Sampling from Discrete Distributions and Computing Fréchet Distances

Abstract of Doctoral Thesis [8]

Karl Bringmann

1 Sampling from Discrete Distributions

Sampling from a probability distribution is a fundamental problem that lies at the heart of randomized computation, and has never been as important as today, as most sciences perform computer simulations of models involving randomness. We approach this problem area from an algorithm theory perspective. The central problem in the first part of this dissertation is *proportional sampling*, defined as follows. We are given non-negative numbers $p_1, \dots, p_n$ that define a probability distribution on $\{1, \dots, n\}$ by picking i with probability proportional to p_i, i.e., the probability of sampling i is $\frac{p_i}{\sum_j p_j}$. The task is to build a data structure that supports sampling from this distribution as a query. The classic solution to this problem is the alias method by Walker from '74 [38], which uses $O(1)$ query time and $O(n)$ preprocessing time, i.e., the time for building the data structure is $O(n)$. It is easy to see that both time bounds of Walker's method are optimal. We extend this classic data structure in various directions as follows.

Succinct Sampling While the time bounds are well understood, space usage of discrete sampling algorithms has received little attention. To bound its space usage, we show that Walker's alias method can be implemented on the Word RAM model of computation (where each cell stores $w = \Omega(\log n)$ bits) with a space usage of $n(w + 2\lg n + O(1))$ bits [12]. Using the terminology of succinct data structures, this solution has a redundancy of $2n\lg n + O(n)$ bits, i.e., it uses $2n\lg n + O(n)$ bits in addition to the information theoretic minimum required for storing the input. We examine whether this space usage can be improved in two common models for data structures from the field of succinct data structures: In the systematic model, in which the input is read-only, we present a novel data structure using $r + O(w)$ redundant bits, $O(n/r)$ expected query time, and $O(n)$ preprocessing time for any r. This is an improvement in redundancy by a factor of $\Omega(\log n)$ over the

alias method for $r = n$, even though the alias method is not systematic. Moreover, we complement this data structure with a lower bound showing that this trade-off is tight for systematic data structures. In the non-systematic model, in which the input numbers may be represented in more clever ways than just storing them one-by-one, we demonstrate a very surprising separation from the systematic case: With only 1 redundant bit, it is possible to support optimal $O(1)$ expected query time and $O(n)$ preprocessing time! On the one hand, these results improve upon the space requirement of the classic solution for a fundamental sampling problem, and on the other hand, they provide the strongest known separation between the systematic and non-systematic model for any data structure problem. Finally, we also believe that these upper bounds are practically efficient and simpler than Walker's alias method.

Restricted Inputs Since the preprocessing and query time bounds of Walker's alias method are optimal in the worst case, we examine the situation where we have additional knowledge about the input distribution [14]. For example, assume that we have the guarantee that the input is sorted. We show that, in this case, the preprocessing time can be reduced to $O(\log n)$ while still achieving expected query time $O(1)$. Moreover, one can further reduce the preprocessing time at the price of increasing the query time, specifically, any expected query time $O(t)$ can be achieved with $O(\log_t n)$ preprocessing time. In particular, we can achieve preprocessing and expected query time $O(\log n / \log\log n)$. We also show tight lower bounds for this trade-off curve at all of its points.

Subset Sampling Let us consider a different sampling problem [14]: In *subset sampling* we are given $p_1, \ldots, p_n$ and consider n independent events, where event i occurs with probability p_i. The task is to sample the set of occurring events. This problem can be seen as a generalization of proportional sampling, since we show that any data structure for subset sampling can be transformed into a data structure for proportional sampling with the same asymptotic running times. As for proportional sampling, we consider sorted and unsorted input sequences and in both cases present data structures with optimal preprocessing-query time trade-offs. The situation is more complex than for proportional sampling, since the running times now also depend on the expected size μ of the sampled subset. For instance, we design a data structure for subset sampling on sorted inputs with preprocessing and expected query time $O(1 + \mu + \frac{\log n}{\log(\log(n)/\mu)})$, which corresponds to one point on an optimal trade-off curve.

Special distributions Particularly fast sampling methods are known for special distributions such as Bernoulli, geometric, or binomial random variables. For

instance, a geometric random variable Geo(p) can be sampled using the simple formula $\lceil \frac{\log R}{\log(1-p)} \rceil$, where R is a uniformly random real in $(0, 1)$. On a Real RAM, this formula can be evaluated in constant time. However, on real-life computers this formula is typically evaluated with the usual floating point precision, so that it is not exact. Hence, we study whether special distributions can be sampled exactly and efficiently on a bounded-precision machine such as the Word RAM. We prove that a geometric random variable Geo(p) can be sampled in expected time $O(1 + \log(1/p)/w)$ on the Word RAM [10]. This is optimal, as it matches the expected number of output words. To this end, we have to avoid the simple formula above, as it is a long-standing open problem to compute logarithms in linear time. We also present optimal sampling algorithms on the Word RAM for Bernoulli and binomial random variables as well as ErdÅŚsâĂŞRÃl'nyi random graphs.

Applications The insights on the above fundamental problems also prove beneficial for sampling more complex random structures motivated by physics. Consider the following simple exemplary process. The Internal Diffusion Limited Aggregation (IDLA [33]) process places particles on the initially empty integer grid $\mathbb{Z}^2$. In every step, a new particle is born at the origin and performs a random walk until it hits an empty grid cell and occupies it. This process models certain chemical and physical phenomena such as corrosion and the melting of a solid around a heat source. The emerging shape is roughly a ball. Proving this rigorously turned out to be a challenging mathematical problem which has only recently been resolved [32]. From a computational perspective, the trivial simulation algorithm takes time $O(n^2)$ to generate an IDLA shape with n particles. We prove that $O(n \log^2 n)$ time and $O(\sqrt{n} \log n)$ space are sufficient for exactly sampling from the IDLA distribution [15], which allows for experiments on a much larger scale.

2 Computing Fréchet Distances

The second part of this dissertation belongs to the area of computational geometry and deals with algorithms for the Fréchet distance, which is a popular measure of similarity of curves. Intuitively, the (continuous) Fréchet distance of two curves P, Q is the minimal length of a leash required to connect a dog to its owner, as they walk along P and Q, respectively, without backtracking.

Alt and Godau introduced the Fréchet distance to computational geometry in 1991 [4, 27]. For polygonal curves P and Q with n and m vertices, respectively, $n \geq m$, they presented an $O(nm \log(nm))$ algorithm. Since Alt and Godau's seminal paper, Fréchet distance has become a rich field of research, with many variants,

extensions, and generalizations (see, e.g., [3, 17, 20, 24]). Being a natural measure for curve similarity, Fréchet distance has found applications in various areas such as signature verification (see, e.g., [34]), map-matching tracking data (see, e.g., [7]), and moving objects analysis (see, e.g., [18]).

A particular variant that we also discuss in this dissertation is the *discrete* Fréchet distance. Here, intuitively the dog and its owner are replaced by two frogs, and in each time step each frog can jump to the next vertex along its curve or stay at its current vertex. Defined in [25], the original algorithm for the discrete Fréchet distance has running time $O(nm)$.

Quadratic time complexity? Recently, improved algorithms have been found for the classic variants. Agarwal et al. [2] showed how to compute the discrete Fréchet distance in (mildly) subquadratic time $O(nm\frac{\log\log n}{\log n})$. Buchin et al. [19] designed algorithms for the continuous Fréchet distance with running time $O(n^2\sqrt{\log n}(\log\log n)^{3/2})$ on the Real RAM and $O(n^2(\log\log n)^2)$ on the Word RAM. However, the problem remained open whether there is a *strongly subquadratic*[1] algorithm for the Fréchet distance, i.e., an algorithm with running time $O(n^{2-\delta})$ for any $\delta > 0$. The only known lower bound shows that the Fréchet distance takes time $\Omega(n\log n)$ (in the algebraic decision tree model) [16]. The typical way of proving (conditional) quadratic lower bounds for geometric problems is via 3SUM [26]. In fact, Helmut Alt conjectured that the Fréchet distance is 3SUM-hard, but this conjecture remains open. Instead of relating the Fréchet distance to 3SUM, we consider the Strong Exponential Time Hypothesis.

Strong Exponential Time Hypothesis (SETH) The hypothesis SETH, introduced by Impagliazzo, Paturi, and Zane [30, 31], provides a way of proving conditional lower bounds. SETH asserts that satisfiability has no algorithms that are much faster than exhaustive search.

Hypothesis SETH: *For no $\varepsilon > 0$,* k-SAT *can be solved in time $O(2^{(1-\varepsilon)N})$ for all $k \geq 3$.*

Note that exhaustive search takes time $O^*(2^N)$ and the best-known algorithms for k-SAT have a running time of the form $O(2^{(1-c/k)N})$ for some constant $c > 0$ [36]. Thus, SETH is a reasonable hypothesis and, due to lack of progress in the last decades, can be considered unlikely to fail. It has been observed before this work that SETH can be used to prove lower bounds for polynomial time problems such

[1] We use the term *strongly subquadratic* to differentiate between this running time and the *(mildly) subquadratic* $O(n^2\log\log n/\log n)$ algorithm from [2].

as k-Dominating Set and others [35], the diameter of sparse graphs [37], and dynamic connectivity problems [1].

Main lower bound Our main result of the second part of this dissertation gives strong evidence that the Fréchet distance may have no strongly subquadratic algorithms by relating it to the Strong Exponential Time Hypothesis. *We prove that there is no $O(n^{2-\delta})$ algorithm for the (continuous or discrete) Fréchet distance for any $\delta > 0$, unless* SETH *fails [9].* Since SETH is a reasonable hypothesis, by this result one can consider it unlikely that the Fréchet distance has strongly subquadratic algorithms. In particular, any strongly subquadratic algorithm for the Fréchet distance would not only give improved algorithms for k-SAT that are much faster than exhaustive search, but also for various other problems such as Hitting Set, Set Splitting, and NAE-SAT via the reductions in [22]. Alternatively, in the spirit of [35], one can view the above theorem as a possible attack on k-SAT, as algorithms for the Fréchet distance now could provide a route to faster k-SAT algorithms. In any case, anyone trying to find strongly subquadratic algorithms for the Fréchet distance should be aware that this is as hard as finding improved k-SAT algorithms, which might be impossible.

We remark that all our lower bounds (unless stated otherwise) hold in the Euclidean plane, and thus also in $\mathbb{R}^d$ for any $d \geq 2$.

Extensions We extend our main lower bound in two important directions: We show approximation hardness and we show tight lower bounds if one curve has much fewer vertices than the other, $m \ll n$. In order to state our result, we first formalize that a statement holds for "$m \approx n^\gamma$ for any γ". We say that a statement holds *for any polynomial restriction of $n^{\gamma_0} \leq m \leq n^{\gamma_1}$* if it holds restricted to instances with $n^{\gamma-\delta} \leq m \leq n^{\gamma+\delta}$ for any constants $\delta > 0$ and $\gamma_0 + \delta \leq \gamma \leq \gamma_1 - \delta$. *We prove that there is no* 1.001*-approximation with running time $O((nm)^{1-\delta})$ for the (continuous or discrete) Fréchet distance for any $\delta > 0$, unless* SETH *fails. This holds for any polynomial restriction of $1 \leq m \leq n$ [9].* In recent work together with Wolfgang Mulzer focussing on the discrete Fréchet distance [13], we improve this result further by excluding 1.399-approximation algorithms even for one-dimensional curves, assuming SETH. Moreover, we present an α-approximation in time $O(n^2/\alpha + n \log n)$ for any $\alpha \geq 1$.

Realistic input curves In attempts to break the apparent quadratic time barrier at least for realistic inputs, various restricted classes of curves have been considered, such as backbone curves [6], κ-bounded and κ-straight curves [5], and ϕ-low density curves [24]. The most popular model of realistic inputs are *c-packed curves*. A curve π is c-packed if for any point $z \in \mathbb{R}^d$ and any radius

$r > 0$ the total length of π inside the ball $B(z, r)$ is at most cr, where $B(z, r)$ is the ball of radius r around z. This model is well motivated from a practical point of view. The model has been used for several generalizations of the Fréchet distance [21, 23, 28, 29]. Driemel et al. [24] introduced c-packed curves and presented a $(1 + \varepsilon)$-approximation for the continuous Fréchet distance in time $O(cn/\varepsilon + cn \log n)$, which works in any $\mathbb{R}^d$, $d \geq 2$.

While this algorithm takes near-linear time for small c and $1/\varepsilon$, is is not clear whether its dependence on c and $1/\varepsilon$ is optimal for c and $1/\varepsilon$ that grow with n. We give strong evidence that the algorithm of Driemel et al. has optimal dependence on c for any constant $0 < \varepsilon \leq 0.001$. *We prove that there is no* 1.001*-approximation with running time* $O((cn)^{1-\delta})$ *for the (continuous or discrete) Fréchet distance on c-packed curves for any* $\delta > 0$*, unless* SETH *fails. This holds for any polynomial restriction of* $1 \leq c \leq n$ *[9].* Since we prove this claim for any polynomial restriction $c \approx n^\gamma$, this result also excludes 1.001-approximations with running time, say, $O(c^2 + n)$.

Regarding the dependence on ε, in any dimension $d \geq 5$ we can prove a conditional lower bound that matches the dependency on ε of the algorithm by Driemel et al. up to a polynomial. *We prove that in* $\mathbb{R}^d$*,* $d \geq 5$*, there is no* $(1 + \varepsilon)$*-approximation for the (continuous or discrete) Fréchet distance on c-packed curves running in time* $O(\min\{cn/\sqrt{\varepsilon}, n^2\}^{1-\delta})$ *for any* $\delta > 0$*, unless* SETH *fails. This holds for sufficiently small* $\varepsilon > 0$ *and any polynomial restriction of* $1 \leq c \leq n$ *and* $\varepsilon \leq 1$ *[9].*

This, however, still leaves open a gap between the best-known upper and (conditional) lower bounds. We resolve this issue positively by giving a faster algorithm with time complexity $O(cn \log^2(1/\varepsilon)/\sqrt{\varepsilon} + cn \log n)$ [11]. This dependence on c, n and ε is optimal in high dimensions apart from lower order factors, unless SETH fails. In fact, the new algorithm was obtained by examining and exploiting properties that prevent a stronger lower bound, thus demonstrating that SETH-based lower bounds may also inspire algorithmic improvements. We leave open the challenging problem of determining the optimal running time in dimensions $d = 2, 3, 4$.

References

[1] A. Abboud and V. Vassilevska Williams. Popular conjectures imply strong lower bounds for dynamic problems. In *Proc. 55th Annual IEEE Symposium on Foundations of Computer Science (FOCS'14)*, 2014. To appear.

[2] P. Agarwal, R. B. Avraham, H. Kaplan, and M. Sharir. Computing the dis-

crete Fréchet distance in subquadratic time. In *Proc. 24th ACM-SIAM Symposium on Discrete Algorithms (SODA'13)*, pages 156–167, 2013.

[3] H. Alt and M. Buchin. Can we compute the similarity between surfaces? *Discrete & Computational Geometry*, 43(1):78–99, 2010.

[4] H. Alt and M. Godau. Computing the Fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 5(1-2):78–99, 1995.

[5] H. Alt, C. Knauer, and C. Wenk. Comparison of distance measures for planar curves. *Algorithmica*, 38(1):45–58, 2004.

[6] B. Aronov, S. Har-Peled, C. Knauer, Y. Wang, and C. Wenk. Fréchet distance for curves, revisited. In *Proc. 14th Annual European Symposium on Algorithms (ESA'06)*, volume 4168 of *LNCS*, pages 52–63. 2006.

[7] S. Brakatsoulas, D. Pfoser, R. Salas, and C. Wenk. On map-matching vehicle tracking data. In *Proc. 31st International Conference on Very Large Data Bases (VLDB'05)*, pages 853–864, 2005.

[8] K. Bringmann. Sampling from discrete distributions and computing Fréchet distances, 2014. Dissertation. `http://scidok.sulb.uni-saarland.de/volltexte/2015/5988/`.

[9] K. Bringmann. Why walking the dog takes time: Frechet distance has no strongly subquadratic algorithms unless SETH fails. In *Proc. 55th Annual IEEE Symposium on Foundations of Computer Science (FOCS'14)*, 2014. To appear.

[10] K. Bringmann and T. Friedrich. Exact and efficient generation of geometric random variates and random graphs. In *Proc. 40th International Colloquium on Automata, Languages, and Programming (ICALP'13)*, volume 7965 of *LNCS*, pages 267–278, 2013.

[11] K. Bringmann and M. Künnemann. Improved approximation for Fréchet distance on c-packed curves matching conditional lower bounds, 2014. Submitted. Preprint at arXiv 1408.1340.

[12] K. Bringmann and K. G. Larsen. Succinct sampling from discrete distributions. In *Proc. 45th Annual ACM Symposium on Symposium on Theory of Computing (STOC'13)*, pages 775–782, New York, NY, USA, 2013. ACM.

[13] K. Bringmann and W. Mulzer. Approximability of the discrete fréchet distance. In *Proc. 31st International Symposium on Computational Geometry (SoCG'15)*, 2015. To appear.

[14] K. Bringmann and K. Panagiotou. Efficient sampling methods for discrete distributions. In *Proc. 39th International Colloquium on Automata, Languages, and Programming (ICALP'12)*, volume 7391 of *LNCS*, pages 133–144, 2012.

[15] K. Bringmann, F. Kuhn, K. Panagiotou, U. Peter, and H. Thomas. Internal DLA: Efficient simulation of a physical growth model. In *Proc. 41th International Colloquium on Automata, Languages, and Programming (ICALP'14)*, volume 8572 of *LNCS*, pages 247–258, 2014.

[16] K. Buchin, M. Buchin, C. Knauer, G. Rote, and C. Wenk. How difficult is it to walk the dog? In *Proc. 23rd European Workshop on Computational Geometry (EWCG'07)*, pages 170–173, 2007.

[17] K. Buchin, M. Buchin, and Y. Wang. Exact algorithms for partial curve matching via the Fréchet distance. In *Proc. 20th ACM-SIAM Symposium on Discrete Algorithms (SODA'09)*, pages 645–654, 2009.

[18] K. Buchin, M. Buchin, J. Gudmundsson, M. Löffler, and J. Luo. Detecting commuting patterns by clustering subtrajectories. *International Journal of Computational Geometry & Applications*, 21(3):253–282, 2011.

[19] K. Buchin, M. Buchin, W. Meulemans, and W. Mulzer. Four soviets walk the dog - with an application to Alt's conjecture. In *Proc. 25th ACM-SIAM Symposium on Discrete Algorithms (SODA'14)*, pages 1399–1413, 2014.

[20] E. W. Chambers, É. Colin de Verdière, J. Erickson, S. Lazard, F. Lazarus, and S. Thite. Homotopic Fréchet distance between curves or, walking your dog in the woods in polynomial time. *Computational Geometry*, 43(3):295–311, 2010.

[21] D. Chen, A. Driemel, L. J. Guibas, A. Nguyen, and C. Wenk. Approximate map matching with respect to the Fréchet distance. In *Proc. 13th Workshop on Algorithm Engineering and Experiments (ALENEX'11)*, pages 75–83, 2011.

[22] M. Cygan, H. Dell, D. Lokshtanov, D. Marx, J. Nederlof, Y. Okamoto, R. Paturi, S. Saurabh, and M. Wahlström. On problems as hard as CNF-SAT. In *Proc. 27th IEEE Conference on Computational Complexity (CCC'12)*, pages 74–84, 2012.

[23] A. Driemel and S. Har-Peled. Jaywalking your dog: computing the Fréchet distance with shortcuts. *SIAM Journal on Computing*, 42(5):1830–1866, 2013.

[24] A. Driemel, S. Har-Peled, and C. Wenk. Approximating the Fréchet distance for realistic curves in near linear time. *Discrete & Computational Geometry*, 48(1):94–127, 2012.

[25] T. Eiter and H. Mannila. Computing discrete Fréchet distance. Technical Report CD-TR 94/64, Christian Doppler Laboratory for Expert Systems, TU Vienna, Austria, 1994.

[26] A. Gajentaan and M. H. Overmars. On a class of $O(n^2)$ problems in computational geometry. *Computational Geometry: Theory and Applications*, 5 (3):165–185, 1995.

[27] M. Godau. A natural metric for curves - computing the distance for polygonal chains and approximation algorithms. In *Proc. 8th Symposium on Theoretical Aspects of Computer Science (STACS'91)*, volume 480 of *LNCS*, pages 127–136, 1991.

[28] J. Gudmundsson and M. Smid. Fréchet queries in geometric trees. In *Proc. 21st Annual European Symposium on Algorithms (ESA'13)*, volume 8125 of *LNCS*, pages 565–576. 2013.

[29] S. Har-Peled and B. Raichel. The Fréchet distance revisited and extended. In *Proc. 27th Annual Symposium on Computational Geometry (SoCG'11)*, pages 448–457, 2011.

[30] R. Impagliazzo and R. Paturi. On the complexity of k-SAT. *Journal of Computer and System Sciences*, 62(2):367–375, 2001.

[31] R. Impagliazzo, R. Paturi, and F. Zane. Which problems have strongly exponential complexity? *Journal of Computer and System Sciences*, 63(4): 512–530, 2001.

[32] D. Jerison, L. Levine, and S. Sheffield. Logarithmic fluctuations for internal DLA. *Journal of the American Mathematical Society*, 25:271–301, 2012.

[33] P. Meakin and J. M. Deutch. The formation of surfaces by diffusion limited annihilation. *The Journal of Chemical Physics*, 85:2320, 1986.

[34] M. E. Munich and P. Perona. Continuous dynamic time warping for translation-invariant curve alignment with applications to signature verification. In

Proc. 7th IEEE International Conference on Computer Vision, volume 1, pages 108–115, 1999.

[35] M. Pătraşcu and R. Williams. On the possibility of faster SAT algorithms. In *Proc. 21nd ACM-SIAM Symposium on Discrete Algorithms (SODA'10)*, pages 1065–1075, 2010.

[36] R. Paturi, P. Pudlák, M. E. Saks, and F. Zane. An improved exponential-time algorithm for k-sat. *Journal of the ACM*, 52(3):337–364, 2005.

[37] L. Roditty and V. Vassilevska Williams. Fast approximation algorithms for the diameter and radius of sparse graphs. In *Proc. 45th Annual ACM Symposium on Symposium on Theory of Computing (STOC'13)*, pages 515–524, 2013.

[38] A. J. Walker. New fast method for generating discrete random numbers with arbitrary distributions. *Electronic Letters*, 10(8):127–128, 1974.