Efficient Algorithms for Discrepancy Subset Selection

PhD Project Proposal

March 19, 2021

1 General Information

- Title: Efficient Algorithms for Discrepancy Subset Selection
- Supervisors:
  - Carola Doerr, LIP6, Sorbonne Université, Paris, France
  - Luís Paquete, Departament of Informatics Engineering, University of Coimbra, Portugal
- Location: Sorbonne Université, Paris, France. Extended research visits to Coimbra are strongly encouraged but not strictly required
- Duration: 36 months
- Earliest starting date: October 2021
- Keywords: Geometric discrepancies, subset selection, optimization, operations research

2 Context: Discrepancy Theory

Discrepancy measures are metrics designed to quantify how regularly a set of points is distributed in a given space. Several discrepancy notions exist, measuring different aspects of "regularity". The arguably most common discrepancy notion is the L_∞ star discrepancy. Intuitively speaking, the L_∞ star discrepancy of a point set $P \subseteq [0,1]^d$ measures how well the volume of a d-dimensional anchored box of the form $[0,q)$ can be approximated by the fraction $|P \cap [0,q)|/|P|$ of points that fall inside this box. More precisely, it measures the largest such deviation between volume and fraction of points. Point sets of low star discrepancy have several important applications, among them Quasi-Monte Carlo integration [12, 3], one-shot optimization [2, 1], financial mathematics [7], design of experiments [13], and many more.

The design of point sets that guarantee small discrepancy values has been an intensively studied topic in numerical analysis in the last decades, and several constructions are known to achieve a smaller L_∞ star discrepancy than randomly sampled points. Among the best-known low-discrepancy constructions are those by Hammersley [11], by Sobol [14], and by Halton [10].

Originally motivated by applications in instance selection, we introduced in [6] the *star discrepancy subset selection problem*, which consists of finding a subset of m out of n points that minimizes the star discrepancy. Our preliminary work showed that the so-generated point sets can be of discrepancy values that are much smaller than those of common low-discrepancy sequences, random point sets, and of Latin Hypercube Sampling. This suggests that subset selection could be an interesting approach for generating point sets of small discrepancy value.

To solve the star discrepancy subset selection problem, we introduced two mixed integer linear formulations (MILP) and a combinatorial branch-and-bound (BB) algorithm for this problem and we evaluated our approaches against random subset selection and a greedy construction on different use-cases in dimension two and three. Our results show that one of the MILPs and BB are efficient in dimension two for large and small m/n ratio, respectively, and for not too large n. However, the performance of both approaches decays strongly for larger dimensions and set sizes.

3 Objectives of the Project

The main goal of this PhD project is the **design efficient approaches to address the discrepancy subset selection** for settings with $m > 100$ points and for dimensions $d > 3$. As discussed in our work [6], we do not expect the MILP formulations and BB approaches to generalize well. Improving the quality of the upper and lower bounds for BB, or the use of other techniques such as column generation or branch and cut/price, should allow for better performance only on slightly larger instances. Given the computational complexity of the star discrepancy evaluation, we can expect that it is impossible to find algorithms that scale polynomially in n and d. Heuristic solutions, tailored to the star discrepancy settings such as the "snapping" procedures in [9] may therefore be needed.

Depending on the interest and the background of the candidate, we will study the **application** of the obtained low-discrepancy point sets to different tasks (e.g., numerical integration, one-shot optimization, design of experiments, initialization of Bayesian Optimization, generation of diverse instances, etc.)

We are also interested in **bounding the loss of point sequences vs. specifically designed point sets**. Low discrepancy point sequences can be important when the budget is not known in advance or when intermediate results need to be communicated, but their incremental design comes at the cost of larger discrepancy values.

Theoretical aspects such as the hardness of approximating the discrepancy subset selection problem or the formal analysis of the algorithms and its components are of great interest as well.

4 Supervisors

Carola Doerr, formerly Winzen, is since 2013 a CNRS researcher within the Operations Research team at LIP6, Sorbonne Université in Paris, France. Before joining the CNRS, she was a PostDoc at IRIF, Université de Paris, and at the Max Planck Institute for Informatics in Saarbrücken, Germany, where she obtained her PhD summa cum laude in 2011 under the supervision of Kurt Mehlhorn. Carola's main research activities are in the analysis of randomized algorithms, with a strong focus on evolutionary algorithms and other sampling-based optimization heuristics. Discrepancy theory is a topic that she has worked on since her diploma thesis back in 2007. Mainly focsing on computational aspects of geometric discrepancies, her contributions to the field comprise: proof of NP-hardness of computing the star discrepancy [8], a heuristic to compute the star discrepancy of a given point set [9], a probabilistic lower bound for the star discrepancy of Latin Hypercube Samples [4], and a survey on computational aspects of geometric discrepancy [5]. Carola is Associate Editor of ACM Transactions on Evolutionary Learning and Optimization and editorial board member of Evolutionary Computation.

Luís Paquete is since 2007 an Associate Professor at the Department of Informatics Engineering at the University of Coimbra, Portugal. He received his Ph.D. in Computer Science with summa cum laude from the Technical University of Darmstadt, Germany, in 2005, under the supervision of Thomas Stützle and Wolfgang Bibel. His research interest is mainly focused on exact and heuristic solution methods for multiobjective combinatorial optimization problems. He is in the editorial board of Operations Research Perspectives and Area Editor at ACM Transactions on Evolutionary Learning and Optimization. Luís has supervised one Master thesis on the star discrepancy subset selection problem, which built the basis for the work presented in [6].

5 Student Profile and Prerequisites

- The candidate must have a research **Master's degree** or equivalent in computer science, in mathematics, or similar.
- The student should have a solid background in the **design and the analysis of algorithms**. Courses in optimization, in operations research, or heuristic search are a plus.
- Some **programming experience** (in any procedural programming language) as well as willingness to conduct empirical work is required.
- The PhD thesis targets publications in journals and conferences of highest international standing. A good command of **written and spoken English** is therefore required. English is also our working language. French language skills are not needed.
- **International students** are very welcome to apply.

6 Application Process

- The applications will be handled by EDITE (application system expected to open on April 21, 2021), but we strongly advise all interested candidates to get in touch with with the supervisors of this proposal prior to submitting their application.
- Applications close on May 16, 2021
- Pre-selection: early June, 2021
- Hearing of the selected candidates mid-June, 2021
- Decisions are expected to be announced in June
- Earliest starting date: October 1, 2021

References

[1] O. Bousquet, S. Gelly, K. Kurach, O. Teytaud, and D. Vincent. Critical hyper-parameters: No random, no cry. *CoRR*, abs/1706.03200, 2017.

[2] M. Cauwet, C. Couprie, J. Dehos, P. Luc, J. Rapin, M. Rivière, F. Teytaud, O. Teytaud, and N. Usunier. Fully parallel hyperparameter search: Reshaped space-filling. In *Proc. of the 37th International Conference on Machine Learning, ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 1338–1348. PMLR, 2020.

[3] J. Dick and F. Pillichshammer. *Digital Nets and Sequences.* Cambridge University Press, 2010.

[4] B. Doerr, C. Doerr, and M. Gnewuch. Probabilistic lower bounds for the discrepancy of latin hypercube samples. In *Contemporary Computational Mathematics - A Celebration of the 80th Birthday of Ian Sloan*, pages 339–350. Springer, 2018.

[5] C. Doerr, M. Gnewuch, and M. Wahlström. Calculation of discrepancy measures and applications. In W. Chen, A. Srivastav, and G. Travaglini, editors, *A Panorama of Discrepancy Theory*, pages 621–678. Springer, 2014.

[6] C. Doerr and L. Paquete. Star discrepancy subset selection: Problem formulation and efficient approaches for low dimensions. *CoRR*, abs/2101.07881, 2021.

[7] S. Galanti and A. Jung. Low-discrepancy sequences: Monte carlo simulation of option prices. *Journal of Derivatives*, pages 63–83, 1997.

[8] M. Gnewuch, A. Srivastav, and C. Winzen. Finding optimal volume subintervals with k points and calculating the star discrepancy are NP-hard problems. *Journal of Complexity*, 25:115–127, 2009.

[9] M. Gnewuch, M. Wahlström, and C. Winzen. A new randomized algorithm to approximate the star discrepancy based on threshold accepting. *SIAM J. Numerical Analysis*, 50:781–807, 2012.

[10] J. H. Halton. Algorithm 247: Radical-Inverse Quasi-random Point Sequence. *Communications of the ACM*, 7(12):701 – 702, 1964.

[11] J. Hammersley. Monte carlo methods for solving multivariable problems. *Annals of the New York Academy of Sciences*, 86, 1960.

[12] H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*, volume 63 of *SIAM CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, 1992.

[13] T. Santner, B. Williams, and W. Notz. *The Design and Analysis of Computer Experiments*. Springer Series in Statistics. Springer, 2003.

[14] I. M. Sobol. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 7(4):86–112, Jan. 1967.