WIP: Towards Optimal Online Approximation of Data Streams

Phillip Sitbon, Nirupama Bulusu, Wu-chi Feng
Portland State University
E-mail: {sitbon,nbulusu,wuchi}@cs.pdx.edu

***Abstract*—In this paper, we provide a basic solution for online compression of data streams using error-bounded piecewise-linear approximation (PLA). We compare this method to the optimal (but offline) solution. Our current work in progress is developing an online PLA method that meets the same optimality constraints as the offline method. Also, the vertices of the constructed approximations are subsets of the sampled data points, which we believe to be a benefit in many scenarios.**

I. Introduction

In many applications, data is presented as a continuous stream in which the most recent data is expected to be available with low latency from the source. For example, stock market data is most useful in a one minute window, and traffic data more than an hour old often cannot represent current conditions accurately. In simplistic terms, network applications can take several online approaches toward handling data streams, some of which may resemble the following:

- Pass every value without manipulation: generates large network traffic.
- Average data over a fixed period: reduces network traffic but also data accuracy corresponding to the averaging period.
- Maintain a *sliding window*: only transmit data when significant changes occur, thus balancing network traffic and data accuracy.

Resources are often constrained in terms of power (wireless sensor networks), cost (3G cellular networks), or capability (acoustic networks). Additionally, many-to-one networking applications can suffer from *DDOS* (Distributed Denial-of-Service) effects as scale increases, thus requiring undesired reductions in data quality in order to maintain scalability. For these reasons, there is significant motivation to reduce data volume while maintaining robustness.

Although many solutions to online stream compression exist, they provide an optimally small number of data points at the cost of throwing out the original data; conversely, optimal approximations using connected segments of actual data points are calculated offline.

Our work in progress is developing a piecewise linear method to approximate data streams with a minimal subset while also providing an accuracy guarantee. Providing a minimal subset requires choosing data points only from the generated sets without introducing averaged or otherwise interpolated points. This allows for the up-front transmission and processing of approximated data; additionally, untransmitted data can be retained for later transmission when network traffic is lower or when power reserves are restored. Because all data values are original sensor readings, any additional data will "fill in the gaps," thus providing additional resolution for statistical analysis and reporting.

In this paper, we define the *optimal* (minimal subset) measure of a data stream, which is calculated in an offline manner using a dynamic programming algorithm. We then devise an online method of piecewise linear approximation. It is a greedy approach in which only changes beyond a given error bound are recorded. To evaluate this method, we use uniform two-dimensional position data; however, these methods can also be applied to time-series sensor readings.

Our preliminary results indicate that the greedy approximation method achieves impressive compression ratios for mobility data, while still providing the benefits of online data streaming. Our eventual goal with this work is to provide a data stream that achieves the optimal compression for any given error bound and, unlike current solutions, is also online. If possible, we will present preliminary mathematical results in our endeavors.

II. Related Work

Significant effort has been dedicated to approximating sequential data within a guaranteed bound, for time-series data [1]–[3] and higher-dimensional data as well [4]. Some approximation methods are more indirect when related to networking, such as fuzzy [5], [6] and aggregation [5], [7], [8] methods, but almost all perform some form of linear fit.

Common terminology for sequential data approximation includes *filters*, such as *swing filters* or *slide filters* [2]. Elmeleegy et al. define slide filters as disjoint piecewise approximations that are sequentially adjusted in order to minimize residual error, and this method is most similar to our greedy approximation method. Kiely et al. propose an "Adaptive Linear Filtering Compression" algorithm as a lossless compression algorithm for sensor networks, in which the filter aspect is used to predict sample values which are corrected in later transmissions if wrong [9]. In our work, we chose to keep all data segments connected without reverse correction or data prediction. Because of this, we are able to choose only actual data points from data sources without interpolating new points to facilitate an increase in approximation accuracy.

Keogh et al. define a *sliding window* method in their discussion and consider it in the more common notion, although it is not similar to *sliding filters* above [1]. It is, however, very similar to the greedy algorithm used here and the authors also mention that it is widely used due to its online nature, for example in frequent-patterns discovery [10]. Keogh et al.'s

approximation method, SWAB (*sliding-window and bottom-up*), performs greedy approximation but keeps a buffer in order to backtrack and refine approximations within a certain window. Gandhi et al. propose a generic form of a greedy bucket merging method in order to approximate time series data quickly and in a near-optimal fashion, and as a result are able to represent provable error bounds [11]. Soroush et al. use a piecewise linear approximation on online data and apply it to an actual sensor testbed [3]. They achieve an optimally small approximation because the resulting line segments are not a subset of the original data. An optimal result consisting of only original data points was proposed by Dunham [12] using the L_∞ norm as an error measure. Additionally, Dunham devised a *scan-along* approach for the optimal method; however, it was not discussed or evaluated.

III. Data Series Approximation

A. Problem Statement

For an ordered set of n vectors

$$D = \{\hat{d}_1, \hat{d}_2 \cdots \hat{d}_n\}, \text{ with } \hat{d}_i = \langle x_i, y_i \rangle$$

define A_ϵ^D as the subset of D approximated with a maximum linear interpolation error of ϵ.

$$A_\epsilon^D = \{\hat{a}_1, \hat{a}_2 \cdots \hat{a}_m\}.$$

Exact points are chosen from the original data set in order to facilitate a finite optimal solution space and make the online approximation efficiently computable. Therefore, $A_\epsilon^D \subseteq D$ and the ordering of D is preserved in A_ϵ^D. Furthermore, we assume $n \geq 2$ in order to require that $\hat{a}_1 = \hat{d}_1$ and $\hat{a}_m = \hat{d}_n$ where $1 < m \leq n$.

The linear interpolation error from any point $\hat{d}_i$ not in A_ϵ^D is the euclidian distance between $\hat{d}_i$ and $\hat{d}'_i$, where $\hat{d}'_i$ is the orthogonal projection of $\hat{d}_i$ onto $\overline{\hat{a}_j\hat{a}_k}$, the line segment between the two elements in A_ϵ^D nearest (by index) to $\hat{d}_i$ in D. Typically, this is defined as $\|\hat{a}_j\hat{d}_i \times \overline{\hat{a}_j\hat{a}_k}\|$; however, this does not limit the distance to the line *segment*. Therefore, we define a function, $dist$, in similar vein but measuring at a fixed maximum distance from the endpoints as well (see [12],Fig. 2 for a similar approach).

B. Optimal Approximation

Here, we define A_ϵ^D as a recursive relation in order to portray the dynamic programming solution to the optimal subset given ϵ. First, we define $A_\epsilon^D = M(1, |D|)$ and

$$M(i,k) = \begin{cases} \{d_i, d_k\} & dist(\overline{\hat{d}_i\hat{d}_j}, \overline{\hat{d}_i\hat{d}_k}) < \epsilon \\ & \forall j \in (i,k) \\ \min_{||}\limits_{j\in(i,k)} (M(i,j) \cup M(j,k)) & \text{otherwise.} \end{cases} \quad (1)$$

where $\min_{||}(M(i,j) \cup M(j,k))$ is defined as the union of $M(i,j)$ and $M(j,k)$ having the smallest set size over j. $M(i,k)$ produces a minimal subset bounded by ϵ as a result of the principle of optimality, given the initial case:

$$M(i, i+1) = \{\hat{d}_i, \hat{d}_{i+1}\},$$

and for three consecutive points, there are two possibilities, both of which are minimal subsets and bounded by ϵ:

$$M(i, i+2) = \begin{cases} \{\hat{d}_i, \hat{d}_{i+2}\} & dist(\overline{\hat{d}_i\hat{d}_{i+1}}, \overline{\hat{d}_i\hat{d}_{i+2}}) < \epsilon \\ \{\hat{d}_i, \hat{d}_{i+1}, \hat{d}_{i+2}\} & \text{otherwise.} \end{cases}$$

Because the construction recursively depends on the last element of D, any additional data points have the possibility of changing the entirety of A_ϵ^D, thus rendering this method of approximation an offline solution. The caveat to this is mentioned by Dunham [12], where there exists no set of bounded lines through certain points after extremes are reached. However, in implementing such a method we found that early segmentation decisions led to slightly sub-optimal results.

C. Greedy Online Approximation

This algorithm applies a *sliding window* to the data set. In In Algorithm 1, the window starts at item $\hat{d}_i$ in the data set and iterates forward to item $\hat{d}_j$. If the points in the interval (i,j) are bounded by the line $\overline{\hat{d}_i\hat{d}_j}$, j is advanced further. Otherwise, $\hat{d}_{j-1}$ is added to A and iteration continues with $i = j-1$. This algorithm is *online* because for any data item $\hat{d}_t$ corresponding to time t, the algorithm only operates on $\hat{d}_n$ where $n < t$.

Algorithm 1 Pseudo code for greedy online approximation.

```
let A_ε^D = {d̂_0}
let start = 1, last_valid = 2

for d_i in D − {d̂_1, d̂_2} do
  if line d̂_start d̂_i bounded by ε for {d̂_{start+1} ... d̂_{i−1}} then
    last_valid = i
  else
    A_ε^D = A_ε^D ∪ {d̂_{last_valid}}
    start = last_valid
    last_valid = last_valid + 1
  end if
end for
```

IV. Results & Outlook

Table I shows approximated set sizes as percentages and number of points for both the greedy approximation method on real GPS mobility traces and the optimal approximation calculated on the entire data sets. For this mobility data, our hybrid online algorithm produces a near-optimal compression ratio. Figure 1 shows the compression ratios for different error bound values for both the greedy online and optimal offline compression methods. The greedy approximation method was nearly always within 3% of optimal for all data sets.

When choosing an optimal approximation for a given error bound, multiple solutions are possible. We provide order to the set of solutions by choosing the solution with the lowest average linear interpolation error. Again for the skiing data set, Figure 2 shows the average linear interpolation error of the entire data set for each value of ϵ, showing that the greedy solution provides near-optimal average error. This, however,

Data Set	Greedy				Optimal			
	$\epsilon = 0.25m$		$\epsilon = 5.0m$		$\epsilon = 0.25m$		$\epsilon = 5.0m$	
	Percent	Points	Percent	Points	Percent	Points	Percent	Points
JOGGING	79.20%	335	17.2%	73	78.72%	333	15.8%	67
CYCLING	43.18%	2262	5.90%	309	39.50%	2069	4.70%	237
DRIVING1	13.48%	577	2.10%	90	12.78%	547	1.92%	82
DRIVING2	16.34%	957	2.15%	126	14.44%	846	1.98%	116
DRIVING3	28.72%	725	3.88%	98	26.19%	661	3.37%	85
DRIVING4	28.51%	825	3.73%	108	25.72%	744	3.35%	97
SKIING	47.26%	5239	4.97%	551	44.03%	4881	3.50%	388

TABLE I
GREEDY AND OPTIMAL APPROXIMATION SET SIZES, REPRESENTED AS PERCENT OF ORIGINAL DATA SET SIZE AND THE EQUIVALENT NUMBER OF POINTS FOR $\epsilon = 0.25m$ AND $\epsilon = 5.0m$.

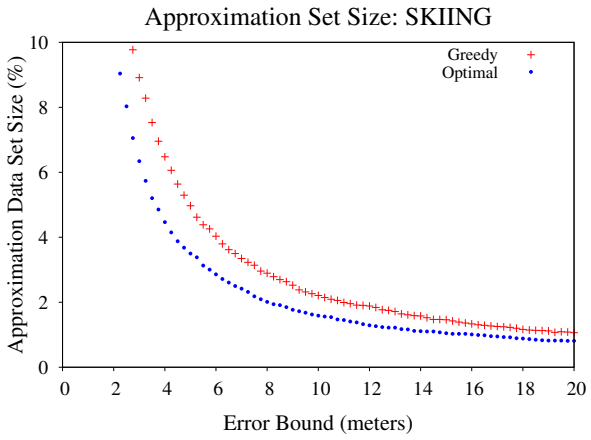

Fig. 1. Individual plots for SKIING approximated data set size given certain error bounds. Sizes above 10% have been removed for readability.

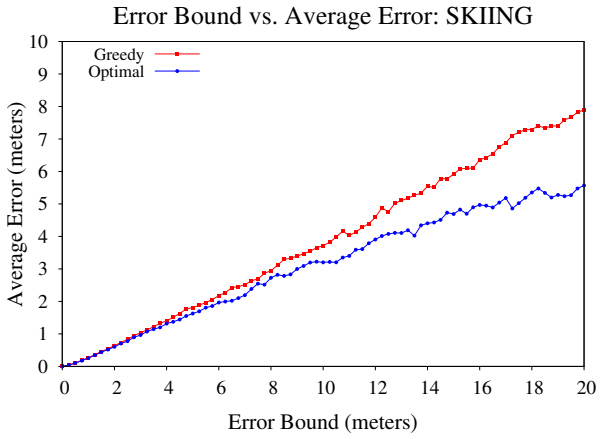

Fig. 2. Individual plots for SKIING approximated data set size given certain error bounds. Sizes above 10% have been removed for readability.

is where there will be notable improvement by developing an online optimal solution.

Our primary goal, and next step, is to design an online version of the optimal approximation dynamic programming algorithm that also minimizes an error measure (such as average) for a given solution. We believe this will be possible by defining data subsets in which future data cannot recursively affect their representation significantly. To our knowledge, this will be the first such algorithm with these characteristics. Later on, we will quantify data savings at a massive scale by simulating the large vehicular network implemented in [13]. Also of interest is study into lowering complexity and parallelizing the approximation methods.

REFERENCES

[1] E. Keogh, S. Chu, D. Hart, and M. Pazzani, "An Online Algorithm for Segmenting Time Series," in *Data Mining, IEEE International Conference on*, vol. 0. Los Alamitos, CA, USA: IEEE Computer Society, 2001, p. 289.

[2] H. Elmeleegy, A. K. Elmagarmid, E. Cecchet, W. G. Aref, and W. Zwaenepoel, "Online piece-wise linear approximation of numerical streams with precision guarantees," in *Proc. VLDB Endow.*, vol. 2, no. 1. VLDB Endowment, 2009, pp. 145–156.

[3] E. Soroush, K. Wu, and J. Pei, "Fast and quality-guaranteed data streaming in resource-constrained sensor networks," in *Proceedings of the 9th ACM international symposium on Mobile ad hoc networking and computing - MobiHoc '08*. New York, New York, USA: ACM Press, 2008, p. 391.

[4] P. S. Heckbert and M. Garland, "Survey of Polygonal Surface Simplification Algorithms, Multiresolution Surface Modeling Course," in *Proceedings of the 24th International Conference on Computer Graphics and Interactive Techniques*, 1997.

[5] S. Dietzel, B. Bako, E. Schoch, and F. Kargl, "A fuzzy logic based approach for structure-free aggregation in vehicular ad-hoc networks," in *VANET '09: Proceedings of the sixth ACM international workshop on VehiculAr InterNETworking*. New York, NY, USA: ACM, 2009, pp. 79–88.

[6] J. Pittman and C. A. Murthy, "Fitting Optimal Piecewise Linear Functions Using Genetic Algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 7, pp. 701–718, 2000.

[7] C. Chen, "Location-Based Data Aggregation in Mobile Ad Hoc Networks," Master's thesis, Institut fr Parallele und Verteilte Systeme, 2003.

[8] I. Timko, M. H. Böhlen, and J. Gamper, "Sequenced spatio-temporal aggregation in road networks," in *EDBT '09: Proceedings of the 12th International Conference on Extending Database Technology*. New York, NY, USA: ACM, 2009, pp. 48–59.

[9] A. B. Kiely, M. Xu, W.-Z. Song, R. Huang, and B. Shirazi, "Adaptive Linear Filtering Compression on realtime sensor networks," *2009 IEEE International Conference on Pervasive Computing and Communications*, pp. 1–10, Mar. 2009.

[10] K.-F. Jea and C.-W. Li, "A Sliding-window Based Adaptive Approximating Method to Discover Recent Frequent Itemsets from Data Streams," in *IMECS 2010: Proceedings of the Internation MultiConference of Engineers and Computer Scientists*, vol. 1. IAE, 2010.

[11] S. Gandhi, L. Foschini, S. Suri, and U. Santa Barbara, "Space-efficient Online Approximation of Time Series Data: Streams, Amnesia, and Out-of-order," in *26th International Conference on Data Engineering*. Long Beach, California, USA: IEEE, 2010.

[12] J. Dunham, "Optimum uniform piecewise linear approximation of planar curves," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 1, pp. 67–75, 1986.

[13] P. Sitbon, W.-c. Feng, and N. Bulusu, "TTN: A time-to-network approach to data reporting in mobile ad hoc networks," in *WoWMoM 2010: International Symposium on a World of Wireless, Mobile and Multimedia Networks*. IEEE, 2010, pp. 1–9.

[14] D. Douglas and T. Peucker, "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature," *Canadian Cartographer*, vol. 10, no. 2, pp. 112–122, 1973.

[15] G. Manis, G. Papakonstantinou, and P. Tsanakas, "Optimal piecewise linear approximation of digitized curves," in *Digital Signal Processing Proceedings, 1997. DSP 97., 1997 13th International Conference on*, vol. 2. IEEE, 1997, pp. 1079–1081.