A Class of Numerical Algorithms for Efficient Approximation of Maximum Entropy Estimates of Probability Density Functions

A. Balestrino, A. Caiti, A. Noe', F. Parenti

Abstract— **A class of algorithms for approximation of the maximum entropy estimate of probability density functions on the basis of a finite number of sampled data is introduced. The algorithms are presented as a finite sequence in order of increasing accuracy and decreasing computational efficience; the last element of the sequence is the exact maximum entropy estimate. Numerical and applicative examples are reported.**

Keywords— **Maximum entropy, probability density function, estimation.**

I. Introduction

MANY engineering and scientific applications require the estimation of the probability density function (p.d.f.) of a random variable from a finite number of realizations. In measurement systems, for instance, such estimate gives a complete sensor characterization at different operating conditions. In other instances, one maybe interested in estimating the interval of variation of a given variable with a prescribed confidence level (as, for instance, "the 95% confidence interval"). In this case the whole probability density function of the variable has to be known a priori (e.g., gaussianity of the p.d.f.) or estimated from the available data.

In its seminal work, Jaynes [1] has introduced the principle of maximum entropy as the underlying theoretical basis to tackle the p.d.f. estimation problem when the a priori knowledge is only available through moments of the p.d.f itself. Jaynes approach leads to the most uniform (or unbiased) p.d.f. estimate conditioned on the available a priori information. From a computational point of view, the application of the maximum entropy principle leads to the casting and solution of a nonlinear optimization problem. Several variations of standard optimization algorithms have been implemented for the solution of such problem [2], [3], [4]. Related problems have also been investigated: [2] discusses the inverse problem of determining the set of constraints that optimally describes the observed samples accordingly to the MinMax measure [5]; [6] and [7] have discussed conditions on the moment constraints that guarantees the existence and uniqueness of a maximum entropy p.d.f.

Much less explored, at least to the authors knowledge, is the study of numerical schemes for the *approximation* of the maximum entropy estimate. Such study can be of interest in situations in which the p.d.f. is required on-line: if this is the case, the solution of the nonlinear optimization problem may be too computational demanding, while approximated solutions, obtained with faster computational schemes, may be more appropriate. One such situation, that has motivated the present research, is that of on-line localization and tracking of autonomous vehicles when measurement errors are unknown but bounded, with known worst case bound. Standard algorithms from set-membership theory are employed to determine the feasible set in which the vehicle is located [8], [9]; however, more information could be obtained by estimating on line the p.d.f. of some of the observed variables within the bounds determined by the set-membership algorithms. Since the estimate has to be produced on-line, efficiency in the numerical computation is critical for the proper integration of the p.d.f. estimate in the localization and tracking algorithms.

With this background and motivations, in this paper a class of approximating schemes for the maximum entropy estimate of the p.d.f. from a finite number of samples is introduced. The proposed algorithms are all based on the construction of the approximating function as a linear combination of basis functions; the basis functions are selected exploiting a partial amount of the available information on the problem. As the amount of problem-dependent information on the basis functions choice increases, the computational efficiency decreases and the approximation accuracy increases. When all the available information on the problem is used, the algorithm is reduced to the exact maximum entropy estimate. Some preliminary results on this line of research have been reported in [10], in particular comparing the approximations obtained and the corresponding computational time with simulated data. In this paper a more general formulation of the approach is proposed, and the methods are applied to field data obtained with a GPS system.

The paper is organized as follows: in the next section the problem is formally stated, the maximum entropy p.d.f. estimation approach is reviewed, and the numerical algorithm for exact estimation described; in section 3 the implemented approximating algorithms are described; in section 4 the algorithms are applied to elevation data obtained from a GPS system; finally, some conclusions are given.

Authors are with DSEA - Dept. Electrical Systems and Automation, University of Pisa, via Diotisalvi 2, 56100 Pisa, Italy; e-mail: aldo.balestrino,andrea.caiti@dsea.unipi.it

II. Problem statement

Let $f(x)$ be an unknown p.d.f. defined over a finite real interval $[a, b]$ and subject to the probabilty constraints:

$$\int_a^b f(x)dx = 1, \quad f(x) \geq 0 \; \forall x \in [a, b] \tag{1}$$

Let us suppose that k moment constraints on f are known in the form:

$$\int_a^b g_r(x)f(x)dx = a_r \quad r = 1, \ldots, k \tag{2}$$

with known functions $g_r(x)$ and known real constants a_r.

The maximum entropy estimate of $f(x)$ is obtained by maximization of the Shannon entropy $S(f)$ associated to f subject to the constraints given by equations (1) and (2), where the Shannon entropy is given by:

$$S(f) = -\int_a^b f(x) \ln f(x)dx \tag{3}$$

Jaynes [1] has shown that the maximization of $S(f)$ with respect to f, subject to the constraints (1) and (2), leads to the following analytical solution:

$$f(x) = e^{-\lambda_0 - \lambda_1 g_1(x) - \cdots - \lambda_k g_k(x)} \tag{4}$$

where the Lagrangian multipliers $\lambda_0, \cdots, \lambda_k$ satisfy the following relations:

$$\int_a^b \exp\left(-\sum_{j=1}^k \lambda_j g_j(x)\right) dx = \exp(\lambda_0) \tag{5}$$

$$\frac{\int_a^b g_r(x) \exp\left(-\sum_{j=1}^k \lambda_j g_j(x)\right) dx}{\exp \lambda_0} = a_r \quad j = 1, \cdots, k \tag{6}$$

From a practical point of view, the determination of a maximum entropy p.d.f. from available data is reduced to the solution of the nonlinear system of k equations (6). Although it has been shown that this is a nonlinear programming problem of polynomial complexity, and that known methods are available for its solution, the computational cost associated to the determination of the maximum entropy p.d.f is such to preclude an on-line use of the estimate (see [3] for a thourough discussion of several computational approaches to the determination of the maximum entropy p.d.f). The results obtained in this paper have been obtained by applying a standard Newton-Raphson method. In the next section some fast computational algorithms for the *approximated* solution of the system (6) are proposed.

III. A class of approximating algorithms

A series of algorithms to obtain approximated solutions to the system (6) are now described. It is assumed throughout the section that the functions g_r in equation (2) have the following form:

$$g_r(x) = x^r, \quad r = 1, \cdots, k \tag{7}$$

The available data are the real constants a_r in equation (2) The proposed algorithms are based on the approximation of the p.d.f. f (see equation (4)) with a linear combination of basis functions:

$$f(x) \approx \sum_{i=1}^k \beta_i f_i(x) = \hat{f}(x) \tag{8}$$

For any given set of basis functions $\{f_i\}$, the coefficients β_i are determined by solving the following system of *linear* equations:

$$\begin{cases} \beta_1 \int_a^b f_1(x)dx + \ldots + \beta_n \int_a^b f_n(x)dx = 1 \\ \beta_1 \int_a^b x f_1(x)dx + \ldots + \beta_n \int_a^b x f_n(x)dx = a_1 \\ \vdots \\ \beta_1 \int_a^b x^n f_1(x)dx + \ldots + \beta_n \int_a^b x^n f_n(x)dx = a_k \end{cases} \tag{9}$$

The algorithms differ in the choice of the set of basis functions. Before describing the choices suggested, it is important to underline the relation between the approximating function $\hat{f}$ and the true maximume entropy p.d.f. estimate f. Let the following notation be used, for any generic function $g(x)$ and any integer i:

$$E(g, x^i) = \int_a^b x^i g(x)dx \tag{10}$$

Then by construction:

$$E(f, x^i) = E(\hat{f}, i), \quad i = 1, \cdots, k \tag{11}$$

Since two functions are equal (but for a set of null measure) if all their moments are equal, it follows that $\hat{f} \to f$ as $k \to \infty$.

Algorithm $\mathcal{A}_0$: the basis functions f_i are taken as the Tchebycheff polynomials, after normalization of the $[a, b]$ interval to the $[0, 1]$ interval:

$$\begin{cases} f_1(x) = 1 \\ f_2(x) = x \\ f_j(x) = 2x f_{j-1}(x) - f_{j-2}(x) \qquad j = 2, 3, \ldots \end{cases} \tag{12}$$

With this choice of basis functions, which is independent from the available data (i.e., from the coefficients a_r), the linear system (9) can be directly solved.

Algorithm $\mathcal{A}_1$: the basis functions f_i are taken so that each of them is the solution of a simplified maximum entropy estimation problem involving *one* of the known moments; in particular, the basis function f_i is taken as solution of the following problem:

$$\left\{ \begin{array}{l} -\max \int_a^b f_i(x) \ln f_i(x) dx \\ \int_a^b f_i(x) dx = 1 \\ \int_a^b x^j f_i(x) dx = a_i \end{array} \right. \tag{13}$$

Applying Jaynes result to the problem (13), one obtains:

$$f_j(x) = e^{-\lambda_{0_j} - \lambda_j x^j} \quad j = 1, \cdots, k \tag{14}$$

where the λ_{0_j} and λ_j are the solution of the following nonlinear system:

$$\left\{ \begin{array}{rcl} \int_a^b e^{-\lambda_j x^j} dx & = & e^{\lambda_{0_j}} \\ a_j \int_a^b e^{-\lambda_j x^j} dx - \int_a^b x^j e^{-\lambda_j x^j} dx & = & 0 \end{array} \right. \tag{15}$$

The computational advantage of this approach is that, instead of solving one nonlinear system in k unknowns, one solves k nonlinear equations in one unknown, each one independent from the others. These equations can be potentially solved in parallel, though we have implemented the algorithm sequentially. After the k nonlinear equations have been solved, the functions f_i are determined, and the linear system (9) can be solved.

Algorithm $\mathcal{A}_2$: the basis functions f_i are taken so that each of them is the solution of a simplified maximum entropy estimation problem involving *two* of the known constraints; in particular, any basis function f_i is taken as solution of the following problem:

$$\left\{ \begin{array}{l} -\max \int_a^b f_i(x) \ln f_i(x) dx \\ \int_a^b f_i(x) dx = 1 \\ \int_a^b x^p f_i(x) dx = a_p \\ \int_a^b x^q f_i(x) dx = a_q \end{array} \right. \quad p, q \in [1, 2, \cdots, k] \tag{16}$$

Of course for each f_i a different couple (a_p, a_q) must be chosen. Each function f_i, $i = 1, \cdots, k$ solution of the problem (13) is again given through Jaynes formalism and the solution of a nonlinear system of dimension two. After each f_i has been determined, the linear system (9) can be solved. As compared to Algorithm $\mathcal{A}_1$, Algorithm $\mathcal{A}_2$ has an additional computational burden due to the need of solving k systems of nonlinear equations of dimension two instead of one; moreover it requires the choice of the couple (a_p, a_q) to be associated to each f_i. In our implementation this choice has been arbitrarily made; however, it may well be the case that some choices are to be preferred in terms of approximating precision or computational efficiency.

Following the same rationale, one can define Algorithms $\mathcal{A}_3, \mathcal{A}_4, \cdots, \mathcal{A}_k$. In the case of Algorithm $\mathcal{A}_k$, one comes back to the original maximum entropy estimate (4). It is interesting to note that, with a fixed number k of available moments, the class of algorithms proposed is a finite sequence coinciding with the true maximum entropy estimate at the end of the sequence. With a fixed algorithm index $j < k$, the algorithm $\mathcal{A}_j$ will produce an approximation $\hat{f}$ which is convergent to the maximum entropy estimate f as the number k of available moments tend to infinity.

One important point to note here is that in determining $\hat{f}$ with the procedures described above, the natural probability constraint of equation (1) is not enforced anymore. This loss of the probability constraints is due to the fact that $\hat{f}$ is an approximation of f. It has to be remarked, though, that $\hat{f}$ is convergent to f as the number of known moments k increases, and f does respect the natural probabilty constraints.

In the following section the approximating capabilities of the proposed algorithms and their computational efficiency will be investigated through application to field data from a GPS system.

IV. Field data application

The techniques described in the previous section have been applied to altitude data collected with the portable GPS 25LP Garmin. The sampling time of the instrument is one reading per second. The data reported in figure 1 have been gathered over a period of about 25 minutes, while the instrument was moving along a flat terrain (i.e., maintaining the same altitude) in the University of Pisa surrounding. The fluctuations in the altitude measurements are due not only to intinsic insturment uncertainties, but also to the different satellite reception quality due to the obstruction of city buildings. This is a common problem for GPS systems in urban environments.

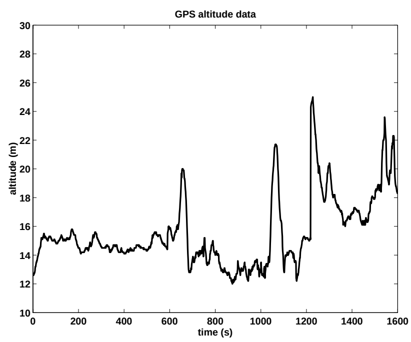

Fig. 1. Data samples obtained from the GPS instrument.

The true maximum entropy estimate has been computed using four data moments (i.e, using mean, variance, skewness and kurtosis information). In figures 2, 3 and 4 the results obtained using algorithms $\mathcal{A}_\prime, \mathcal{A}_\infty$ and $\mathcal{A}_\in$ respectively are shown; for each algorithm, the approximations obtained using two, three and four moments are shown. It can be seen that increasing the number of moments employed the approximation error decreases. Moreover, as expected, the smallest error with an equal number of mo-

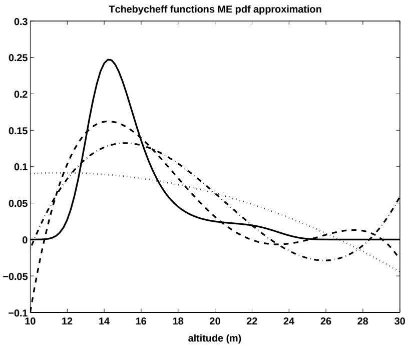

Fig. 2. Probability density function estimates from the data reported in figure 1 obtained with Algorithm 0, based on the use of Tchebycheff functions. Black line is the exact maximum entropy estimate obtained using four moments. Dotted, dash-dotted and dashed lines are the Tchebicheff approximations obtained using two, three and four data moments, respectively.

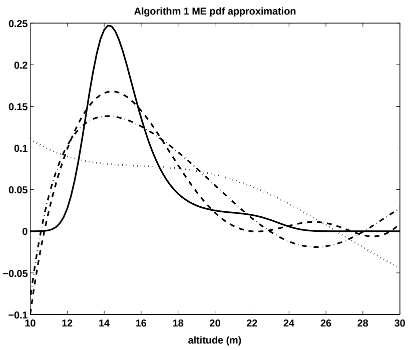

Fig. 3. Probability density function estimates from the data reported in figure 1 obtained with Algorithm 1. Black line is the exact maximum entropy estimate obtained using four moments. Dotted, dash-dotted and dashed lines are the Algorithm 1 approximations obtained using two, three and four data moments, respectively.

ments employed is obtained with algorithm $\mathcal{A}_{\in}$. Finally, in figure 5 the relative computational time of the approximating algorithms proposed are shown with respect to the computational time of the true maximum entropy estimate. It can be seen that in all case the reduction in computational time is quite effective, and it appears, at least in this case, that algorithm $\mathcal{A}_{\infty}$ offers the best compromise between accuracy and computational efficiency.

V. Conclusions

A class of algorithms for computationally efficient approximation of the maximum entropy estimate of probability density functions has been introduced. The proposed algorithms are based on the construction of the approximating function as a linear combination of basis functions; the basis functions are selected exploiting a partial amount of the available information on the problem. Application of the algorithms to experimental data has confirmed the theoretical expectations.

Acknowledgments

This work has been partially supported by MIUR.

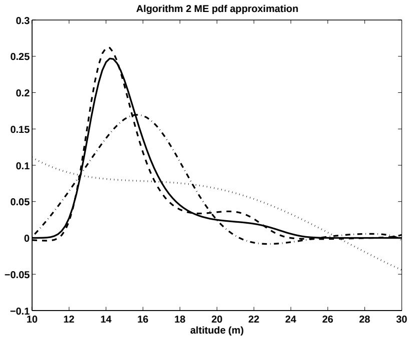

Fig. 4. Probability density function estimates from the data reported in figure 1 obtained with Algorithm 2. Black line is the exact maximum entropy estimate obtained using four moments. Dotted, dash-dotted and dashed lines are the Algorithm 3 approximations obtained using two, three and four data moments, respectively.

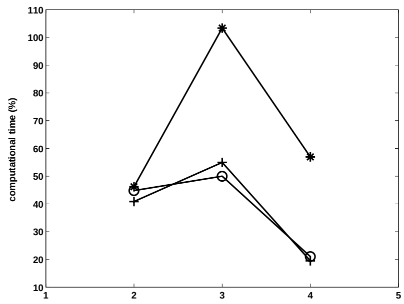

Fig. 5. Relative computational time of the approximating algorithms as a function of number of moments used, with the respect to the true Maximum Entropy estimate computed with the same number of moments. Cross is the computation time of Algorithm 0 (Tchebycheff functions), circle the computation time of Algorithm 1 and asterisk the computation time of Algorithm 2).

References

[1] E.T. Jaynes, "Information theory and statistical mechanics",*Phys. Rev*, vol. 106, pp. 361-373.
[2] M. Srikanth, H.K. Kesavan, P.H. Roe, "Probability density function estimation using the MinMax measure", *IEEE Trans. Sys. Man Cyber. - part C*, vol.30, n.1, pp. 77-82, 2000.
[3] X. Wu, "Calculation of maximum entropy densities with application to income distribution", in revision to *J. Econometrics*, available on line at http://are.berkeley.edu/x̃iming/, 2002.
[4] D. Ormoneit, H. White, "An efficient algorithm to compute maximum entropy densities", *Econometric reviews*, vol.18, n.2, pp.127-140, 1999.
[5] J.N. Kapur, G. Baciu, H.K. Kesavan, "The MinMax information measure", *Int. J. Sys. Sci.*, vol. 26, n.1, pp. 1-12, 1995.
[6] L.R. Mead, N. Papanicolau, "Maximum entropy in the problem of moments", *J. Math. Phys.*, vol.25, n.8, pp.2404-2417, 1984.
[7] A. Tagliani, "Maximum entropy in the discrete generalized moment problem", *Statistica*, vol. LX, pp. 59-72, 2000.
[8] M.Milanese, A. Vicino, "Information based complexity and nonparametric worst-case system identification", *J. Complexity*, vol. 9, pp. 427-446, 1993.
[9] A.Caiti, A.Garulli, F.Livide, D. Prattichizzo, "Set-membership acoustic tracking of autonomous underwater vehicles", *Acta Acustica/Acustica*, vol. 88, pp. 648-652, 2002.
[10] A.Balestrino, A.Caiti, A.Noe', F.Parenti, "Maximum entropy based numerical algorithms for approximation of probability density functions", submitted to *European Control Conference 2003*, Cambridge, UK, 2003.