems for the dual space of the Lagrange multipliers in the unconstrained form [14,18]. In particular, the constrained optimization problems of the first and second kind are reduced to the unconstrained minimization of the Lagrangian functions in the form

$$\mathscr{L}_S(\vec{\gamma}) = \ln\left[\int_{\mathbb{R}^N} \exp\left(\sum_{i=1}^{L} \gamma_i C_i(\vec{\lambda})\right)\mathrm{d}\vec{\lambda}\right] - \sum_{i=1}^{L} \gamma_i f_i, \tag{4a}$$

$$\mathscr{L}_P(\vec{\theta}) = \ln\left[\int_{\mathbb{R}^N} \Pi(\vec{\lambda}) \exp\left(\sum_{i=1}^{L} \theta_i C_i(\vec{\lambda})\right)\mathrm{d}\vec{\lambda}\right] - \sum_{i=1}^{L} \theta_i f_i, \tag{4b}$$

over their sets of Lagrange multipliers $\vec{\gamma}$ and $\vec{\theta}$, respectively. The Hessians of (4a) and (4b) are positive definite [18], and therefore the optimization problems of both the first and second kind are convex with unique optima. The optimal probability densities p_S and p_P are then

$$p_S(\vec{\lambda}) = \exp\left(\sum_{i=0}^{L} \gamma_i C_i(\vec{\lambda})\right), \tag{5a}$$

$$p_P(\vec{\lambda}) = \Pi(\vec{\lambda}) \exp\left(\sum_{i=0}^{L} \theta_i C_i(\vec{\lambda})\right), \tag{5b}$$

where γ_0 and θ_0 are chosen to meet the normalization requirement for probability densities:

$$\gamma_0 = -\ln \int_{\mathbb{R}^N} \exp\left(\sum_{i=1}^{L} \gamma_i C_i(\vec{\lambda})\right) \mathrm{d}\vec{\lambda}, \tag{6a}$$

$$\theta_0 = -\ln \int_{\mathbb{R}^N} \Pi(\vec{\lambda}) \exp\left(\sum_{i=1}^{L} \theta_i C_i(\vec{\lambda})\right) \mathrm{d}\vec{\lambda}. \tag{6b}$$

It is easy to see that if the optima for (4a) and (4b) are reached (i.e., the gradients of (4a) and (4b) are zero), then the constraints in (1) are satisfied by (5a) and (5b) exactly.

3. The unified optimization framework for moment constraints

For practical applications it is common for the constraints in (1) to be various products of powers of phase space coordinates (moments). Since optimization problems with moment constraints are less general than those in (4a) and (4b), additional simplifications can be made in the theoretical formulation of the optimization problems before proceeding with their numerical implementation. First, however, we need to introduce a suitable written form of an arbitrary moment by denoting powers of phase space coordinates via an index of integers:

- Let $\vec{i} = (i_1, \ldots, i_N)$ be an index of nonnegative integers with $|\vec{i}| = \sum_{k=1}^{N} i_k$, where N is the dimension of phase space.
- The $\geqslant$ and $>$ relations are defined as
  - $\{\vec{i} \geqslant \vec{j}\} \Longleftrightarrow \{i_k \geqslant j_k \ \forall k\}$, i.e., $\vec{i}$ being greater than or equal to $\vec{j}$ means that this relation holds independently for all their components i_k and j_k;
  - $\{\vec{i} > \vec{j}\} \Longleftrightarrow \{\vec{i} \geqslant \vec{j}; \exists k^* : i_{k^*} > j_{k^*}\}$, i.e. $\vec{i}$ being strictly greater than $\vec{j}$ means that, first, $\vec{i} \geqslant \vec{j}$ holds (this relation is already defined above), plus there exists at least one index k^*, such that i_{k^*} is strictly greater than j_{k^*}.

Thus, an arbitrary moment of $\vec{\lambda}$ (a product of arbitrary powers of the components of $\vec{\lambda}$) is compactly written as

$$\vec{\lambda}^{\vec{i}} = \prod_{k=1}^{N} \lambda_k^{i_k}, \tag{7}$$

where the moment order is the total power of all vector components, i.e., $|\vec{i}|$. Using the above notation, for a probability density p, we write a set of arbitrary moment constraints up to the total power L as

$$f_0 = 1, \tag{8a}$$

$$f_{\vec{i}} = \int_{\mathbb{R}^N} \vec{\lambda}^{\vec{i}} p(\vec{\lambda}) \mathrm{d}\vec{\lambda}, \quad |\vec{i}| = 1, \tag{8b}$$

$$f_{\vec{i}} = \int_{\mathbb{R}^N} (\vec{\lambda} - \bar{\vec{\lambda}}_p)^{\vec{i}} p(\vec{\lambda}) \mathrm{d}\vec{\lambda}, \quad |\vec{i}| = 2 \ldots L, \tag{8c}$$

where $\bar{\vec{\lambda}}_p$ denotes the mean state vector for p (the moments of second and higher orders are centered at the mean state).

With the set of moment constraints in (8), the unconstrained optimization problems in (4a) and (4b) become

$$\mathscr{L}_S(\vec{\gamma}) = \ln\left[\int_{\mathbb{R}^N} \exp\left(\sum_{|\vec{i}|=1}^{L} \gamma_{\vec{i}}\vec{\lambda}^{\vec{i}}\right) \mathrm{d}\vec{\lambda}\right] - \sum_{|\vec{i}|=2}^{L} \gamma_{\vec{i}} f_{\vec{i}}, \tag{9a}$$

$$\mathscr{L}_P(\vec{\theta}) = \ln\left[\int_{\mathbb{R}^N} \Pi(\vec{\lambda}+\bar{\vec{\lambda}}_p) \exp\left(\sum_{|\vec{i}|=1}^{L} \theta_{\vec{i}}\vec{\lambda}^{\vec{i}}\right) \mathrm{d}\vec{\lambda}\right] - \sum_{|\vec{i}|=2}^{L} \theta_{\vec{i}} f_{\vec{i}}, \tag{9b}$$

with corresponding optimal probability densities

$$p_S(\vec{\lambda}) = \exp\left(\sum_{i=0}^{L} \gamma_i(\vec{\lambda}-\bar{\vec{\lambda}}_p)^{\vec{i}}\right), \tag{10a}$$

$$p_P(\vec{\lambda}) = \Pi(\vec{\lambda}) \exp\left(\sum_{i=0}^{L} \theta_i(\vec{\lambda}-\bar{\vec{\lambda}}_p)^{\vec{i}}\right). \tag{10b}$$

Note that the probability densities in (10a) and (10b) structurally differ solely due to the $\Pi(\vec{\lambda})$ factor in the latter. In particular, if $\Pi(\vec{\lambda})$ itself is of the form (10a), then (10b) also takes the form in (10a). This observation leads to the following theorem.

Theorem 1. *If the prior probability density Π is of the form (10a), then the optimization problem of the second kind in (9b) can be reduced to the optimization problem of the first kind in (9a).*

Proof. Let Π be of the form

$$\Pi(\vec{\lambda}) = \exp\left(\sum_{|\vec{m}|=0}^{M} \alpha_{\vec{m}}(\vec{\lambda}-\bar{\vec{\lambda}}_\Pi)^{\vec{m}}\right), \tag{11}$$

where $\bar{\vec{\lambda}}_\Pi$ denotes the mean state for Π, and $M \leqslant L$. This choice of Π is natural if Π is itself the optimal probability density for the optimization problem of the first kind in (9a) with different constraints. Then the optimization problem of the second kind in (9b) can be written as

$$\mathscr{L}_P(\vec{\theta}) = \ln\left[\int \exp\left(\sum_{|\vec{i}|=1}^{L} \theta_{\vec{i}}\vec{\lambda}^{\vec{i}} + \sum_{|\vec{m}|=1}^{M} \alpha_{\vec{m}}(\vec{\lambda}+\bar{\vec{\lambda}}_p-\bar{\vec{\lambda}}_\Pi)^{\vec{m}}\right) \mathrm{d}\lambda\right] + \alpha_0 - \sum_{|\vec{i}|=2}^{L} \theta_{\vec{i}} f_{\vec{i}}. \tag{12}$$

The following change of variables

$$\sum_{|\vec{i}|=1}^{L} \theta_{\vec{i}}\vec{\lambda}^{\vec{i}} + \sum_{|\vec{m}|=1}^{M} \alpha_{\vec{m}}(\vec{\lambda}+\bar{\vec{\lambda}}_p-\bar{\vec{\lambda}}_\Pi)^{\vec{m}} = C_0 + \sum_{|\vec{k}|=1}^{L} \gamma_{\vec{k}}\vec{\lambda}^{\vec{k}} \quad \forall\vec{\lambda}, \tag{13a}$$

$$C_0 = \sum_{|\vec{k}|=1}^{M} \alpha_{\vec{k}}(\bar{\vec{\lambda}}_p-\bar{\vec{\lambda}}_\Pi)^{\vec{k}}, \tag{13b}$$

$$\gamma_{\vec{k}} = \theta_{\vec{k}} + \sum_{\vec{m}\geqslant\vec{k}} \binom{\vec{m}}{\vec{k}} \alpha_{\vec{m}}(\bar{\vec{\lambda}}_p-\bar{\vec{\lambda}}_\Pi)^{\vec{m}-\vec{k}}, \tag{13c}$$

$$\binom{\vec{m}}{\vec{k}} = \prod_{i=1}^{N} \binom{m}{k} = \prod_{i=1}^{N} \frac{m_i!}{k_i!(m_i-k_i)!} \tag{13d}$$

converts the optimization problem of the second kind in (9b) into

$$\mathscr{L}_P(\vec{\gamma}) = \ln\left[\int \exp\left(\sum_{|\vec{k}|=1}^{L} \gamma_{\vec{k}} \vec{\lambda}^{\vec{k}}\right) \mathrm{d}\vec{\lambda}\right] - \sum_{|\vec{k}|=2}^{L} \gamma_{\vec{k}} f_{\vec{k}} + \sum_{|\vec{k}|=2}^{L} \sum_{\vec{m} \geqslant \vec{k}} \binom{\vec{m}}{\vec{k}} \alpha_{\vec{m}} (\bar{\vec{\lambda}}_p - \bar{\vec{\lambda}}_\Pi)^{\vec{m}-\vec{k}} f_{\vec{k}} + \alpha_0 + C_0. \tag{14}$$

Now it can be seen that $\mathscr{L}_P(\vec{\gamma})$ in (14) and $\mathscr{L}_S(\vec{\gamma})$ in the optimization problem of the first kind (9a) differ by a constant, which means that, for the same set of initial constraints, the optimal set of Lagrange multipliers for (14) coincides with that for (9a). □

Theorem 1 establishes the common algorithmic approach for the optimization problems of both the first and second kind. Once the optimal sets of Lagrange multipliers for $p_S(\vec{\lambda})$ and $p_P(\vec{\lambda})$ are found, the corresponding Shannon and relative entropies are computed as follows: the explicit formula for the Shannon entropy in (2) is

$$S(p) = -\gamma_0 - \sum_{|\vec{k}|=2}^{L} \gamma_{\vec{k}} f_{\vec{k}} \tag{15}$$

and for the relative entropy in (3) the formula consists of three parts:

$$P(p, \Pi) = \mathscr{S} + \mathscr{D} + \mathscr{C}\mathscr{T}. \tag{16}$$

The three terms in the right-hand side of (16) are the signal, dispersion and cross-term:

$$\mathscr{S} = -C_0 = -\sum_{|\vec{k}|=1}^{M} \alpha_{\vec{k}} (\bar{\vec{\lambda}}_p - \bar{\vec{\lambda}}_\Pi)^{\vec{k}}, \quad \text{signal}, \tag{17a}$$

$$\mathscr{D} = \gamma_0 - \alpha_0 + \sum_{|\vec{k}|=2}^{L} (\gamma_{\vec{k}} - \alpha_{\vec{k}}) f_{\vec{k}}, \quad \text{dispersion}, \tag{17b}$$

$$\mathscr{C}\mathscr{T} = -\sum_{|\vec{k}|=2}^{L} \sum_{\vec{m} > \vec{k}} \binom{\vec{m}}{\vec{k}} \alpha_{\vec{m}} (\bar{\vec{\lambda}}_p - \bar{\vec{\lambda}}_\Pi)^{\vec{m}-\vec{k}} f_{\vec{k}}, \quad \text{cross-term}. \tag{17c}$$

Note that the signal is zero when the mean state of p coincides with the mean state of Π, even if the second and higher moments are different; whereas the dispersion is zero when the second and higher moments of p coincide with those of Π, even if their mean states are different. This observation allows to distinguish between different sources of predictability [3,4]. The cross-term is zero if Π is Gaussian. The properties of the signal–dispersion–cross-term decomposition are discussed in detail in [5,6].

4. Preconditioning and numerical method of solution

It is shown in Theorem 1 that, for the moment constraints and the special form of a prior probability density Π, the optimization problem of the second kind in (9b) can be reduced to the optimization problem of the first kind in (9a), thus establishing a unified computational approach for both problems. In this section we describe how one can implement this approach in practice.

Before proceeding with the numerical algorithm for optimization, it is desirable to simplify and standardize the initial problem as much as possible via usual arithmetic means (as partially done in Theorem 1). This removes many implementation problems and improves dynamic range of applicability and convergence of optimization algorithms. Standard ways of boosting convergence usually involve a different coordinate basis in the space of Lagrange multipliers [8], which is of significant help if moments of

high ($\geqslant$10) order are considered on a one-dimensional domain, which is often the case in solid state physics. Here, instead, the situation is complicated by the fact that a higher-dimensional phase space is used, whereas moments do not go up to high orders. The latter necessitates a phase space change of coordinates (which would be unnecessary in 1D), while not necessarily requiring Lagrange multiplier modification as in [8]. As one can see, the optimization problem in (9a) is already formulated in such a way that the mean state of the objective probability density does not directly participate as a constraint, and is added directly into the objective probability density in (10) after its shape is computed. In other words, the optimization problem in (9a) is solved as if its mean state constraint is always zero, which in a certain sense standardizes the problem.

Here, we propose further formalization via the preconditioning of the constraints by rotating and stretching the coordinate system in such a way that the constraint covariance matrix (the matrix of second moments) becomes the identity matrix in new coordinates. This is achieved by setting

$$\vec{\lambda}_{\text{old}} = E\Lambda^{1/2}\vec{\lambda}_{\text{new}}, \tag{18}$$

where E is the matrix of eigenvectors, and Λ is the diagonal matrix of eigenvalues for the constraint covariance matrix. If the covariance matrix is found to be singular, the target probability density is supported on a subspace of smaller dimension. Thus, one finds a new reduced coordinate system by dropping eigenvectors with zero eigenvalues. Note that in the case of a linear coordinate transformation, the recomputation of the moment constraints of order l in the new coordinates requires only the moment constraints up to order l in the old coordinates. Since no extra information beyond the set of moment constraints has to be provided for this preconditioning to work, its user-transparent encapsulation into the main optimization routine makes it very convenient in practical use (for example, it can be incorporated into an existing code without violating its intrinsic dependencies). After the optimal Lagrange multipliers in the new coordinates $\vec{\lambda}_{\text{new}}$ are found, they can be similarly recomputed in the original coordinates $\vec{\lambda}_{\text{old}}$.

In particular, in the case of two-dimensional phase space, the rotation-stretching transformation is

$$\begin{aligned}\lambda_{1,\text{new}} &= a_{11}\lambda_{1,\text{old}} + a_{12}\lambda_{2,\text{old}}\\ \lambda_{2,\text{new}} &= a_{21}\lambda_{1,\text{old}} + a_{22}\lambda_{2,\text{old}}\end{aligned}, \qquad \Lambda^{-1/2}E^{\mathrm{T}} = \begin{bmatrix} a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}. \tag{19}$$

Then the constraints in the new coordinates are computed from the old coordinates through

$$\left\langle \lambda_{1,\text{new}}^{k}\lambda_{2,\text{new}}^{l}\right\rangle = \sum_{m=0}^{k}\sum_{n=0}^{l}\binom{k}{m}\binom{l}{n}a_{11}^{m}a_{21}^{n}a_{12}^{k-m}a_{22}^{l-n}\left\langle \lambda_{1,\text{old}}^{m+n}\lambda_{2,\text{old}}^{k+l-m-n}\right\rangle, \tag{20}$$

where by $\langle\ldots\rangle$ we denote an integral over the phase space weighted by the objective probability density. As for the Lagrange multipliers, one can write the argument of the exponent in (10) as

$$\sum_{k,l}\gamma_{kl}^{\text{new}}\lambda_{1,\text{new}}^{k}\lambda_{2,\text{new}}^{l} = \sum_{k,l}\sum_{m=0}^{k}\sum_{n=0}^{l}\gamma_{kl}^{\text{new}}\times\binom{k}{m}\binom{l}{n}a_{11}^{m}a_{12}^{k-m}a_{21}^{n}a_{22}^{l-n}\lambda_{1,\text{old}}^{m+n}\lambda_{2,\text{old}}^{k+l-m-n}, \tag{21}$$

then $\gamma_{k,l}^{\text{old}}$ are simply the sums of the constant coefficients in (21) with same combinations of powers $\lambda_{1,\text{old}}$ and $\lambda_{2,\text{old}}$. This implicit operation is easy to implement algorithmically. For an N-dimensional phase space with $N > 2$, N-term sums can be recursively split into pairs of partial sums, such that the binomial expansions in (20) and (21) are utilized for these pairs on each step of the recursion.

The standard Newton method [16] with an inexact line search is used to solve the convex optimization problem with the Lagrangian function in (9a), or, equivalently, in (14). The gradient and Hessian of the Lagrangian function in (9a) are computed at each Newton iteration step as

$$\frac{\partial \mathscr{L}_S}{\partial \gamma_{\vec{i}}} = \frac{Q(\vec{i})}{Q(0)} - f_{\vec{k}}, \tag{22a}$$

$$\frac{\partial \mathscr{L}_S}{\partial \gamma_{\vec{i}} \partial \gamma_{\vec{j}}} = \frac{Q(\vec{i}+\vec{j})}{Q(0)} - \frac{Q(\vec{i})Q(\vec{j})}{Q(0)^2}, \tag{22b}$$

where the building blocks $Q(\vec{k})$ are given by the integrals of the form

$$Q(\vec{k}) = \int_{\mathbb{R}^N} \vec{\lambda}^{\vec{k}} \exp\left(\sum_{|\vec{i}|=1}^{L} \gamma_{\vec{i}} \vec{\lambda}^{\vec{i}}\right) \mathrm{d}\vec{\lambda}. \tag{23}$$

A similar approach in the one-dimensional setting was described in [10]. Note that the set of the building blocks in (23) for the gradient in (22a) is a subset of the set of those building blocks for the Hessian in (22b), hence the gradient is automatically computed in the process of the Hessian computation.

The integrals in (23) are computed using the Gauss–Hermite quadrature [19]. Recall that the optimization problem is preconditioned by (18) to have a zero mean state and identity covariance matrix, which means that in case of Gaussian constraints the integrals in (23) are computed exactly. In addition, the Gaussian probability density with zero mean state and identity covariance matrix is used as the starting point for the Newton iterations.

At first, for non-Gaussian constraints the Gauss–Hermite quadrature might appear to have no particular advantage over any other high-order quadrature. However, the location of the abscissas and magnitude of the weights are important. The integrands in (23) are centered at zero and have similar widths for low and moderate $|\vec{k}|$ (because of the identity covariance matrix). The Gauss–Hermite abscissae are concentrated near the origin and become sparse away from it, and the Gauss–Hermite weights approach zero away from the origin (where the integrands in (23) approach zero as well). Thus, the Gauss–Hermite quadrature correctly samples an integrand of the form (23), which makes it suitable for our problem. In the examples presented below the two-dimensional nested Gauss–Hermite quadrature is used with 100 discretization points in each dimension.

5. Experience of practical application and examples

In this section, we show some practical applications of the optimization algorithm described in Sections 2–4. Since the 4-moment approximation is only a moderate improvement over the Gaussian 2-moment approximation, we here do not perform a systematic study to test the algorithm, but instead describe some of the results already obtained in the studies where this algorithm was used as a tool [5–7]. The reason that only moments up to order 4 are used is that in practical full-scale atmospheric general circulation models (GCM) only few-member ensembles are computationally feasible, which means that higher order moments cannot be estimated reliably. This undersampling issue is systematically addressed in [7] by estimating the credibility of a measurement based on its sample size through information theory.

Two versions of the optimization algorithm have been numerically implemented so far: one-dimensional and two-dimensional, both with an arbitrary number of input constraints (moments up to order 4 have been used in practice). In fact, the total number of constraints in the two-dimensional case is 14, since all cross-moments (mixed powers of two state variables) of the order of less than or equal to 4 are used.

The work [5], which is the first study where the optimization algorithm is used extensively, employs the Lorenz 96 model. The optimization algorithm is used for evaluation of the optimal probability density functions of statistical forecast ensembles, as well as the climatological steady states (averaged long-time series of a single solution) which are given in a basis of Fourier harmonics (wavenumbers). Thus

constructed probability density functions are used in the study of predictability of the Lorenz 96 model. In Fig. 1, we demonstrate the joint two-dimensional PDF's of the statistical steady state for the Fourier wavenumbers $k = 0$ and $k = 8$. The PDF on the left-hand plot in Fig. 1 is obtained by bin-counting the long time series of a single mixing solution, thus showing the projection of the statistical steady state on the wavenumbers $k = 0$ and $k = 8$. The right-hand plot in Fig. 1 shows the PDF reconstructed by the optimization algorithm from the measured statistical cross-moments of up to order 4. As we can see, the optimized PDF is close in shape to the actual one. The PDF on the left-hand plot in Fig. 1 appears to be close to a Gaussian in shape, however upon closer inspection some non-Gaussian distortions are noticeable, especially in the lowest contour at the base of the PDF. These distortions are captured by the optimization algorithm in the estimated PDF on the right-hand plot in Fig. 1.

Another example of a successful application of the optimization algorithm, described in Sections 2–4, is the work [6]. Again, the optimization algorithm is used for evaluation of the optimal probability density functions of statistical forecast ensembles, as well as climatological steady states (averaged long-time series of a single solution). The properties of two-dimensional barotropic flow on a sphere are modeled via the standard triangular T21 truncation on spherical harmonics, with realistic Earth-like topography and realistic constant forcing designed to mimic the variability of actual air flow in the atmosphere at a certain height. Unlike the results displayed for the Lorenz 96 model in Fig. 1 (where weak non-Gaussianity was detected), the statistical steady state of the T21 truncation, corresponding to the 500 mbar potential height, is highly non-Gaussian.

Note that all the results for the T21 truncation are given in the standard (within the atmosphere/ocean science community) orthogonal coordinate system, which is chosen so that the mean of the statistical steady state is zero, and its covariance matrix is diagonal. This coordinate system is called the empirical orthogonal basis, whereas its spectral coefficients are called principal components (PCs). The ability of the two-dimensional optimization algorithm to capture non-Gaussian features in a PDF is shown in Fig. 2. The left-hand plot in Fig. 2 depicts the actual joint two-dimensional PDF of the first and second principal components for the T21 truncation. The structure of this two-dimensional PDF is highly non-Gaussian with one high peak and two little humps. It is obvious that a Gaussian with the same mean and variance cannot approximate such a PDF.

The right-hand plot in Fig. 2 demonstrates the PDF, which is estimated by the two-dimensional optimization algorithm from the recorded moments of up to order 4. It can be seen that most of the features of the original PDF are captured, and, in particular, the location and size of two small humps coincides with the original. On the other hand, the sharp peak in the original PDF on the left-hand plot in Fig. 2 is not

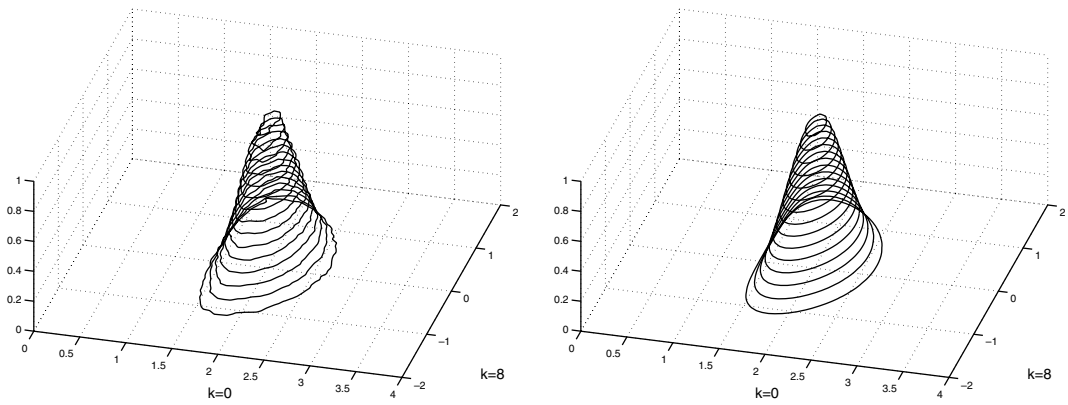

Fig. 1. The joint two-dimensional statistical steady-state PDF's for the Fourier modes $k = 0$ and $k = 8$ for the Lorenz 96 model [5]. Left: the PDF found by bin-counting. Right: the optimal PDF determined under the moment constraints up to order 4.

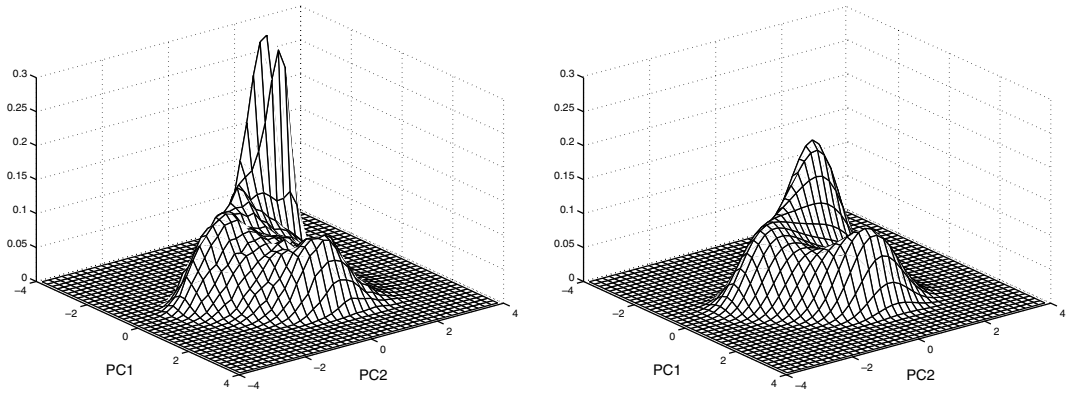

Fig. 2. The joint two-dimensional PDF's for the statistical steady state of the first and second principal components (PC1 and PC2) in the standard empirical orthogonal basis of the T21 spherical truncation of the barotropic flow with realistic topography [6]. Left: the PDF found by bin-counting. Right: the optimal PDF determined under the moment constraints up to order 4.

captured well (due to the lack of order of approximation), however there is still a smaller, smoother peak in its place in the estimated PDF on the right-hand plot in Fig. 2. The entropy of the actual PDF on the left-hand plot relative to the estimated PDF on the right-hand plot in Fig. 2 is 5.736×10^{-2}, which indicates that such a sharp and narrow peak does not contribute much of information, although attracting visual attention.

The results in Fig. 3 are the two estimated PDF's, with their samples drawn from a forecast ensemble with high mutual information in the principal components PC1 and PC3 for the T21 truncation. The reason for such high mutual information is the specific sample structure of this particular forecast ensemble, consisting of a narrow band oriented at an angle with respect to both the PC1 and PC3 coordinates, such that the second cross-moment between the PC1 and PC3 is unusually high. The joint PDF for the PC1 and PC3 on the left-hand plot in Fig. 3 is obtained through the two-dimensional optimization algorithm using moments up to order 4, whereas the PDF on the right-hand plot in Fig. 3 is a direct product of two

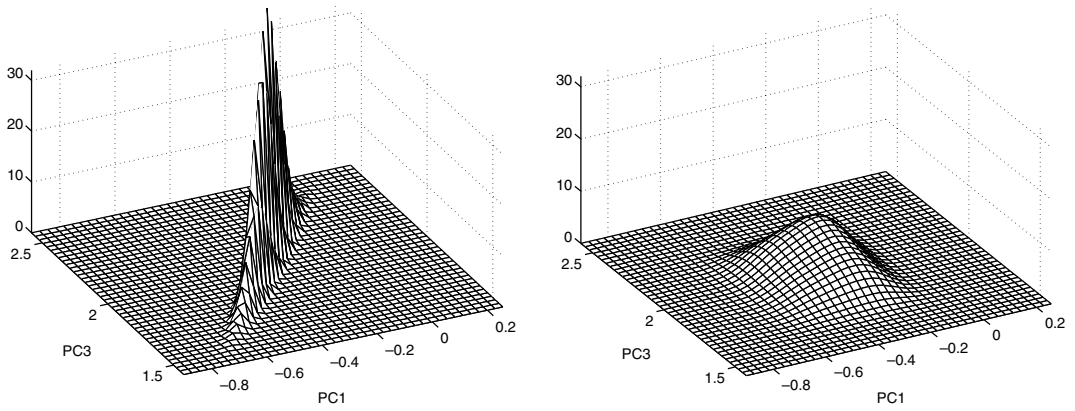

Fig. 3. The illustration of the advantage of the two-dimensional optimization algorithm over the direct product of two one-dimensional optimization algorithms in the case of two-dimensional distributions. Left – the optimal two-dimensional PDF, right – the direct product of one-dimensional PDF's. Moment constraints up to order 4 are used in both cases. The results are sampled from a forecast ensemble with high mutual information [6].

corresponding one-dimensional PDF's of the PC1 and PC3, also estimated using moments up to order 4. The inability of the latter method to make use of the second cross-moment between the PC1 and PC3 results in an inadequate approximation (a smooth PDF with no signs of mutual information), whereas the two-dimensional algorithm readily captures narrow band-like structure of the original sample. The rotation-stretching preconditioning of the constraints, described in Section 4, plays an important role by aligning one of the two coordinates with the narrow band of this sample.

6. Summary

The current work sets forth a practical computational approach for the moment-constrained maximum entropy principle in multidimensional phase space. First, the two optimization problems for general constraints are formulated in Section 2: the optimization problem of the first kind for the Shannon entropy, and the optimization problem of the second kind for the relative entropy. These two problems are later reformulated in Section 3 for the moment constraints with an arbitrary number of phase space coordinates and an arbitrary order of constraints. It is then shown in Theorem 1 that, for an appropriate form of the prior probability density in relative entropy, the optimization problem of the second kind can be reduced to the optimization problem of the first kind, thus justifying the common computational approach for both types of entropy. For the moment-constrained optimization problem of the first kind, further systematic formalization is used in Section 4 to achieve good dynamic range and robustness. Namely, encapsulated user-transparent diagonalization and normalization of the constraint covariance matrix are used as a preconditioning algorithm, and the Gauss–Hermite quadrature is used for calculation of both the gradient and Hessian of the optimization problem. The general suitability of this algorithm for practical applications is confirmed in the concurrent works [5–7]. Several interesting numerical results from [5,6] are presented in Section 5.

Acknowledgements

The author thanks Andrew Majda, Gregor Kovačič, Richard Kleeman, Joseph Biello, Kyle Haven, Christian Franzke and Daan Crommelin for helpful advice, comments and suggestions. The author also thanks Andrew Majda for supporting the author as a postdoctoral fellow during the last several years through the NSF Grants DMS-9972865 and CMG DMS-0222133, and the ONR Grant N00014-96-1-0043.

References

[1] T. Schneider, S. Griffies, A conceptual framework for predictability studies, J. Clim. 12 (1999) 3133–3155.
[2] M. Roulston, L. Smith, Evaluating probabilistic forecasts using information theory, Mon. Wea. Rev. 130 (2002) 1653–1660.
[3] R. Kleeman, Measuring dynamical prediction utility using relative entropy, J. Atmos. Sci. 59 (2002) 2057–2072.
[4] A. Majda, R. Kleeman, D. Cai, A framework for predictability through relative entropy, Meth. Appl. Anal. 9 (2002) 425–444.
[5] R. Abramov, A. Majda, Quantifying uncertainty for non-Gaussian ensembles in complex systems, SIAM J. Sci. Comp. 26 (2) (2003) 411–447.
[6] R. Abramov, A. Majda, R. Kleeman, Information theory and predictability for low frequency variability, J. Atmos. Sci. 62 (1) (2005) 65–87.
[7] K. Haven, A. Majda, R. Abramov, Quantifying predictability through information theory: small sample estimation in a non-Gaussian framework, J. Comp. Phys. 206 (2005) 334–362.
[8] I. Turek, A maximum-entropy approach to the density of states with the recursion method, J. Phys. C 21 (1988) 3251–3260.
[9] D. Ormoneit, H. White, An efficient algorithm to compute maximum entropy densities, Econ. Rev. 18 (2) (1999) 127–140.

[10] X. Wu, Calculation of maximum entropy densities with application to income distribution, J. Econ. 115 (2003) 347–354.
[11] P. Rao, Nonparametric functional estimation, Academic Press, New York, 1983.
[12] B. Silverman, Density estimation for statistics and data analysis, Chapman and Hall, London, 1986.
[13] D. Scott, Multivariate density estimation: theory, practice and visualization,, Wiley, New York, 1992.
[14] L. Mead, N. Papanicolaou, Maximum entropy in the problem of moments, J. Math. Phys. 25 (1984) 2404–2417.
[15] A. Tagliani, Hausdorff moment problem and maximum entropy: a unified approach, Appl. Math. Comp. 105 (1999) 291–305.
[16] R. Fletcher, Practical methods of optimization, Wiley, New York, 1987.
[17] R. Byrd, P. Lu, J. Nocedal, A limited memory algorithm for bound-constrained optimization, SIAM J. Sci. Comp. 16 (6) (1995) 1190–1208.
[18] Z. Wu, G. Phillips, R. Tapia, Y. Zhang, A fast Newton algorithm for entropy maximization in phase determination, SIAM Rev. 43 (4) (2001) 623–642.
[19] M. Abramowitz, I. Stegun, Handbook of mathematical functions with formulas, graphs, and mathematical tables, Dover, New York, 1972.