Real-Time Construction of Optimized Predictors from Data Streams

Frank Kwasniok and Leonard A. Smith

Department of Statistics, London School of Economics, London, United Kingdom

(Received 13 May 2003; published 23 April 2004)

A new approach to the construction and optimization of local models is proposed in the context of data streams, that is, unlimited sources of data where the utilization of all observations is impractical. Real-time revision of the learning set allows selective coverage of regions in state space which contribute most to reconstructing the underlying dynamical system.

DOI: 10.1103/PhysRevLett.92.164101 PACS numbers: 05.45.Tp, 05.45.Ac

Introduction.—Predicting the evolution of a dynamical system is a common goal in science. Often the underlying dynamical equations are unknown and only a time series of observations is available. In the case of deterministic dynamical systems, there are well-established methodologies [1] based on state space reconstructions using local statistical models (e.g., local polynomial models) constructed from the observations. Given the fact that properties of nonlinear dynamical systems such as local curvature of the solution manifold and the local density of the invariant measure may vary enormously across state space, the question of optimizing such prediction schemes arises. Previous studies have optimized local model structure [2], and varied the complexity of the model (or data weighting) in a partitioned state space [3–5]. The present Letter considers the setting of a data stream, a continuing source of data such that retaining and processing all observations is impractical [6]. Our aim is to extract a learning data set of limited size optimized in terms of predictive power either in the context of a data stream or for a huge observational database. Two conceivable applications among others are the prediction of turbulent gusts in surface wind velocities [7] and grid frequency.

Our approach is contrasted with the traditional method in which the learning set is uniformly distributed with respect to the system's invariant measure. Our refined learning set adapts both to local curvature and to local data density; it is illustrated in the context of local linear prediction using the Ikeda map.

Methodology.—Consider a data stream of scalar values $\{s_n\}$ measured at equally spaced times $\{t_n\}$. The (unknown) underlying dynamical system may be either a discrete map or a continuous flow. We restrict attention to a single-channel measurement for simplicity; generalization to a multichannel situation is straightforward. The prediction problem at time t_n consists of predicting s_{n+1} from an m-dimensional time-delay vector $(s_{n-m+1}, \ldots, s_n)$ (see [1]). Let $\hat{s}_{n+1}$ denote the predicted value as opposed to the observed value s_{n+1}. We consider local linear modeling, that is, $\hat{s}_{n+1} = a_0 + \sum_{i=1}^{m} a_i s_{n-m+i}$, where the prediction coefficients $\{a_i\}_{i=0}^{m}$ are determined by least-squares regression on the points in some neighborhood of $(s_{n-m+1}, \ldots, s_n)$. These neighbors in delay space are chosen from a learning data set of prescribed size N, collected at times prior to t_n. Denote the learning set by $\mathcal{L} = \{(\mathbf{x}_\alpha, y_\alpha)\}_{\alpha=1}^{N}$ with $\mathbf{x}_\alpha = (s_{n_\alpha-m+1}, \ldots, s_{n_\alpha})$ and $y_\alpha = s_{n_\alpha+1}$.

A traditional learning set, hereafter $\mathcal{L}_0$, consists of N points distributed uniformly in time (say, $n_\alpha = \alpha + m - 1$), and thus (roughly) according to the invariant measure. We take $\mathcal{L}_0$ as a starting point for our learning set, $\mathcal{L}$ (that is, $\mathcal{L} = \mathcal{L}_0$ initially), and then refine $\mathcal{L}$ keeping N fixed via the following algorithm.

(1) Read the next point $(\mathbf{v}, w)$ from the data stream, where $\mathbf{v} = (s_{n'-m+1}, \ldots, s_{n'})$ and $w = s_{n'+1}$.

(2) Calculate the prediction $\hat{s}_{n'+1}$ based on the current refined learning set $\mathcal{L}$ and the absolute prediction error $\varepsilon' = |s_{n'+1} - \hat{s}_{n'+1}|$.

(3) Draw a point at random from $\mathcal{L}$, each point being equally likely, and denote it as $(\mathbf{x}_{\alpha^*}, y_{\alpha^*})$.

(4) Calculate the prediction $\hat{s}_{n_{\alpha^*}+1}$ with the learning set $\mathcal{L}^* = \mathcal{L} \cup \{(\mathbf{v}, w)\} \setminus \{(\mathbf{x}_{\alpha^*}, y_{\alpha^*})\}$ [that is, include $(\mathbf{v}, w)$ while excluding $(\mathbf{x}_{\alpha^*}, y_{\alpha^*})$] and corresponding prediction error $\varepsilon^* = |s_{n_{\alpha^*}+1} - \hat{s}_{n_{\alpha^*}+1}|$.

(5) If $\varepsilon^* < \varepsilon'$ then exchange $(\mathbf{x}_{\alpha^*}, y_{\alpha^*})$ for $(\mathbf{v}, w)$, that is take $\mathcal{L}^*$ as the new refined learning set; otherwise do not alter $\mathcal{L}$. Proceed to Step (1).

This algorithm aims to selectively include points in $\mathcal{L}$ from regions in reconstruction space with the largest errors, at the cost of removing points in regions where prediction is relatively good; it immediately generalizes to analogue prediction, higher order local polynomial prediction, or other local models. The ultimate precision of the model is limited by technological constraints, effectively the value of N. The algorithm can be run indefinitely.

Let ρ denote the exchange probability, that is, the probability that a new point is included in $\mathcal{L}$. Let $p(\varepsilon)$ be the probability distribution of out-of-sample absolute prediction errors at some stage of the algorithm and $\Phi(\varepsilon)$ the corresponding cumulative distribution,

 0031-9007/04/92(16)/164101(4)$22.50

while $p^*(\varepsilon)$ and $\Phi^*(\varepsilon)$ denote the corresponding in-sample distributions. The exchange probability is then $\rho = \int_0^\infty p(\varepsilon)\Phi^*(\varepsilon)\,d\varepsilon$ and, since p and p^* are initially the same, the initial exchange probability is $\rho_0 = \int_0^\infty p(\varepsilon)\Phi(\varepsilon)\,d\varepsilon = \frac{1}{2}$, independent of the distribution. Depending on operational constraints, one could, of course, repeat steps (3), (4), and (5) above, thereby testing each new point against several in $\mathcal{L}$ and increasing the probability of changing $\mathcal{L}$. In the cases presented below, this modification had little effect and is not pursued further in this Letter.

Results.—We illustrate the methodology with the Ikeda map [8]. In the complex plane, the map is $z_{n+1} = p + Bz_n \exp[i(\kappa - [\alpha/(1 + |z_n|^2)])]$, with $p = 1$, $B = 0.9$, $\kappa = 0.4$, and $\alpha = 6$. The data stream consists of the scalar observable $s_n = \mathrm{Re}(z_n)$, generated by numerical iteration of the map. In the noise-free case, local linear models were constructed from time-delay vectors formed with $m = 4$ using $K = 10$ nearest neighbors; K is then twice the number of free parameters in the model. We use the Euclidean norm throughout. The variance of the prediction error is drastically enhanced by occasional ill-conditioned fits. Eigenvectors of the predictor-predictor covariance matrix corresponding to eigenvalues smaller than 10^{-5} times the largest eigenvalue were omitted in the noise-free case (in the cases with observational noise discussed below, only the three largest eigenvalues were included) [9].

The results shown are means over 40 independent realizations using different data streams; also the initial learning sets and the test sets differed in each realization. Figure 1 outlines the one-step ahead performance of the algorithm with $N = 1500$. The mean absolute (MA) error and the root mean square (rms) error are given as a function of k, the number of points processed after the initial learning set. Thus, $k = 0$ corresponds to the traditional learning set $\mathcal{L}_0$. Both the MA error and the rms error are calculated out-of-sample using 5000 points. Sampling fluctuations in the estimated MA and rms errors were of order 0.0001; hence, the visible differences are almost entirely due to the two different algorithms used. Figure 1 shows the mean error and the central interval containing 95% of the realizations. The mean prediction error of a learning set consisting of all $N + k$ data points is also shown as an indication of ideal model performance; in realistic applications with large k this comparison is impractical. Initially, both the MA error and the rms error decrease rapidly as the algorithm proceeds, saturating after about 10 000 and 15 000 points, respectively, having dropped by 51% and 74%, respectively. Moreover, our algorithm has the additional advantage that the variability of both quantities between different realizations decreases with increasing k. Note that, with respect to rms error, the refined method remains fairly close to the learning set retaining all the available data;

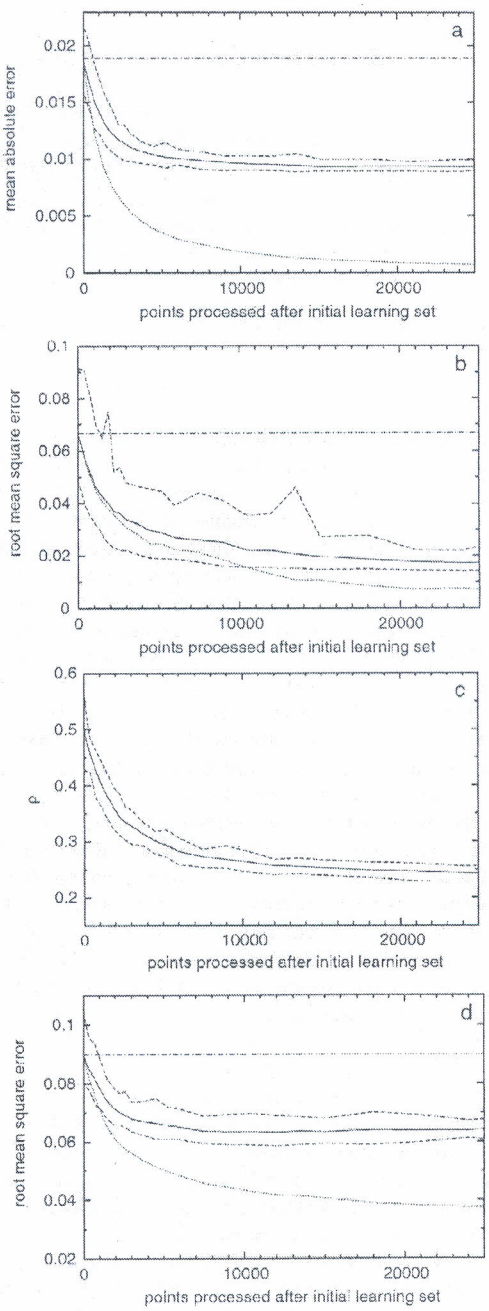

FIG. 1. Mean absolute error (a), rms error (b), and exchange probability (c) as a function of k for one-step prediction with $N = 1500$. Dotted curves reflect error when keeping all observations. The error in the traditional method is denoted by the horizontal dash-dotted line. Solid, dotted, and dash-dotted curves are means over 40 realizations; dashed curves capture the central 95% of the distribution. rms error (d) with 2% observational noise on the data stream.

the refined learning set with $N = 1500$ points yields forecasts with accuracy similar to that of a traditional learning set a factor of 7 larger (10 500 points).

Figure 2 illustrates the evolution of the cumulative distribution of absolute prediction errors. The distribution of errors obtained with $\mathcal{L}_0$ is contrasted with that reached after 1500 ($k = N$), 15 000 ($k = 10N$), and 30 000 ($k = 20N$) additional observations. Large errors decrease greatly and small errors increase slightly, converging in distribution after about 15 000 points. The refined learning set $\mathcal{L}$ provides a more homogeneous error distribution and substantially reduced mean error which would almost always be preferable to that of $\mathcal{L}_0$.

Figure 3 contrasts both the one-step and the (iterated) two-step forecast quality of models using $\mathcal{L}_0$ with that of models using $\mathcal{L}$ at $k = 20N$ (when the error distribution has stabilized). The ratio between the forecast error using $\mathcal{L}$ and that using $\mathcal{L}_0$ is plotted as a function of N for three error measures: the MA error, the rms error, and the 90% quantile of the error distribution. With $N = 5000$, the refined method improves the one-step MA error by 57% while the rms error is reduced by 87%; as shown, a significant reduction remains at two steps. With $N = 1500$, $\mathcal{L}$ maintains a systematic advantage over $\mathcal{L}_0$ as the forecasts are iterated into the future (not shown); at one, two, three, and four steps, the rms error is reduced by 74%, 49%, 15%, and 4%, respectively. As expected, larger values of N enhance these reductions, as well as introducing significant reductions at longer lead times.

There remains the question of noise. Given a data stream contaminated with additive Gaussian noise ($\sigma = 0.0097$, corresponding to a noise level of 2%), local linear models were formed by using the ten nearest neighbors and all other points (if any) within a distance of $\sqrt{m}\sigma$. For $N = 1500$, Fig. 1(d) illustrates that by $k \approx 10\,000$ the rms error has saturated at a value 29% below that of $\mathcal{L}_0$. Iterated forecasts again show improvement; in the noisy case using $\mathcal{L}$ with $N = 1500$ yields reductions of rms error by 29%, 22%, 14%, and 5% at one, two, three, and four steps, respectively. Even at only four steps ahead, rms and MA error are misleading tools for model evaluation; evaluating probabilistic forecasts or shadowing times yields more insight [10]. Nevertheless, these systematic forecast improvements are substantial, indicating that our method captures the dynamics of the system noticeably better than the traditional approach.

In its present form, the algorithm has no mechanism to distinguish whether a large prediction error is due to local nonlinearity or a particularly noisy observation. Hence, the algorithm is expected to accumulate observations with extreme realizations of the noise. This is indeed visible at $k \approx 50\,000$ (not shown). An elegant solution to this problem is under investigation; however, simply monitoring the prediction error over a large window provides a simple stopping criterion. Alternatively, reasons for not employing any stopping criteria are considered below. In any event, the algorithm is robust to low levels of additive noise.

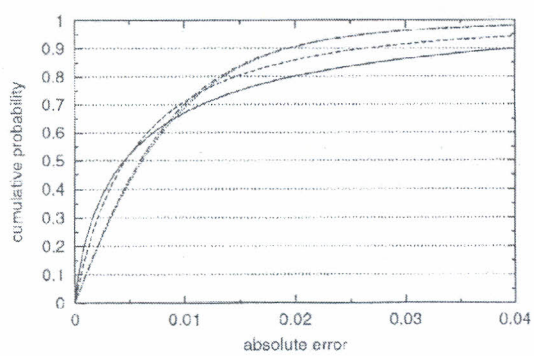

FIG. 2. Cumulative probability distribution of one-step absolute prediction errors for $N = 1500$ with $k = 0$ (solid line), $k = 1500$ (dashed line), $k = 15\,000$ (dash-dotted line), and $k = 30\,000$ (dotted line). Each curve represents an average over 40 realizations.

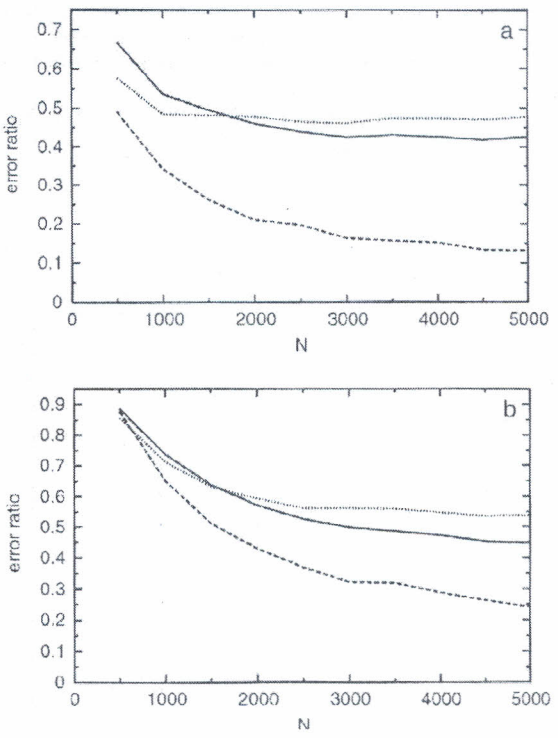

FIG. 3. Ratio of prediction errors of the refined method to the traditional method as a function of N for one-step (a) and two-step iterated (b) prediction showing mean absolute error (solid line), rms error (dashed line), and 90% quantile of absolute error (dotted line).

Discussion.—We have shown that learning sets that selectively sample state space can improve prediction. This approach is complementary to methods which explicitly partition the state space and either (i) increase the complexity of the model in poorly forecast regions by growing cell structures [4], using cluster-weighted models [5] or employing multiple models, or (ii) fix the model complexity and reweight the data [3]. Network approaches [4,5] have the advantage that clustering (or vector quantization) offers efficient noise reduction. Alternatively, the high degree of locality inherent to nearest-neighbor schemes yields a very high effective model complexity which is an advantage, as long as it is not employed to fit the noise. A systematic comparison of these approaches is desirable.

Another advantage of local models is the ease with which they can be updated by merely changing the learning set. Our method is easily generalized to consider the age (or the expected value of inclusion) of a point in $\mathcal{L}$, allowing application to nonstationary systems, including those undergoing gradual parameter shifts, as well as addressing the noise issues above. By construction, our method continuously adapts the learning set to the system.

We have introduced a general method for refining local models of data streams, and demonstrated improved short-term prediction. We hope this approach will prove useful in practical application.

We thank P. E. McSharry and K. Judd for detailed discussions and implementations of earlier algorithms.

[1] J. D. Farmer and J. J. Sidorowich, Phys. Rev. Lett. **59**, 845 (1987); G. Sugihara and R. M. May, Nature (London) **344**, 734 (1990); T. Sauer, J. A. Yorke, and M. Casdagli, J. Stat. Phys. **65**, 579 (1991); H. D. I. Abarbanel, R. Brown, J. J. Sidorowich, and L. S. Tsimring, Rev. Mod. Phys. **65**, 1331 (1993); H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, England, 1997).

[2] L. A. Smith, Philos. Trans. R. Soc. London, Ser. A **348**, 371 (1994); E. M. Bollt, Int. J. Bifurcation Chaos Appl. Sci. Eng. **10**, 1407 (2000); K. Judd and M. Small, Physica (Amsterdam) **136D**, 31 (2000).

[3] L. A. Smith, Physica (Amsterdam) **58D**, 50 (1992).

[4] B. Fritzke, Neural Networks **7**, 1441 (1994).

[5] N. Gershenfeld, B. Schoner, and E. Metois, Nature (London) **397**, 329 (1999).

[6] P. E. McSharry, Ph.D. thesis, Oxford, 1999).

[7] M. Ragwitz and H. Kantz, Europhys. Lett. **51**, 595 (2000).

[8] K. Ikeda, Opt. Commun. **30**, 257 (1979).

[9] Predictions outside the observed range were rescaled: Those smaller than $s_{\min} = -1$ or larger than $s_{\max} = 2.5$ were set to $s_{\min}$ and $s_{\max}$, respectively. This tends to *underestimate* the relative value of our approach.

[10] C. Grebogi, S. M. Hammel, J. A. Yorke, and T. Sauer, Phys. Rev. Lett. **65**, 1527 (1990); T. Palmer, Rep. Prog. Phys. **63**, 71 (2000); L. A. Smith, in *Nonlinear Dynamics and Statistics*, edited by A. I. Mees (Birkhauser, Boston, 2000); M. S. Roulston and L. A. Smith, Mon. Weather Rev. **130**, 1653 (2002).