

Extending Enclosure Arithmetic to Statistical Regular Sub-pavings

Principal Investigator: Dr. Raazesh Sainudiin
Department of Mathematics and Statistics,
University of Canterbury,
Private Bag 4800, Christchurch, New Zealand
r.sainudiin@math.canterbury.ac.nz
Telephone: +64 3 364 2987 ext 7691

December 15, 2009

1 Research Abstract and Goals

The quantity and complexity of data collected in most disciplines today is overwhelming. Computer-aided statistical decisions based on standard methods and theory will often be impractical, inappropriate or ineffective for such massive datasets. Limitations on machine memory and speed of access and operations pose physical challenges for massive data processing, especially in the presence of updates. There is an urgent need to develop theory and methods that account for the engineering constraints on representing and processing massive data in computing machines as highlighted in [7]. We propose to use set-valued mathematics [1, 6, 12], recursively computable statistics [2, 4, 9], multi-dimensional metric data-structures [6, 15, 16] and extended enclosure arithmetics to produce memory-limited and statistically consistent enclosures of non-parametric density and plug-in estimates for static as well as dynamic massive data problems. We restrict our proposed study to problems involving observations x_1, \dots, x_n in some m -dimensional metric-space, i.e. $x_i \in \mathbb{R}^m$, where $m \leq 50$ and n is typically more than the available primary machine memory. Our goal is to build a stable version of *MRS: A C++ class library for statistical set processing* that is generic, thread-safe and exactly implements our computational statistical theory.

keywords: Information retrieval, extraction, and organization; Machine learning and data mining; Structured data and database management.

2 Technical description

Ongoing research (e.g. [5, 8, 10, 11, 14, 18]) seeks efficient data-structures to organize and extract useful information from massive datasets for specific decision problems. Our proposal is similar to the above approaches, in terms of using an intermediary data-structure. We represent massive multi-dimensional metric data using dynamic statistical regular sub-pavings (*SRSPs*). Regular sub-pavings [6, 15], also called n -trees [16], are a class of partitions of a root box \mathbb{X} in \mathbb{R}^m . Recursive bisections of the root box \mathbb{X} along the first longest dimension allow regular sub-paving to be equivalent to ordered binary trees that are closed under efficient union, intersection and non-minimal union operations. As data passes through \mathbb{X} , we adaptively and regularly pave or depave \mathbb{X} by bisecting its leaves or reuniting its cherries (a pair of sibling leaves), respectively, while mutably caching the recursively computable statistics, such as counts, mean vector and/or variance-covariance matrix, at the nodes of the tree corresponding to the sub-boxes of \mathbb{X} (akin to [11]). Thus, *SRSP* augments the regular sub-paving with recursively computable statistics in order to dynamically and sufficiently condense enclosures of massive datasets for our decision problems. In our proposed study, we will extend enclosure arithmetic to functions and functionals over *SRSPs* by exploiting the structure of the underpinning ordered binary trees. Our approach will differ from other ongoing research in the following three ways.

Firstly, our *SRSP* is data-driven and stochastically dynamic. It is constructed as a discrete time irreducible Markov chain $\{S(i)\}_{i \in \mathbb{Z}_+}$ over $\mathbb{S}_{\hat{L}}$, the space of all *SRSPs* with a maximum of \hat{L} leaves. The leaves are split and the cherries are pruned as data passes through \mathbb{X} using proposal distributions over leaves and cherries that are functions of the recursively computed statistics and/or the functional(s) of interest. Proposal distributions are cast as randomized

priority queues with appropriate priority functions. In the context of density estimation, the stationary distribution of $\{S(i)\}_{i \in \mathbb{Z}_+}$ is the posterior density over histograms on $\mathbb{S}_{\hat{L}}$ and the proposal distributions use randomized priority queues to enforce the principle of statistically equivalent blocks [3] to dynamically optimize available primary machine memory. Thus, we develop a Markov chain over $\mathbb{S}_{\hat{L}}$ that is driven by statistically consistent randomized priority-queues capable of adaptively pruning and growing the $SRSP$ tree by using the recursively computable statistics at the leaf and cherry nodes/sub-boxes while maintaining containers of pointers to actual data in external memory.

Secondly, interval-valued and box-valued enclosure arithmetics are formally extended to $SRSP$ s. Enclosure arithmetic over $SRSP$ with L leaves can be thought of as rigorously discretized function arithmetic over different adaptive partitions of \mathbb{X} . In the case of density estimation, this allows for fast multi-dimensional adaptive histogram averaging as the data-driven partition is grown/pruned. The histogram average over the posterior distribution corresponds to the Bayes estimator. In the case of plug-in estimation of a function $g : \mathbb{X} \rightarrow \mathbb{R}$ with well-defined interval extension, the interval-valued range enclosure of g over each leaf box of $SRSP$ along with the corresponding count can be used to enclose the plug-in estimate of g using L interval evaluations as opposed to n punctual evaluations of g without resorting to non-rigorous sub-sampling or average-based estimates.

Thirdly, the asymptotic considerations in our mathematical statistical model account for the engineering constraints on massive data processing. The size of the dataset n approaches infinity while available primary machine memory \hat{L} for the currently active chain $\{S(i)\}_{i \in \mathbb{Z}_+}$ is finite. We use Lauritzen's notions of sufficiency [9] and enclosure arithmetic over $SRSP$ s to address *memory-limited consistency and efficiency* of enclosures of classical nonparametric plug-in estimators and density estimators for massive datasets. In effect, our approach transfers the bounds on primary machine memory to the sharpness of enclosures of point estimates with a mathematical formalism not unlike Neumaier's clouds [13].

We propose to implement an efficient and templated way of extending arithmetic over regular sub-pavings using the concept of *ranged sub-pavings*, a natural generalization of the arithmetic for histogram averaging as well as discretized function arithmetic for enclosing plug-in estimates. This will allow us to perform enclosure arithmetic generically over $SRSP$ s and also speed-up interval constraint propagation algorithms [6, 17] in the context of sharper range enclosures of (i) functionals in plug-in estimation as well as (ii) posterior densities in Bayesian model selection for massive datasets.

3 Expected outcomes and results

- Stable GPL release of *MRS: A C++ class library for statistical set processing* that is generic, thread-safe and exactly implements our computational statistical theory with ranged sub-paving and $SRSP$ arithmetics.
- Presenting the theory and methods at an international conference and publishing in an appropriate journal.

4 Budget

Support for the salary of one full-time research associate or for the stipend and tuition fees of one graduate student who can focus on the task of implementing the classes in the MRS library is requested. This would cost about NZ\$ 60,000 or US\$ 43,673 at today's exchange rate. The budget includes my institution's personnel over-head cost of 20%.

5 Google Contact

Andrew Moore (Google Pittsburgh / Carnegie Mellon U.) was interested in the use of set-valued enclosure arithmetics in conjunction with his *Cached-Sufficient Statistics Algorithms* when we chatted over telephone in February 2007.

6 Condensed CV

- Employment & Education:
 - 2007–current Lecturer (akin to Tenured US Assistant Professor), University of Canterbury, NZ
 - 2005–2007 Research Fellow of the Royal Commission for the Exhibition of 1851, Department of Statistics, University of Oxford, UK

- 2005 Postdoctoral Research Associate, Department of Mathematics, Cornell University
- 2005 PhD in Statistics, Cornell University, USA; 1999 BS in Biology & Maths, Minnesota State University, USA
- Relevant Publications:
 - *Machine Interval Experiments: Accounting for the Physical Limits on Empirical and Numerical Resolutions*, Raazesh Sainudiin, LAP Academic Publishers, Köln, Germany, 2009.
 - *Auto-validating von Neumann rejection sampling from small phylogenetic tree spaces*, Raazesh Sainudiin and Thomas York, *Algorithms for Molecular Biology* 4:1, 2009.
 - *Applications of interval methods to phylogenetics*, Raazesh Sainudiin and Ruriko Yoshida, In L. Pachter and B. Sturmfels (Eds.), *Algebraic Statistics for Computational Biology*, Cambridge University Press, 2005.
- Relevant Talks:
 - February 2009 on *An auto-validating trans-dimensional von Neumann rejection sampler* at The 3rd Workshop on High-Dimensional Approximation, Sydney, Australia ([PDF 2MB](#)).
 - December 2009 *Statistical Regular Sub-pavings in Multi-variate Density Estimation*, Department of Mathematics, Ångström Laboratory, Uppsala University, Sweden ([PDF 5.8MB](#))
- Full CV link: <http://www.math.canterbury.ac.nz/~r.sainudiin/RaazeshSainudiinCV.pdf>.

References

- [1] G. Alefeld and J. Herzberger, *An introduction to interval computations*, Academic press, 1983.
- [2] R. R. Bahadur, *Sufficiency and statistical decision functions*, The Annals of Mathematical Statistics **25** (1954), no. 3, 423–462.
- [3] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*, Springer-Verlag New York Inc., 1996.
- [4] R. A. Fisher, *Theory of statistical estimation*, Proceedings of Cambridge Philosophy Society **22** (1925), 700–725.
- [5] P. B. Gibbons and Y. Matias, *Synopsis data structures for massive data sets*, pp. 39–70, American Mathematical Society, Boston, MA, USA, 1999.
- [6] L. Jaulin, M. Kieffer, O. Didrit, and E. Walter, *Applied interval analysis with examples in parameter and state estimation, robust control and robotics*, Springer, April 2001.
- [7] J. R. Kettenring, *Massive datasets*, Wiley Interdisciplinary Reviews: Computational Statistics **1** (2009), 25–32.
- [8] P. Komarek and A. W. Moore, *A dynamic adaptation of AD-trees for efficient machine learning on large data sets*, ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning (San Francisco, CA, USA), Morgan Kaufmann Publishers Inc., 2000, pp. 495–502.
- [9] S. L. Lauritzen, *Notions of sufficiency*, 44th Session of the International Statistical Institute, Madrid, Spain, 1983.
- [10] B. G. Lindsay, J. Kettenring, and O. Siegmund, *A report on the future of statistics*, Statistical Science **19** (2004), no. 3, 387–407 (eng).
- [11] A. W. Moore and M. Soon Lee, *Cached sufficient statistics for efficient machine learning with large datasets*, Journal of Artificial Intelligence Research **8** (1998), 67–91.
- [12] R. E. Moore, *Interval analysis*, Prentice-Hall, 1967.
- [13] A. Neumaier, *Clouds, fuzzy sets and probability intervals*, Reliable Computing, Kluwer Academic Publishers **10** (2004), 249–272.
- [14] D. Pelleg and A. W. Moore, *Dependency trees in sub-linear time and bounded memory*, VLDB Journal **15** (2006), no. 3, 250–262.
- [15] D. Sam-haroud and B. Faltings, *Consistency techniques for continuous constraints*, Constraints **1** (1996), 85–118.
- [16] H. Samet, *The design and analysis of spatial data structures*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1990.
- [17] H. Schichl and A. Neumaier, *Interval analysis on directed acyclic graphs for global optimization*, J. of Global Optimization **33** (2005), no. 4, 541–562.
- [18] E. Segal, D. Pe'er, A. Regev, D. Koller, and N. Friedman, *Learning module networks*, J. Mach. Learn. Res. **6** (2005), 557–588.