

ROBUST ONLINE SAMPLING FROM POSSIBLY MOVING TARGET DISTRIBUTIONS

FRANÇOIS CLÉMENT AND STEFAN STEINERBERGER

ABSTRACT. We suppose we are given a list of points $x_1, \dots, x_n \in \mathbb{R}$, a target probability measure μ and are asked to add additional points $x_{n+1}, \dots, x_{n+m}$ so that $x_1, \dots, x_{n+m}$ is as close as possible to the distribution of μ ; additionally, we want this to be true uniformly for all m . We propose a simple method that achieves this goal. It selects new points in regions where the existing set is lacking points and avoids regions that are already overly crowded. If we replace μ by another measure μ_2 in the middle of the computation, the method dynamically adjusts and allows us to keep the original sampling points. x_{n+1} can be computed in $\mathcal{O}(n)$ steps and we obtain state-of-the-art results. It appears to be an interesting dynamical system in its own right; we analyze a continuous mean-field version that reflects much of the same behavior.

1. INTRODUCTION AND RESULTS

1.1. The problem. We consider the setting where we have a (known) probability measure μ supported on $\mathbb{R}$ as well as a finite number of points $x_1, \dots, x_n \in \mathbb{R}$. We want to add additional points $x_{n+1}, x_{n+2}, \dots, x_{n+m}$, the goal is to end up with a finite set of points $x_1, \dots, x_{n+m}$ that approximates the measure μ . Since the problem setup is one-dimensional, this problem allows for a relatively easy solution when m , the number of points to be added, is fixed: one can place them easily in a way to minimize virtually any metric. We are interested in the setting when m itself is *not known*. The assumption of m being unknown means we want to have *uniformly* good approximation of the prescribed measure μ for all m .

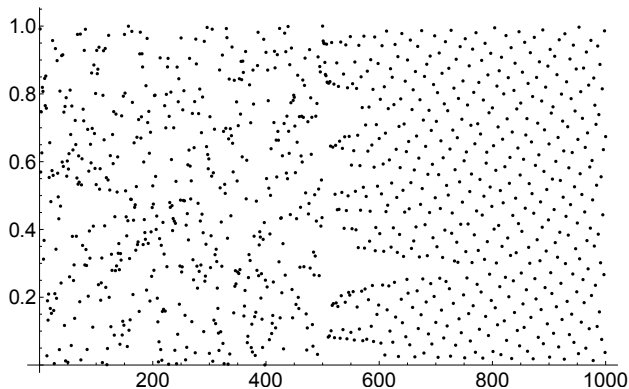


FIGURE 1. 500 uniformly distributed random variables followed by 500 new points: they initially fix gaps before, around point ~ 700 , reaching an equilibrium and refining at smaller scales.

Besides being of intrinsic interest, there are several applications.

- (1) We had an initial sampling budget but based on first results would like to sample more points. This could be the Bayes setting, where we have been sampling with respect to a measure μ_1 and found out that we really should sample with respect to a nearby measure μ_2 .
- (2) We are uncertain about the total sampling budget, or may wish to stop the experiment early based on partial results. Initial samples must still be as uniformly spread out as possible. This is in particular the case in hyperparameter tuning [5, 33] and Bayesian optimization [19].
- (3) We have a set of sampling procedures which individually require samples to be uniformly distributed, and that keep diversity relative to each other. This is for example the case in sequential iterations of CMA-ES [35] or in batch sampling [21].

We note that the problem of *uniformly* distributing points is already difficult if we start with no points at all: finding a sequence of points $(x_k)_{k=1}^\infty$ with the property that $\{x_1, \dots, x_n\}$ is very evenly distributed in the unit interval $[0, 1]$ *uniformly* in n is a nontrivial problem (for which we obtain state-of-the-art results, see §5.2).

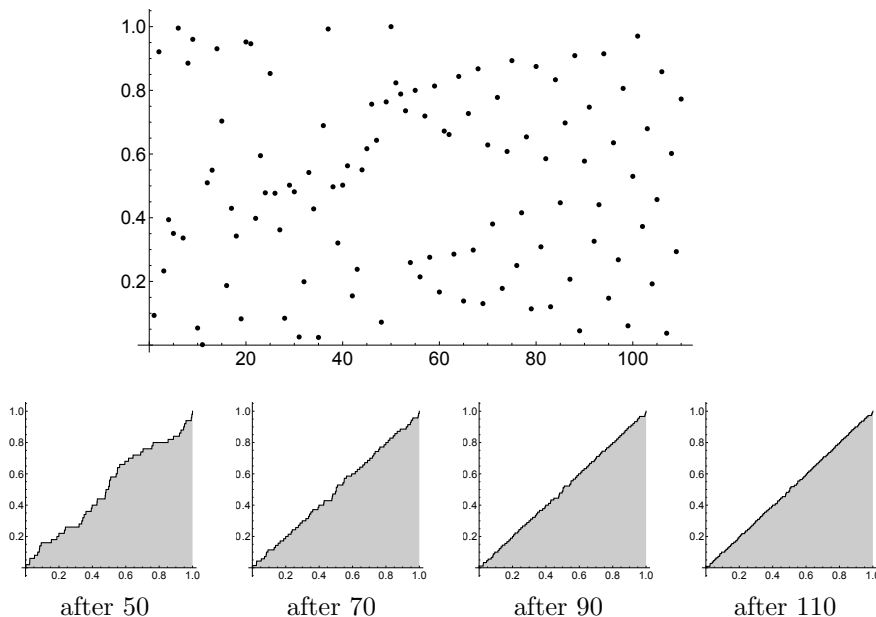


FIGURE 2. 50 iid random points in $[0, 1]$ followed by 60 new points. The new points fill the gaps, the CDF converges quickly to x .

1.2. The method. The main purpose of this paper is to present a simple algorithm that performs exceedingly well (see §5.2). Using the cumulative distribution function of the target measure, it is enough to solve the problem for the uniform distribution on $[0, 1]$. The case of general measures is discussed in Fig. 3 and §1.3.

To solve the problem for the uniform measure on $[0, 1]$, we define a notion of ‘energy’ for any set of n points in the unit interval via

$$E(\mathbf{x}) = \min_{\pi \in S_n} \sum_{i=1}^n \left| x_{\pi(i)} - \frac{i}{n} \right|,$$

where $\pi : \{1, 2, \dots, n\} \rightarrow \{1, 2, \dots, n\}$ ranges over all permutations. The role of the permutation is to sort the points; if the points are already ordered, $x_1 \leq x_2 \leq \dots \leq x_n$, then the energy simplifies to

$$E(\mathbf{x}) = \sum_{i=1}^n \left| x_i - \frac{i}{n} \right|.$$

We propose adding x_{n+1} so that it minimizes the energy $E(x_1, \dots, x_n, x_{n+1})$, i.e.

$$x_{n+1} = \arg \min_{0 \leq x \leq 1} E(x_1, \dots, x_n, x).$$

Greedy methods of this type are easy to implement. Greedy methods tend to work for a while but usually do not perform very well in the long run: errors and mistakes accumulate over time, they have a tendency to get stuck in strange places in the configuration space. Somewhat to our surprise, this does not appear to be the case here; the arising method is simple, fast and completely robust. The robustness stems from an underlying variational structure that we analyze for a continuous limit (see §1.6). Moreover, the minimum is easy to find and easy to store because it is a rational number with a priori known denominator.

Theorem 1. *Given $0 \leq x_1 \leq x_2 \leq \dots \leq x_n \leq 1$, then*

$$x_{n+1} = \arg \min_{0 \leq x \leq 1} E(x_1, \dots, x_n, x)$$

can be computed using $\mathcal{O}(n)$ operations. Moreover, $x_{n+1} = k/(n+1)$ is a rational number with $1 \leq k \leq n+1$. If $1 \leq k \leq n-1$, then $x_k < (k+1)/(n+1) < x_{k+1}$.

The result tells us that the optimization problem can be solved fairly quickly without any sophisticated numerical methods, the result is easy to store and reuse. Finally, the new point is either going to be at the very beginning, $1/(n+1)$, or at the very end, 1, or it is going to end up in a location $x \sim k/(n+1)$ so that $[0, x]$ contains k out of $n+1$ points.

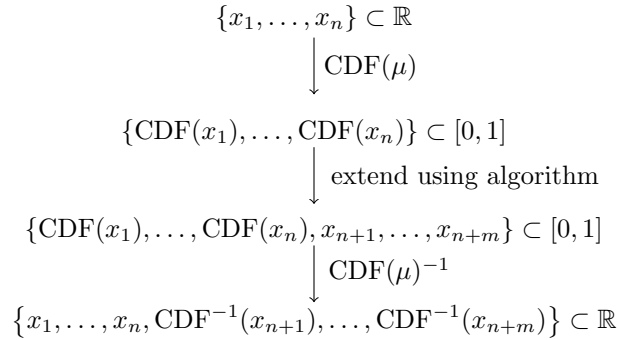


FIGURE 3. Using the method to extend a set of points on $\mathbb{R}$ with respect to an arbitrary probability measure μ .

1.3. General distributions. The solution for the uniform distribution in $[0, 1]$ immediately generalizes to all other probability measures μ . If one is given $x_1, \dots, x_n \in \mathbb{R}$ and if the points are supposed to approximate the probability measure μ , then we may use the cumulative distribution function $\text{CDF}(\mu)$, apply it to $x_1, \dots, x_n$ and then work in the unit interval. The algorithm is outlined in Figure 3, an explicit example is given in Fig. 4.

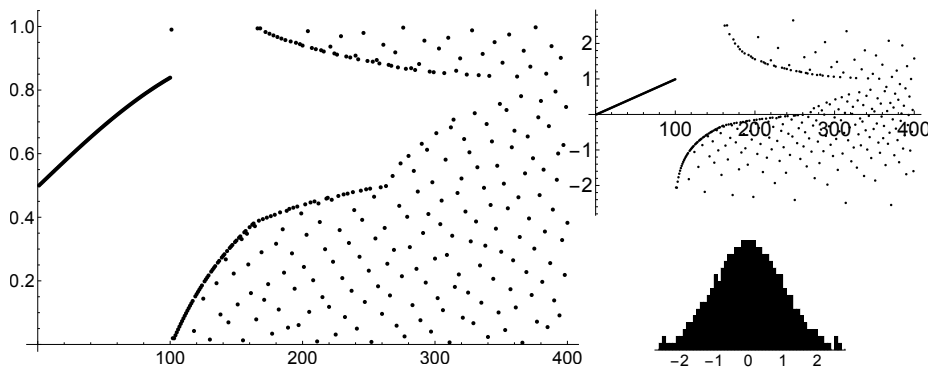


FIGURE 4. (Left): the algorithm working to fix things in CDF-space. (Right, top): the new points being added in real space. (Right, bottom): the final set of 400 points.

Suppose we try to approximate the uniform distribution in $[0, 1]$ with 100 points: this is easy, we use $0/100, 1/100, 2/100, \dots, 99/100$. Imagine we sample in these points and then realize that the goal is to approximate the standard Gaussian $\mathcal{N}(0, 1)$ instead. Is there a way to add new points so that we end up with a Gaussian distribution without throwing away any of the points we already have? Note that there is a severe imbalance: the Gaussian distribution has only

$$\text{a fraction of } \int_0^1 \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \sim 0.341 \dots \text{ of all its points in } [0, 1].$$

This means that unless we have at least $100/0.341 \sim 293$ points, i.e. we add at least 193 points, this severe imbalance cannot be corrected. The situation is a bit more complicated: the Gaussian restricted to $[0, 1]$ is far from uniform, so even more points will be required. The behavior of the algorithm is shown in Figure 4. We use the cumulative distribution function to map the points to a set of points in $[\Phi(0), \Phi(1)] = [0.5, 0.84] \subset [0, 1]$ and run the method in the unit interval. We observe that in ‘CDF-space’, we reach a nice uniform distribution when adding another ~ 250 points. After that, the uniform distribution is then maintained and even further refined. The uniform distribution in $[0, 1]$ in CDF-space is mirrored by a high quality approximation of the Gaussian in real space.

1.4. **Where is x_{n+1} going to end up?** A single step of the algorithm is elementary but x_{n+1} is not easy to predict. The purpose of this section is to explain, roughly, how x_{n+1} behaves and where we can expect to find it. We introduce the discrepancy function

$$\Delta(x) = \frac{\#\{1 \leq i \leq n : x_i \leq x\}}{n} - x.$$

It measures, approximately, whether there are ‘too many’ or ‘not enough’ points in $[0, x]$ compared to what one would expect from the uniform distribution. The second is a one-parameter family of functions $h_x : [0, 1] \rightarrow [0, 1]$ defined by

$$h_x(y) = \begin{cases} y & \text{if } y \leq x \\ y - 1 & \text{if } y > x. \end{cases}$$

A graph of the function $h_{1/3}(y)$ on the interval $[0, 1]$. The function is a line from $(0, 0)$ to $(1/3, 1/3)$, then a horizontal line at $y = 1/3$ from $x = 1/3$ to $x = 1$, and finally a line from $(1, 0)$ to $(1/3, 1/3)$.

A graph of the function $h_{2/3}(y)$ on the interval $[0, 1]$. The function is a line from $(0, 0)$ to $(2/3, 2/3)$, then a horizontal line at $y = 2/3$ from $x = 2/3$ to $x = 1$, and finally a line from $(1, 0)$ to $(2/3, 2/3)$.

With these two definitions in place, we can ‘to first order’ (see §2.3) predict that the next point will be added so as to maximize a certain integral

$$x_{n+1} \sim \arg \max_{0 \leq x \leq 1} \int_0^1 \text{sign}[\Delta(y)] \cdot h_x(y) dy.$$

We quickly illustrate this for 4 examples of 1000 random points in the unit interval.

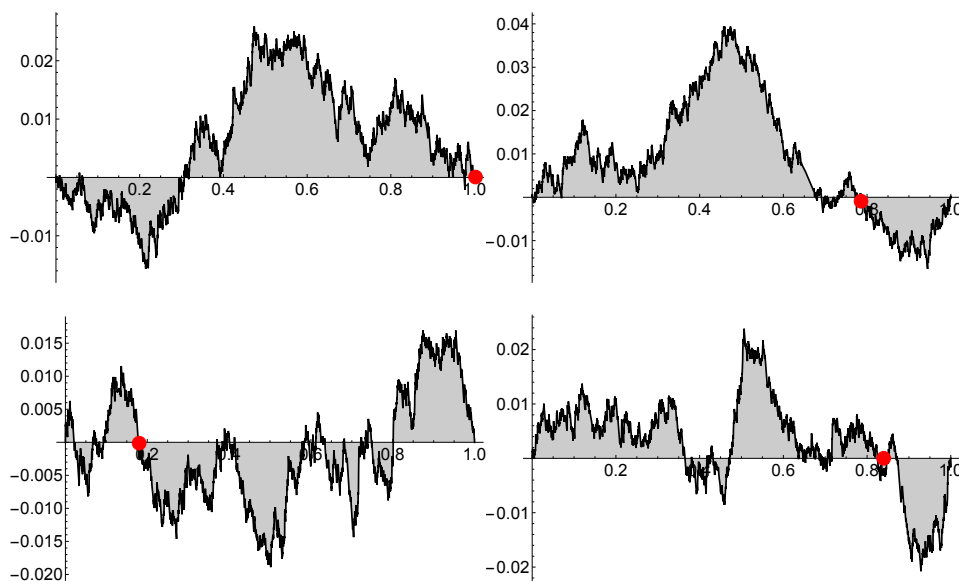


FIGURE 5. Test your intuition: four examples of 1000 random points, their discrepancy function $\Delta(x)$ (black) and x_{1001} (red).

1.5. Behavior of the energy. When placing x_{n+1} , we do not consider the actual numerical value of the energy, we only minimize it. As it turns out, the behavior of the energy $E(x_1, \dots, x_n)$ as a sequence in n is quite interesting. Note that if $x_i = i/n$, then the energy is identically 0. However, in that case it will not be 0 in the next step. On average, the energy is at least size $\gtrsim 1$.

Proposition 1. *For any $x_1, \dots, x_n, x \in [0, 1]$, we have*

$$\frac{E(x_1, \dots, x_n) + E(x_1, \dots, x_n, x)}{2} \geq \frac{1}{8}.$$

Empirically (see Fig. 6) this is the correct scale. To make matters more interesting, we also know that the energy level cannot be uniformly bounded. This follows from a deep result of Halász [24] and some classic arguments (see also Graham [23]).

Theorem 2. *There exists $c > 0$ such that if $(x_k)_{k=1}^\infty$ is any infinite sequence in $[0, 1]$, then there are infinitely many $n \in \mathbb{N}$ such that*

$$\min_{\pi \in S_n} \sum_{i=1}^n \left| x_{\pi(i)} - \frac{i}{n} \right| \geq c\sqrt{\log n}.$$

A nontrivial rate of this type indicates a level of unavoidable complexity and seems to rule out certain naive approaches when trying to understand the behavior of the energies of the greedy sequence. It is known that the rate $\sqrt{\log n}$ is attained by some sequences [45]. We do not know whether the greedy procedure also attains that rate. However, it is easy to do a numerical simulation. Starting with the set $\{1/3, 1/2\}$, we computed the first 500 000 elements and their energies.

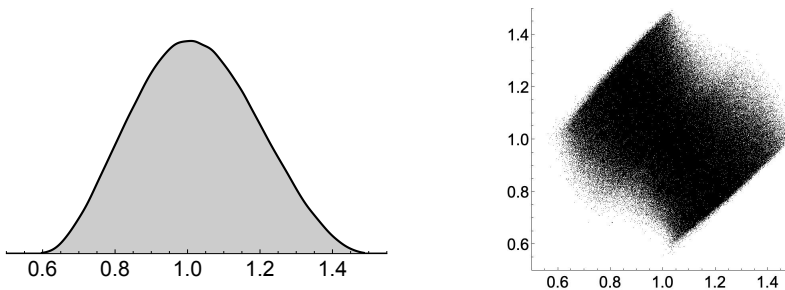


FIGURE 6. Left: distribution of the first 500.000 energies. Right: distribution of pairs of consecutive energies $(E_n, E_{n+1})_{n=1}^{5 \cdot 10^5}$.

The energies are all in the range $[0.5, 1.5]$ and seem to be essentially stationary: $\sqrt{\log n}$ grows very slowly and this setting is an example of that. Plotting the distribution of two consecutive energy values (E_n, E_{n+1}) in the xy -plane shows an intricate structure: small values tend to be followed by fairly large values, while fairly large values tend to be followed by small values (see Fig. 6). To make matters even more intriguing, if we plot three consecutive energies, that is (E_n, E_{n+1}, E_{n+2}) as a point in three dimensions, we observe a clear separation into two distinct clusters. A projection onto a suitable one-dimensional axis shows an approximately bimodal behavior that seems to roughly describe the two ‘modes’ (small, large,

small) and (large, small, large). This is evidence of intricate dynamical behavior. We do not know whether this is what one might call a ‘small n phenomenon’, something that disappears asymptotically, or whether there is a more concrete Theorem lurking in the background.

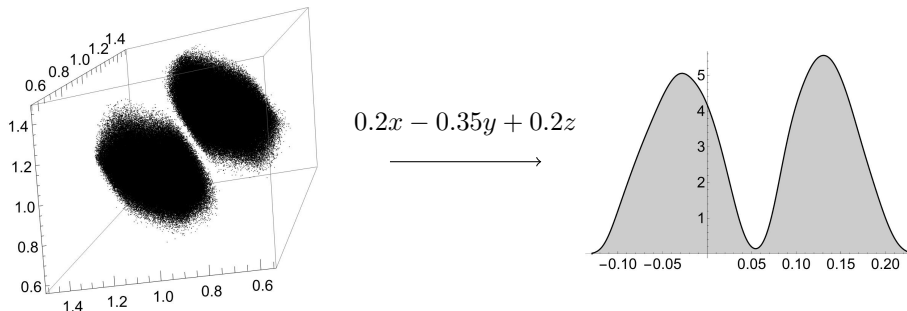


FIGURE 7. Left: the point cloud (E_n, E_{n+1}, E_{n+2}) . Right: a projection onto a one-dimensional line.

Lyapunov functions. We propose a greedy method to add x_{n+1} to an existing set of points $x_1, \dots, x_n$. One way to control such processes is to find a Lyapunov function that can help control the dynamical behavior. An obvious candidate is the energy itself since that is what is being minimized. It would be helpful if there was a statement of the form ‘if $E(x_1, \dots, x_n)$ is very large, then $E(x_1, \dots, x_{n+1})$ is smaller’. Unfortunately, this is false (albeit usually true).

Proposition 2. *There exists a set of n points such that*

$$E(x_1, \dots, x_n) \geq \frac{n}{100}$$

as well as, for all $0 \leq x \leq 1$,

$$E(x_1, \dots, x_n, x) \geq E(x_1, \dots, x_n) + \frac{1}{10}.$$

The example is completely explicit. This should not be misconstrued to mean that the method fails. Indeed, in this particular example, we also have

$$E(x_1, \dots, x_{n+3}) < E(x_1, \dots, x_n) - \frac{1}{10}$$

and the method performs quite beautifully after that, it does seem to be a purely local obstruction. However, it shows that the energy landscape is at least somewhat intricate and non-convex. This limits what one can hope to achieve with arguments that only consider the choice of the next point, x_{n+1} . Lyapunov functions may exist, we have been unable to find any. The continuous setting is much better behaved and more can be said, see §1.6.

1.6. A continuous limit. The difficulties in rigorously understanding the process seem to be rooted in the fact that the energy can be as small as ~ 1 . Adding a single point can lead to global changes and may change the energy also at scale ~ 1 . Adding a single point, even if there are already *many* points present, may not be a ‘small’ change. As is often the case, if a process is extremely adaptive to minuscule changes in the local structure, this makes the process more difficult to analyze. The purpose of this section is to introduce a continuous limit of the process that behaves ‘essentially’ like the discrete process, can be analyzed and may, at least partially, explain why the discrete process works so well.

The setting. Let μ be a probability measure on $[0, 1]$. For ease of exposition, we assume that it is absolutely continuous with respect to the Lebesgue measure and that $\phi(x) = [d\mu/dx]$ has a smooth density that is bounded away from 0 and from above. If $\mu = \phi(x)dx$ and if we denote its cumulative distribution function by

$$\Phi(x) = \int_0^x \phi(y)dy,$$

then the continuous analogue of our energy (see §4 for a derivation) is

$$E(\mu) = \int_0^1 |\Phi^{-1}(x) - x| dx = \int_0^1 |x - \Phi(x)| \phi(x) dx.$$

Clearly, $E(\mu) \geq 0$ with equality if and only if μ is the uniform distribution. One could consider the behavior of the functional under general perturbations of the measure $\mu \rightarrow \mu + \varepsilon\nu$, however, this would not accurately represent the discrete process: in the discrete process we add a single point at a time. To account for that, we revisit the discrete setting where we have an empirical measure

$$\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i} \quad \text{and replace it by} \quad \mu_{n+1} = \frac{1}{n+1} \left(\sum_{i=1}^n \delta_{x_i} + \delta_{x_{n+1}} \right)$$

in the next step. The new measure can be rewritten as

$$\mu_{n+1} = \left(1 - \frac{1}{n+1} \right) \mu_n + \frac{1}{n+1} \delta_{x_{n+1}}.$$

This suggests the appropriate type of perturbation: rescaling the original measure and accounting for the loss of mass in the direction of a Dirac measure. Formally, given the probability measure $\mu = \phi(x)dx$ on $[0, 1]$ and a point $0 \leq x \leq 1$, we define the change of energy under adding an infinitesimal amount of mass in x as

$$\frac{\partial E(\mu)}{\partial x} := \lim_{\varepsilon \rightarrow 0^+} \frac{1}{\varepsilon} (E((1 - \varepsilon)\mu + \varepsilon\delta_x) - E(\mu)).$$

This formal derivative $\partial E(\mu)/\partial x$ can be computed exactly and has some *very* interesting properties. For example, the derivative can only be minimal in fixed points $\Phi(x) = x$ (compare with Theorem 1). These properties make it possible for the continuous functional to have all the structural properties that we were able to establish in the discrete setting (see Theorem 3). Most importantly, the continuous limit does not allow for spurious local minima. Moreover, if one is far from the uniform distribution (in two concrete ways), then there is an upper bound on the gradient and more rapid improvement.

Theorem 3. *Let $\mu = \phi(x)dx$ be a probability measure on $[0, 1]$ with $0 < c < \phi(x) < C < \infty$ such that $x = \Phi(x)$ has only finitely many solutions. Then*

$$\min_{0 \leq x \leq 1} \frac{\partial E(\mu)}{\partial x} < 0.$$

The global minimum of $\partial E(\mu)/\partial x$ is attained in a fixed point of the cumulative distribution function, it satisfies $\Phi(x) = x$. If ϕ is ‘far from the uniform distribution’, then the gradient is ‘very’ negative in a quantitative sense:

- (1) *for every interval $J \subset [0, 1]$ where either $\Phi(x) > x$ or $\Phi(x) < x$,*

$$\min_{0 \leq x \leq 1} \frac{\partial E(\mu)}{\partial x} \leq -\frac{|J|^2}{4}$$

- (2) *and if the equation $\Phi(x) = x$ has only S solutions, then*

$$\min_{0 \leq x \leq 1} \frac{\partial E(\mu)}{\partial x} \leq -\frac{1}{4S}.$$

The result shows an absence of spurious local minima. Moreover, the best way to decrease the energy is to add additional mass into one of the solutions of $\Phi(x) = x$. We note that this equation always has at least two solutions, 0 and 1. Furthermore, if there is a large gap between two consecutive solutions of $\Phi(x) = x$, then the gradient is quite negative and dramatic improvement is possible; the same is true if there are only few solutions of $\Phi(x) = x$.

1.7. Related results. The special case where we start with an empty set of points and try to place a sequence of points in $[0, 1]$ so that the first n points approximate the uniform distribution uniformly in n is an old problem. It has a well-developed theory, we refer to the classic books by Beck and Chen [2], Chazelle [8], Dick and Pillichshammer [14], Drmota and Tichy [16] and Kuipers and Niederreiter [28]. A very rough summary of results would be

- (1) The situation is somewhat understood in $[0, 1]$ with regards to the leading order of magnitude of many ‘error’ quantities [1, 12, 13, 23, 42].
- (2) The problem of finding the ‘best’ example with regards to a natural metric is wide open even in $d = 1$ dimension [30, 36]
- (3) and there are few examples of ‘nice’ sequences in $[0, 1]$ whose construction usually relies on either combinatorial or number-theoretic principles.
- (4) The situation is not understood in $[0, 1]^d$ when $d \geq 2$ [4, 41]
- (5) and, when $d \geq 2$, it is not even clear what to conjecture [3].

Our work grew out of a desire to understand whether constructions based on either greedy or potential-theoretic ‘energy-based’ methods could be effective in constructing new types of such sequences in $[0, 1]$ with the ultimate goal being new such examples in $[0, 1]^d$. Related results in this direction are [6, 7, 9, 10, 11, 27, 37, 43, 44, 46, 47]. We are not aware of this type of question (starting with a given initial set, updating in a way to be close to a given target measure) having been studied anywhere. If we start with an empty set and know in advance how many points to add, this leads us to the question of quantization of probability measures which is very well studied [22]. If the number of points is not known, this leads to questions of uniform distribution and the results cited above. A greedy variant of quantization was studied by Luschgy-Pagès [34] (see also [17]). Some related ideas are presented in [20, 39, 40].

2. PROOF OF THEOREM 1

2.1. Balancing property. We first show that the new point is always a rational number of the form $x_{n+1} = k/(n+1)$ with $1 \leq k \leq n$.

Proof. We may assume that the first n points are ordered $0 \leq x_1 \leq x_2 \leq \dots \leq x_n \leq 1$. Suppose we now add a point so as to minimize the energy for $n+1$ points. If the new point being added satisfies $x \leq x_1$, then the energy is

$$E(x, x_1, \dots, x_n) = \left| x - \frac{1}{n+1} \right| + \sum_{i=1}^n \left| x_i - \frac{i+1}{n+1} \right|.$$

If $1/(n+1) \leq x_1$, then the obviously optimal choice is $x = 1/(n+1)$. We will now argue that in the case $1/(n+1) > x_1$, adding the new point in $[0, x_1]$ cannot be optimal. For this sake, let us define $\ell \in \mathbb{N}_{\geq 0}$ so that $x_1 = \dots = x_{1+\ell} < x_{2+\ell}$ keeps track of how many points are located in x_1 . If $1/(n+1) > x_1$ and we add the new point in $[0, x_1]$, then the energy is minimized by making x as large as the restriction allows and we get

$$\begin{aligned} E_1(x, x_1, \dots, x_n) &= \left| x - \frac{1}{n+1} \right| + \sum_{i=1}^n \left| x_i - \frac{i+1}{n+1} \right| \\ &= \left| x_1 - \frac{1}{n+1} \right| + \sum_{i=1}^{1+\ell} \left| x_i - \frac{i+1}{n+1} \right| + \sum_{i=2+\ell}^n \left| x_i - \frac{i+1}{n+1} \right| \\ &= \left| x_1 - \frac{1}{n+1} \right| + \sum_{i=1}^{1+\ell} \left| x_1 - \frac{i+1}{n+1} \right| + \sum_{i=2+\ell}^n \left| x_i - \frac{i+1}{n+1} \right|. \end{aligned}$$

We will compare this energy to what happens if we add x in the interval $(x_{1+\ell}, x_{2+\ell})$. This would make x the $(2+\ell)$ -th largest point and the energy becomes

$$E_2(x_1, \dots, x, \dots, x_n) = \sum_{i=1}^{1+\ell} \left| x_1 - \frac{i}{n+1} \right| + \left| x - \frac{\ell+2}{n+1} \right| + \sum_{i=2+\ell}^n \left| x_i - \frac{i+1}{n+1} \right|.$$

We first keep $x \in (x_{1+\ell}, x_{2+\ell})$ arbitrary and will then show that $x = x_{1+\ell} + \varepsilon$ for $0 < \varepsilon \ll 1$ sufficiently small is an admissible choice that ensures that $E_2 < E_1$ which is a contradiction. The difference between these two energies is

$$E_1 - E_2 = \left| x_1 - \frac{1}{n+1} \right| - \left| x - \frac{\ell+2}{n+1} \right| + \sum_{i=1}^{1+\ell} \left(\left| x_1 - \frac{i+1}{n+1} \right| - \left| x_1 - \frac{i}{n+1} \right| \right).$$

By assumption, we have $x_1 < 1/(n+1)$, therefore

$$\begin{aligned} \sum_{i=1}^{1+\ell} \left(\left| x_1 - \frac{i+1}{n+1} \right| - \left| x_1 - \frac{i}{n+1} \right| \right) &= \sum_{i=1}^{1+\ell} \left(\frac{i+1}{n+1} - x_1 \right) - \left(\frac{i}{n+1} - x_1 \right) \\ &= \sum_{i=1}^{1+\ell} \frac{i+1}{n+1} - \frac{i}{n+1} = \frac{1+\ell}{n+1}. \end{aligned}$$

Therefore, using the triangle inequality,

$$\begin{aligned} E_1 - E_2 &= \frac{1+\ell}{n+1} + \left| x_1 - \frac{1}{n+1} \right| - \left| x - \frac{\ell+2}{n+1} \right| \\ &\geq \frac{1+\ell}{n+1} + \left| x_1 - \frac{1}{n+1} \right| - \left| x - \frac{1}{n+1} \right| - \frac{1+\ell}{n+1} \\ &= \left| x_1 - \frac{1}{n+1} \right| - \left| x - \frac{1}{n+1} \right| \end{aligned}$$

However, since $x_1 < 1/(n+1)$ and $x > x_1$ is an admissible choice, we see that we can make the expression strictly positive and $E_2 < E_1$. This concludes this case. If the new point being added is larger than all the existing points $x_n \leq x$, then

$$E(x_1, \dots, x_n, x) = \sum_{i=1}^n \left| x_i - \frac{i}{n+1} \right| + |x-1|$$

which is now clearly minimized if we choose $x = 1$ (which is guaranteed to not violate the condition $x_n \leq x$). It remains to deal with the case where the new point being added satisfies $x_k \leq x < x_{k+1}$. The energy is then given by

$$E(x_1, \dots, x, \dots, x_n) = \sum_{i=1}^k \left| x_i - \frac{i}{n+1} \right| + \left| x - \frac{k+1}{n+1} \right| + \sum_{i=k+1}^n \left| x_i - \frac{i+1}{n+1} \right|.$$

If

$$x_k \leq \frac{k+1}{n+1} \leq x_{k+1},$$

then we are done and $x = (k+1)/(n+1)$ is the optimal choice. Suppose now that $(k+1)/(n+1) < x_k$. If that is the case, then the best choice for x subject to the constraints is $x = x_k$. The same computation as above shows that choosing $x = x_k - \varepsilon$ for ε sufficiently small leads to a smaller energy. The case $x_{k+1} \leq (k+1)/(n+1)$ is completely analogous. This shows that x_{n+1} is always a rational number with denominator $n+1$. It also shows that if $x_{n+1} = (k+1)/(n+1)$, then

$$x_k \leq \frac{k+1}{n+1} \leq x_{k+1}$$

which was the desired result. $\square$

2.2. Fast computation. Assume that the first n points are ordered $0 \leq x_1 \leq x_2 \leq \dots \leq x_n \leq 1$. We already know that the next point being added is going to be of the form $k/(n+1)$ for some $1 \leq k \leq n$. Supposing the points are sorted $x_1 \leq x_2 \leq \dots \leq x_n$, it suffices to start the computation from an extremal interval (either $0 \leq x \leq x_1$ or $x_n \leq x \leq 1$), evaluate the sum, then update *only* the sum-term that changes. Indeed, the next point added does not contribute to the new energy, we only need to track the contributions of previous points. If $0 \leq x_{n+1} \leq x_1$, we know that $x_{n+1} = 1/(n+1)$ which means that the energy would then be

$$\left| x_{n+1} - \frac{1}{n+1} \right| + \sum_{i=1}^n \left| x_i - \frac{i+1}{n+1} \right| = \sum_{i=1}^n \left| x_i - \frac{i+1}{n+1} \right|.$$

Optimizing for the best choice $x_1 \leq x_{n+1} \leq x_2$, the change in the energy sum is that the term $|x_1 - 2/(n+1)|$ is replaced by $|x_1 - 1/(n+1)|$. Only this term needs to be updated to obtain the new energy, this can be done in $\mathcal{O}(1)$ time. This process can be repeated for all the intervals: when moving x to the $x_i \leq x \leq x_{i+1}$ interval from

Algorithm 1 Fast computation of the sequence

Require: Starting set L , of initial size l , target number of points n

```

1: Sort  $L$ 
2: for  $i = l$  to  $i = n - 1$  do
3:   Set  $s = \sum_{k=1}^i |L[k] - \frac{k+1}{i+1}|$ , the initial sum if the point is added in  $[0, L[1]]$ .
4:   Define  $c = s$ , current lowest sum value found.
5:   Define  $p = 0$ , position where we found the current lowest sum  $c$ .
6:   for  $j = 1$  to  $j = i$  do
7:     We now consider  $x \in [L[j]; L[j+1]]$ , with  $L[i+1] = 1$ .
8:     Update the sum value  $s = s - |L[j] - \frac{j+1}{i+1}| + |L[j] - \frac{j}{i+1}|$ 
9:     if  $s < c$  then
10:      Update the current best sums and associated position:  $c = s$  and
       $p = j$ .
11:   end if
12: end for
13:   Add the new point  $x = \frac{p+1}{i+1}$  to  $L$ :  $L = L[: p+1] + [x] + L[p+1 :]$ 
14: end for
15: Return  $L$ 

```

the $x_{i-1} \leq x \leq x_i$ interval, we replace $|x_i - (i+1)/(n+1)|$ by $|x_i - i/(n+1)|$. The initial sum requires $\mathcal{O}(n)$ operations, while the n different updates require $\mathcal{O}(1)$ other operations each. Overall, only a linear number of operations are required to compute the next element in the sequence.

2.3. Heuristic derivation of the behavior. We now derive the heuristic

$$x_{n+1} \sim \arg \max_{0 \leq x \leq 1} \int_0^1 \text{sign}[\Delta(y)] \cdot h_x(y) dy.$$

If we add a new point in $0 \leq x \leq 1$, then the change in energy can be written as

$$\Delta E = \sum_{x_i < x} \left(\left| x_i - \frac{i}{n+1} \right| - \left| x_i - \frac{i}{n} \right| \right) + \sum_{x_i > x} \left(\left| x_i - \frac{i+1}{n+1} \right| - \left| x_i - \frac{i}{n} \right| \right).$$

A useful first order asymptotic is

$$\begin{aligned} \left| x_i - \frac{i}{n+1} \right| - \left| x_i - \frac{i}{n} \right| &\sim \frac{1}{n} \text{sign} \left(x_i - \frac{i}{n} \right) \cdot \frac{i}{n} \\ \left| x_i - \frac{i+1}{n+1} \right| - \left| x_i - \frac{i}{n} \right| &\sim \frac{1}{n} \text{sign} \left(x_i - \frac{i}{n} \right) \cdot \frac{i-n}{n}. \end{aligned}$$

This suggests introducing the function $h_x : [0, 1] \rightarrow [0, 1]$ defined by

$$h_x(y) = \begin{cases} y & \text{if } y \leq x \\ y-1 & \text{if } y > x. \end{cases}$$

This allows to write the change in energy, at least approximately, as

$$\Delta E \sim \frac{1}{n} \sum_{x_i < x} \text{sign} \left(x_i - \frac{i}{n} \right) h_x(i/n) + \frac{1}{n} \sum_{x_i > x} \text{sign} \left(x_i - \frac{i}{n} \right) h_x(i/n).$$

Given $0 \leq x_1, \dots, x_n \leq 1$, we introduce the discrepancy function

$$\Delta(x) = \frac{\#\{1 \leq i \leq n : x_i \leq x\}}{n} - x$$

and observe that $\Delta(x) > 0$ indicates that there are ‘too many’ points in $[0, x]$ which shows that, locally, $x_i < i/n$, whereas $\Delta(x) < 0$ indicates that there are ‘not enough’ points in $[0, x]$ which is indicative of $x_i > i/n$. Therefore, we expect

$$\text{sign} \left[\left(x_i - \frac{i}{n} \right) \right] = -\text{sign} [\Delta(x_i)].$$

Thus minimizing ΔE is roughly the same as maximizing

$$-\Delta E \sim \frac{1}{n} \sum_{x_i < x} \text{sign} [\Delta(x_i)] h_x(i/n) + \frac{1}{n} \sum_{x_i > x} \text{sign} [\Delta(x_i)] h_x(i/n)$$

and assuming that the points are approximately uniformly distributed, we may approximate the quantity of interest a little bit further and write

$$x_{n+1} \sim \arg \max_{0 \leq x \leq 1} \int_0^1 \text{sign} [\Delta(y)] \cdot h_x(y) dy.$$

3. PROOFS

3.1. Proof of Proposition 1.

Proof. We know $x_{n+1} = k/(n+1)$ for some $k \in \{1, \dots, n+1\}$ and its contribution to the energy is 0. Therefore

$$\begin{aligned} E(x_1, \dots, x_n) + E(x_1, \dots, x_{n+1}) &= \sum_{\substack{1 \leq i \leq n \\ x_i \leq x_{n+1}}} \left| x_i - \frac{i}{n} \right| + \left| x_i - \frac{i}{n+1} \right| \\ &\quad + \sum_{\substack{1 \leq i \leq n \\ x_i > x_{n+1}}} \left| x_i - \frac{i}{n} \right| + \left| x_i - \frac{i+1}{n+1} \right|. \end{aligned}$$

We have, for all $x \in \mathbb{R}$,

$$\left| x - \frac{i}{n} \right| + \left| x - \frac{i}{n+1} \right| \geq \frac{i}{n(n+1)}$$

as well as

$$\left| x_i - \frac{i}{n} \right| + \left| x_i - \frac{i+1}{n+1} \right| \geq \frac{n-i}{n(n+1)}.$$

Therefore

$$E(x_1, \dots, x_n) + E(x_1, \dots, x_{n+1}) \geq \sum_{i=1}^n \min \left(\frac{i}{n(n+1)}, \frac{n-i}{n(n+1)} \right).$$

A short computation then implies the result. $\square$

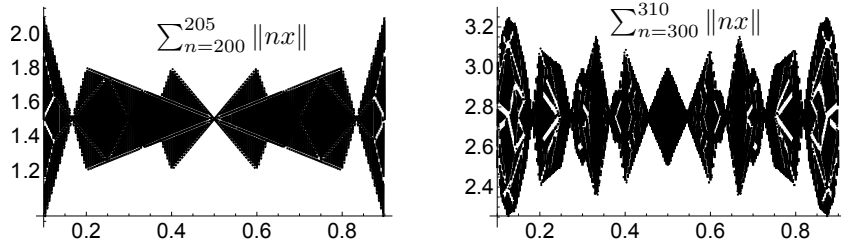
There is an alternative argument that implies slightly larger constants when taking averages over more consecutive elements. Consider an individual point, say $0 < x < 1$. This point can be of the form i/n and it could be of the form $j/(n+1)$ but it is unlikely to be both simultaneously. Using $\|\cdot\|$ to denote the distance to the nearest integer, we have that

$$\min_{k \in \mathbb{N}} \left| x - \frac{k}{n} \right| = \frac{\|nx\|}{n}.$$

Then,

$$\sum_{r=n}^{n+\ell} \min_{k \in \mathbb{N}} \left| x - \frac{k}{r} \right| = \sum_{r=n}^{n+\ell} \frac{\|rx\|}{r} \geq \frac{1}{n+\ell} \sum_{r=n}^{n+\ell} \|rx\|.$$

The sum $\sum_{r=n}^{n+\ell} \|rx\|$ is interesting in itself. It evaluates to 0 when $x = 0, 1$ and, by continuity, for values close to those two values. However, if $\sum_{i=1}^n |x_i - i/n| \lesssim 1$, then ‘most’ elements are far away from 0 and 1. It is not clear to us whether the sum has been studied before (the distribution of the summands has been considered by Don [15]). This can be used to prove that averages over more terms have an average larger than $1/8$, however, we have not pursued this further.



3.2. Proof of Theorem 2.

Proof. The notion of energy is naturally related to Optimal Transport in the Wasserstein W_1 metric. This quantity, in turn, has a simple formula in terms of the cumulative distribution function (see below). Since the original measure is discrete, there is an explicit formula for the W_1 -distance: assuming without loss of generality that $0 \leq x_1 \leq \dots \leq x_n \leq 1$, we have

$$W_1 \left(\frac{1}{n} \sum_{k=1}^n \delta_{x_k}, 1_{[0,1]}(x) dx \right) = \sum_{i=1}^n \int_{(i-1)/n}^{i/n} |x_i - y| dy.$$

Taking the lower bound

$$\int_{(i-1)/n}^{i/n} |x_i - y| dy \geq \frac{1}{n} \left| x_i - \frac{i-0.5}{n} \right| \geq \frac{1}{n} \left| x_i - \frac{i}{n} \right| - \frac{1}{n^2}$$

and summing over all n terms implies

$$W_1 \left(\frac{1}{n} \sum_{k=1}^n \delta_{x_k}, 1_{[0,1]}(x) dx \right) \geq \frac{E(x_1, \dots, x_n)}{n} - \frac{1}{n}.$$

Simultaneously, we have the upper bound

$$\int_{(i-1)/n}^{i/n} |x_i - y| dy \leq \frac{1}{n} \left| x_i - \frac{i}{n} \right| + \frac{1}{n^2}$$

from which we deduce

$$W_1 \left(\frac{1}{n} \sum_{k=1}^n \delta_{x_k}, 1_{[0,1]}(x) dx \right) \leq \frac{E(x_1, \dots, x_n)}{n} + \frac{1}{n}$$

and thus

$$\left| n \cdot W_1 \left(\frac{1}{n} \sum_{k=1}^n \delta_{x_k}, 1_{[0,1]}(x) dx \right) - E(x_1, \dots, x_n) \right| \leq 1.$$

A standard result in optimal transport in one dimension is an exact formula for the W_1 -cost (see [49, Theorem 2.18]): if $F(x) = \int_{-\infty}^x d\mu$ is the CDF of a probability measure μ supported on $[0, 1]$, then

$$W_1(\mu, dx) = \int_0^1 |F(x) - x| dx.$$

Therefore

$$W_1 \left(\frac{1}{n} \sum_{k=1}^n \delta_{x_k}, 1_{[0,1]}(x) dx \right) = \int_0^1 \left| \frac{\#\{1 \leq i \leq n : x_i \leq x\}}{n} - x \right| dx.$$

This quantity is also known as the L^1 -star discrepancy. At this point, we invoke a powerful result of Halász [24]. It says that if $P \subset [0, 1]^2$ is a set of $\#\mathcal{P}$ points $\mathcal{P} = \{(x_1, y_1), \dots, (x_n, y_n)\}$, then

$$\int_0^1 \int_0^1 \left| \frac{\#\{1 \leq i \leq \#\mathcal{P} : x_i \leq x \wedge y_i \leq y\}}{\#\mathcal{P}} - xy \right| dx dy \geq c \sqrt{\log(\#\mathcal{P})}.$$

It is well understood that the problem of understanding sequences in $[0, 1]$ is ‘equivalent’ to understanding the behavior of sets in $[0, 1]^2$ by using a standard construction. The details are as follows. The goal is to show that for every one-dimensional sequence $(x_k)_{k=1}^\infty$ there are infinitely many values $n \in \mathbb{N}$ such that

$$\int_0^1 \left| \frac{\#\{1 \leq i \leq n : x_i \leq x\}}{n} - x \right| dx \geq c \sqrt{\log n}.$$

Suppose this is false. Then, for every $\varepsilon > 0$ we can find a sequence $(x_k)_{k=1}^\infty$ such that the L^1 -discrepancy is $\leq \varepsilon \sqrt{\log n}$ for all but finitely many terms. Given a one-dimensional sequence of points $x_1, \dots, x_n$ in $[0, 1]$, we can turn them into a two-dimensional set of points in $[0, 1]^2$ by writing

$$\mathcal{P} = \left\{ (x_1, \frac{1}{n}), (x_2, \frac{2}{n}), \dots, (x_n, \frac{n}{n}) \right\}.$$

We assume that n is extremely large, much larger than all the exceptional values and so large that the exceptional values contribute an arbitrarily small amount in the subsequent computation. The number of elements of $\mathcal{P}$ that can be found in $[0, x] \times [0, y]$ is really the number of elements of the first $y\#\mathcal{P}$ elements of the sequence that are in $[0, x]$, formally

$$\begin{aligned} \frac{\#\{1 \leq i \leq \#\mathcal{P} : x_i \leq x \wedge y_i \leq y\}}{\#\mathcal{P}} &= \frac{\#\{1 \leq i \leq \lfloor y\#\mathcal{P} \rfloor : x_i \leq x\}}{\#\mathcal{P}} \\ &= \frac{\#\{1 \leq i \leq \lfloor y\#\mathcal{P} \rfloor : x_i \leq x\}}{\lfloor y\#\mathcal{P} \rfloor} \frac{\lfloor y\#\mathcal{P} \rfloor}{\#\mathcal{P}} \\ &= \frac{\#\{1 \leq i \leq \lfloor y\#\mathcal{P} \rfloor : x_i \leq x\}}{\lfloor y\#\mathcal{P} \rfloor} \frac{y\#\mathcal{P} + \mathcal{O}(1)}{\#\mathcal{P}} \\ &= \frac{\#\{1 \leq i \leq \lfloor y\#\mathcal{P} \rfloor : x_i \leq x\}}{\lfloor y\#\mathcal{P} \rfloor} y + \mathcal{O}(\#\mathcal{P}^{-1}). \end{aligned}$$

Then the integral over x satisfies

$$\begin{aligned} I &= \int_0^1 \left| \frac{\#\{1 \leq i \leq \#\mathcal{P} : x_i \leq x \wedge y_i \leq y\}}{\#\mathcal{P}} - xy \right| dx \\ &= \int_0^1 \left| \frac{\#\{1 \leq i \leq \lfloor y\#\mathcal{P} \rfloor : x_i \leq x\}}{\lfloor y\#\mathcal{P} \rfloor} y - xy \right| dx + \mathcal{O}(\#\mathcal{P}^{-1}) \\ &= y \int_0^1 \left| \frac{\#\{1 \leq i \leq \lfloor y\#\mathcal{P} \rfloor : x_i \leq x\}}{\lfloor y\#\mathcal{P} \rfloor} - x \right| dx + \mathcal{O}(\#\mathcal{P}^{-1}). \end{aligned}$$

Now, for our infinite hypothetical sequence, we can integrate this over y and get a bound that is smaller than $\leq \varepsilon \sqrt{\log \#\mathcal{P}}$ which contradicts Halász' result and the argument follows. $\square$

The argument generalizes to other W_p distances, we refer to Graham [23], and it is philosophically connected to a result of Proinov [38] (see also Kirk [26]).

3.3. Proof of Proposition 3.

Proof. Proposition 3 follows from the following explicit example. Let $n = 4m$ be a multiple of 4 and set

$$x_k = \begin{cases} 0 & \text{if } k \leq m \\ k/n & \text{if } m < k \leq 3m \\ 1 & \text{if } k > 3m. \end{cases}$$

Some fairly standard computations show that

$$E(x_1, \dots, x_n) = \frac{n + o(n)}{8}$$

and that the next point x_{n+1} is being added in the middle, $x_{n+1} = (2m)/(4m+1)$ and that the energy increases by, asymptotically, $1/8$. Some more work shows that $x_{n+2} \sim 3/4$ (with energy comparable to $E(x_1, \dots, x_{n+1})$) and that $x_{n+3} \sim 1/4$ with the energy satisfying

$$E(x_1, \dots, x_{n+3}) \sim E(x_1, \dots, x_n) - \frac{1 + o(1)}{12}.$$

$\square$

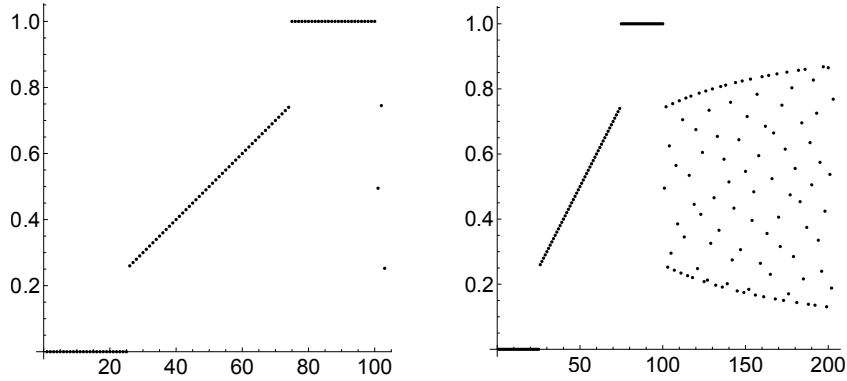


FIGURE 8. Example from Proposition 3 when $m = 25$.

4. THE CONTINUOUS LIMIT

4.1. Deriving the functional. Suppose that μ is a probability measure on $[0, 1]$ that admits a smooth density $\mu = \phi(x)dx$ and that this density is bounded away from 0. We also introduce the cumulative distribution function

$$\Phi(x) = \text{CDF}[\mu](x) = \int_0^x d\mu = \int_0^x \phi(y)dy.$$

If we were to sample $n \gg 1$ elements from that distribution, then the i -th largest element would, as $n \rightarrow \infty$, be located near $x = \Phi^{-1}(i/n)$. The theory of order statistics suggests the approximation

$$\sum_{i=1}^n \left| x_i - \frac{i}{n} \right| \sim \sum_{i=1}^n \left| \Phi^{-1}\left(\frac{i}{n}\right) - \frac{i}{n} \right|.$$

Taking a continuous limit, we arrive at

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \left| \Phi^{-1}\left(\frac{i}{n}\right) - \frac{i}{n} \right| = \int_0^1 |\Phi^{-1}(t) - t| dt.$$

We use that $\Phi : [0, 1] \rightarrow [0, 1]$ is a smooth bijective map allowing us to change variables to obtain the alternative representation

$$\int_0^1 |\Phi^{-1}(t) - t| dt = \int_0^1 |t - \Phi(t)| \cdot \phi(t) dt.$$

4.2. The derivative of the functional. In the discrete setting, we are given a set of n points $x_1, \dots, x_n$ and the corresponding probability measure

$$\mu_1 = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}.$$

The new point x_{n+1} is being added so as to minimize the energy functional and leads to a new measure μ_2 that can be rewritten as

$$\mu_2 = \frac{1}{n+1} \left(\delta_{x_{n+1}} + \sum_{i=1}^n \delta_{x_i} \right) = \frac{n}{n+1} \mu_1 + \frac{\delta_{x_{n+1}}}{n+1}.$$

In the continuous limit, the natural scaling is therefore to replace μ by

$$\mu \rightarrow \mu_\varepsilon = (1 - \varepsilon)\mu + \varepsilon\delta_x$$

where $0 < x < 1$ is chosen so that the energy functional is as small as possible and we should think of $\varepsilon > 0$ as infinitesimally small. In terms of order of quantifiers, we think of the optimal x as a function of $x = x(\varepsilon)$ and let $\varepsilon \rightarrow 0$. We note that, outside of the new point x , the probability density function of μ_ε is given by $(1 - \varepsilon)\phi(x)$. The probability density function of μ_ε is not formally defined in x , however, it is defined in the usual weak sense as Dirac measure $\varepsilon\delta_x$. The change in the cumulative distribution function is also completely explicit and

$$\text{CDF}[(1 - \varepsilon)\mu + \varepsilon\delta_x](y) = (1 - \varepsilon)\Phi(y) + \varepsilon \cdot 1_{y > x}.$$

To analyze the change in the energy functional

$$E(\mu) = \int_0^1 |t - \Phi(t)| \cdot \phi(t) dt,$$

we split the integral into the regions $\{t < x\}$, $\{t = x\}$ and $\{t > x\}$. The first is

$$(I) = \int_0^x |t - (1 - \varepsilon)\Phi(t)| \cdot (1 - \varepsilon)\phi(t)dt.$$

Using that $x - \Phi(x)$ vanishes only a finite number of times and that $\varepsilon \rightarrow 0^+$, we may think of $A = t - \Phi(t)$ as being much bigger in absolute value, $|A| \gg |B|$, than $B = \varepsilon\Phi(t)$ for ‘most’ values of $0 \leq t \leq 1$. Using the identity $|A+B| = |A| + \text{sign}(AB)|B|$, valid for $|A| \geq |B| \geq 0$ the term can be rewritten to leading order as

$$\begin{aligned} (I) &= (1 - \varepsilon) \int_0^x |t - (1 - \varepsilon)\Phi(t)|\phi(t)dt \\ &= (1 - \varepsilon) \int_0^x |t - \Phi(t) + \varepsilon\Phi(t)|\phi(t)dt \\ &= (1 - \varepsilon) \int_0^x |t - \Phi(t)|\phi(t)dt + (1 - \varepsilon) \cdot \varepsilon \int_0^x \text{sign}(t - \Phi(t))\Phi(t)\phi(t)dt, \end{aligned}$$

The second term comes from adding ε of mass in point x . This has very little impact on $|t - \Phi(t)|$ around $t = x$ (there is a jump of order ε) but it does impact $\phi(t)$ which, suitably interpreting the Dirac measures, leads to a contribution of

$$(II) = \varepsilon|x - \Phi(x)| + \mathcal{O}(\varepsilon^2).$$

The third term behaves very much like the first and leads to

$$\begin{aligned} (III) &= \int_x^1 |t - (1 - \varepsilon)\Phi(t) - \varepsilon|(1 - \varepsilon)\phi(t)dt \\ &= (1 - \varepsilon) \int_x^1 |t - \Phi(t) + \varepsilon(\Phi(t) - 1)|\phi(t)dt \\ &= (1 - \varepsilon) \int_x^1 |t - \Phi(t)|\phi(t)dt \\ &\quad + (1 - \varepsilon) \cdot \varepsilon \int_x^1 \text{sign}[(t - \Phi(t))(\Phi(t) - 1)](1 - \Phi(t))\phi(t)dt. \end{aligned}$$

Note that $\Phi(t) - 1$ is always negative and thus

$$\text{sign}[(t - \Phi(t))(\Phi(t) - 1)](1 - \Phi(t)) = \text{sign}[t - \Phi(t)] \cdot (\Phi(t) - 1).$$

Adding all three together, we get

$$\begin{aligned} (I) + (II) + (III) &= (1 - \varepsilon) \int_0^1 |t - \Phi(t)|\phi(t)dt + \varepsilon \int_0^x \text{sign}(t - \Phi(t))\Phi(t)\phi(t)dt \\ &\quad + \varepsilon \int_x^1 \text{sign}(t - \Phi(t))(\Phi(t) - 1)\phi(t)dt + \varepsilon \cdot |x - \Phi(x)| + \mathcal{O}(\varepsilon^2). \end{aligned}$$

Note that the first term and ε should be thought of as externally given: our goal is to choose x to make the remaining terms as small as possible. Thus, formally, letting $\varepsilon \rightarrow 0$, we may define the function $g : [0, 1] \rightarrow \mathbb{R}$ via

$$\begin{aligned} \frac{\partial E(\mu)}{\partial x} &= \int_0^x \text{sign}(\Phi(t) - t)\Phi(t)\phi(t)dt \\ &\quad + \int_x^1 \text{sign}(\Phi(t) - t)(\Phi(t) - 1)\phi(t)dt + |x - \Phi(x)|. \end{aligned}$$

$\partial E(\mu)/\partial x$ describes the minuscule change of the overall energy under adding ε mass to x and rescaling the rest of the measure. The goal is therefore to find the value x that minimizes $\partial E(\mu)/\partial x$. Another change of variables $z = \Phi(t)$, gives $t = \Phi^{-1}(z)$ as well as the transformation

$$\frac{dz}{dt} = \phi(t).$$

A short computation shows that this change of variables leads to

$$\begin{aligned} \frac{\partial E(\mu)}{\partial x} &= \int_0^{\Phi^{-1}(x)} \text{sign}(z - \Phi^{-1}(z))z \, dz \\ &\quad + \int_{\Phi^{-1}(x)}^1 \text{sign}(z - \Phi^{-1}(z))(z - 1) \, dz + |\Phi^{-1}(x) - x|. \end{aligned}$$

For simplicity of expression, we perform another substitution $\Phi(\alpha) = x$ and consider the function

$$\begin{aligned} h(\alpha) &= \int_0^\alpha \text{sign}(z - \Phi^{-1}(z))z \, dz \\ &\quad + \int_\alpha^1 \text{sign}(z - \Phi^{-1}(z))(z - 1) \, dz + |\alpha - \Phi(\alpha)|. \end{aligned}$$

In particular, if h attains a global minimum in α^* , then the optimal point to add mass will be in $x^* = \Phi(\alpha^*)$.

4.3. Simplification I. At this point, we show that the overall algebraic structure implies that minimizers can only be attained in fixed points of $\Phi(x)$.

Lemma 1. *Let $\mu = \phi(x)dx$ be a probability measure on $[0, 1]$ with $0 < c < \phi(x) < C < \infty$ such that $x = \Phi(x)$ has only finitely many solutions. The minimal value of*

$$\begin{aligned} h(\alpha) &= \int_0^\alpha \text{sign}(z - \Phi^{-1}(z))z \, dz \\ &\quad + \int_\alpha^1 \text{sign}(z - \Phi^{-1}(z))(z - 1) \, dz + |\alpha - \Phi(\alpha)|. \end{aligned}$$

is attained in a point satisfying $\alpha = \Phi(\alpha)$.

We note that since $\Phi(0) = 0$ and $\Phi(1) = 1$, such fixed points always exist. In particular, if these are the only two fixed points of the cumulative distribution function, then the minimal value is attained at the boundary.

Proof. Suppose the minimizer is attained in β . We distinguish the two cases $\beta < \Phi^{-1}(\beta)$ and $\beta > \Phi^{-1}(\beta)$ and start with the first case. Suppose the smallest solution of $z = \Phi^{-1}(z)$ that is larger than β is in γ . γ exists because 1 has that property. Then, because β is a global minimizer, we have $h(\beta) \leq h(\gamma)$ and thus

$$\begin{aligned} 0 \geq h(\beta) - h(\gamma) &= - \int_\beta^\gamma \text{sign}(z - \Phi^{-1}(z))z \, dz + \int_\beta^\gamma \text{sign}(z - \Phi^{-1}(z))(z - 1)dz \\ &\quad + |\beta - \Phi^{-1}(\beta)| - |\gamma - \Phi^{-1}(\gamma)| \\ &= \int_\beta^\gamma z \, dz - \int_\beta^\gamma (z - 1)dz + |\beta - \Phi^{-1}(\beta)| - |\gamma - \Phi^{-1}(\gamma)| \\ &= (\gamma - \beta) + |\beta - \Phi^{-1}(\beta)| - |\gamma - \Phi^{-1}(\gamma)| \end{aligned}$$

This implies, recalling that $\gamma = \Phi^{-1}(\gamma)$,

$$0 = |\gamma - \Phi^{-1}(\gamma)| \geq |\beta - \Phi^{-1}(\beta)| + (\gamma - \beta)$$

which is a contradiction. Suppose now that the minimizer happens in β and $\beta > \Phi^{-1}(\beta)$. Suppose the largest solution of $z = \Phi^{-1}(z)$ that is smaller than β happens in α . This α exists because 0 has that property. Then

$$\begin{aligned} 0 \geq h(\beta) - h(\alpha) &= \int_{\alpha}^{\beta} \text{sign}(z - \Phi^{-1}(z))z \, dz - \int_{\alpha}^{\beta} \text{sign}(z - \Phi^{-1}(z))(z - 1)dz \\ &\quad + |\beta - \Phi^{-1}(\beta)| - |\alpha - \Phi^{-1}(\alpha)| \\ &= (\beta - \alpha) + |\beta - \Phi^{-1}(\beta)| - |\alpha - \Phi^{-1}(\alpha)|. \end{aligned}$$

Since $\alpha - \Phi^{-1}(\alpha) = 0$ and $\beta > \alpha$, this is a contradiction. $\square$

This simplifies the problem: the minimizer of the functional has to be a fixed point, a solution of $\alpha = \Phi(\alpha)$. We can thus drop the $|\alpha - \Phi(\alpha)|$ and consider the restricted optimization problem

$$\min_{\substack{0 \leq \alpha \leq 1 \\ \alpha = \Phi(\alpha)}} \int_0^{\alpha} \text{sign}(z - \Phi^{-1}(z))z \, dz + \int_{\alpha}^1 \text{sign}(z - \Phi^{-1}(z))(z - 1)dz.$$

4.4. Simplification II. The result in the previous section shows that it is enough to consider fixed points $\Phi(\alpha) = \alpha$ and study the optimization problem

$$\min_{\substack{0 \leq \alpha \leq 1 \\ \alpha = \Phi(\alpha)}} \int_0^{\alpha} \text{sign}(z - \Phi^{-1}(z))z \, dz + \int_{\alpha}^1 \text{sign}(z - \Phi^{-1}(z))(z - 1)dz.$$

This is a simpler functional but has a more complicated restriction, $\Phi(\alpha) = \alpha$. The purpose of this short section is to show that this new functional has again a remarkable algebraic structure that ensures that any minimizer is automatically a fixed point. This allows us to go from restricted optimization of the simpler functional to unrestricted optimization of the simpler functional.

Lemma 2. *If ϕ, Φ are as above and if*

$$X = \min_{\substack{0 \leq \alpha \leq 1 \\ \alpha = \Phi(\alpha)}} \int_0^{\alpha} \text{sign}(z - \Phi^{-1}(z))z \, dz + \int_{\alpha}^1 \text{sign}(z - \Phi^{-1}(z))(z - 1)dz$$

$$Y = \min_{0 \leq \alpha \leq 1} \int_0^{\alpha} \text{sign}(z - \Phi^{-1}(z))z \, dz + \int_{\alpha}^1 \text{sign}(z - \Phi^{-1}(z))(z - 1)dz,$$

then we have $X = Y$.

Proof. If Y attains its minimum in $\alpha = 0$ or $\alpha = 1$, then there is nothing to prove. If the minimum is attained inside $(0, 1)$ and not in fixed point $\Phi(\alpha) = \alpha$ then, because there are only finitely many fixed points, we are a positive distance from the nearest fixed point. Abbreviating

$$g(\alpha) = \int_0^{\alpha} \text{sign}(z - \Phi^{-1}(z))z \, dz + \int_{\alpha}^1 \text{sign}(z - \Phi^{-1}(z))(z - 1)dz$$

and differentiating leads to

$$g'(\alpha) = \text{sign}(\alpha - \Phi^{-1}(\alpha))\alpha - \text{sign}(\alpha - \Phi^{-1}(\alpha))(\alpha - 1) = \text{sign}(\alpha - \Phi^{-1}(\alpha)).$$

We see that this is ± 1 which ensures that we have not yet found the minimum. $\square$

We see that this is ± 1 unless we are in a fixed point. Therefore, the numerical value of the minimum is the same as the numerical value of the unrestricted problem

$$\min_{0 \leq \alpha \leq 1} \int_0^\alpha \text{sign}(z - \Phi^{-1}(z))z \, dz + \int_\alpha^1 \text{sign}(z - \Phi^{-1}(z))(z - 1)dz$$

and we may work with the unrestricted problem instead. Naturally, when actually trying to solve the problem, the restriction $\alpha = \Phi(\alpha)$ is very useful and should not be omitted. In contrast, when trying to prove that the minimum is quite negative somewhere, it is easier to work without the restriction.

4.5. Simplification III. We will now prove the main structural inequality. A variant of it first appeared in [46] where it had the following form.

Lemma 3 ([46]). *Let $g : [0, 1] \rightarrow \mathbb{R}$ be continuous and $g \not\equiv 0$. Then*

$$\max_{0 \leq z \leq 1} \int_0^z g(x)x \, dx + \int_z^1 g(x)(x - 1) \, dx \geq \frac{1}{16} \frac{1}{\|g\|_{L^2}^2} \left(\max_{0 \leq x \leq 1} \left| \int_0^x g(y)dy \right| \right)^3.$$

Taking g to be the characteristic function on $[0, \varepsilon]$ shows that the inequality is sharp up to constants. The setting in [46] required $g : [0, 1] \rightarrow \mathbb{R}$ to be a function that could attain all possible real values and this allows for examples that are supported on very small intervals. Our setting is much nicer insofar as we can restrict ourselves to studying the same problem for functions $g : [0, 1] \rightarrow \{-1, 1\}$ which excludes these counterexamples and allows for results that have a more concrete interpretation in terms of the geometry of the cumulative distribution function.

Lemma 4. *Let $g : [0, 1] \rightarrow \{-1, 1\}$ with finitely many sign changes. Then*

$$-X = \min_{0 \leq \alpha \leq 1} \int_0^\alpha g(z)z \, dz + \int_\alpha^1 g(z)(z - 1)dz \leq 0.$$

Using $\ell_1, \ell_2, \dots$ to denote the length of the intervals on which g is constant,

(1) *the presence of a long interval ensures that X is large*

$$X \geq \max_{k \in \mathbb{N}} \frac{\ell_k^2}{4}.$$

(2) *there cannot be too many intervals with lengths much larger than X*

$$\sum_{\ell_k > X} (\ell_k - X)^2 \leq 2X$$

(3) *if g has S sign changes, then*

$$X \geq \frac{1}{4S}.$$

Proof. We first prove (1), then give an independent argument for (2) which will imply (3). We abbreviate

$$h(\alpha) = \int_0^\alpha g(z)z \, dz + \int_\alpha^1 g(z)(z - 1)dz.$$

The same computation as above shows that whenever $h'(\alpha)$ exists, it satisfies $h'(\alpha) \in \{-1, +1\}$. We first rewrite the function as

$$h(\alpha) = \int_0^1 g(z)z \, dz - \int_\alpha^1 g(z)dz.$$

Introducing the anti-derivative

$$G(z) = \int_0^z g(y)dy$$

and integrating by parts, we see that

$$\int_0^1 g(z)z \, dz = G(z)z \Big|_0^1 - \int_0^1 G(z)dz = G(1) - \int_0^1 G(z)dz.$$

Combining this with $\int_\alpha^1 g(z)dz = G(1) - G(\alpha)$, we arrive that

$$h(\alpha) = G(\alpha) - \int_0^1 G(x)dx.$$

This representation shows that h has mean value 0 since

$$\int_0^1 h(\alpha)d\alpha = \int_0^1 G(\alpha)d\alpha - \int_0^1 G(x)dx = 0.$$

A non-constant function assumes values both above and below the mean value and h is not constant because $h'(\alpha) \in \{-1, 1\}$ for all but finitely many values. Therefore

$$-X = \min_{0 \leq \alpha \leq 1} \int_0^\alpha g(z)z \, dz + \int_\alpha^1 g(z)(z-1)dz$$

satisfies $X > 0$. It remains to make this step quantitative.

Suppose that $I_k \subset [0, 1]$ is an interval on which g is constant and suppose that interval has length $|I_k| = \ell_k$. Then

$$\max_{a \in I_k} h(a) - \min_{a \in I_k} h(a) = \ell_k.$$

Therefore

$$\max_{a \in I_k} h(a) \geq \ell_k - X.$$

Provided $\ell_k \geq X$, we can deduce that, since h is 1-Lipschitz, one has

$$\int_0^1 \max \{h(x), 0\} \, dx \geq \frac{1}{2}(\ell_k - X)^2.$$

Simultaneously, we have the trivial estimate

$$\int_0^1 \min \{h(x), 0\} \, dx \geq -X$$

and the fact that h has mean value 0 then implies that

$$\frac{1}{2}(\ell_k - X)^2 - X \leq 0 \quad \text{meaning that} \quad (\ell_k - X)^2 \leq 2X.$$

If $X \geq 1/4$, then we are done since $\ell_k \leq 1$. It remains to deal with the case $X \geq 1/4$. Assuming $\ell_k \geq X$, this implies, using $\sqrt{2X} + X \leq 2\sqrt{X}$ for all $0 \leq X \leq 0.35$ that

$$\ell_k \leq \sqrt{2X} + X \leq 2\sqrt{X} \quad \text{and thus} \quad X \geq \frac{\ell_k^2}{4}.$$

Therefore we either have $\ell_k \leq X$, in which case we certainly also have $\ell_k^2/4 \leq X$ because $\ell_k \leq 1$, or we have $\ell_k \geq X$ in which case we have $X \geq \ell_k^2/4$. This means we have $X \geq \ell_k^2/4$ in either case. This proves (1).

As for the second statement, we argue as above and note that if $I_k \subset [0, 1]$ is an interval on which g is constant and if it has length $|I_k| = \ell_k$, then

$$\max_{a \in I_k} h(a) \geq \ell_k - X.$$

If $\ell_k > X$, then

$$\int_{I_k} \max \{h(x), 0\} dx \geq \frac{(\ell_k - X)^2}{2}.$$

Now since,

$$\begin{aligned} 0 &= \int_0^1 h(x) dx = \int_0^1 \min \{h(x), 0\} dx + \int_0^1 \max \{h(x), 0\} dx \\ &\geq -X + \int_0^1 \max \{h(x), 0\} dx, \end{aligned}$$

we may deduce that

$$\sum_{\ell_k > X} \frac{(\ell_k - X)^2}{2} \leq \int_0^1 \max \{h(x), 0\} dx \leq X.$$

Suppose now there are only $S - 1$ sign changes. This leads to S intervals. Maybe only very few, say $F \leq S$ of them, have length larger than X . Then the $S - F$ intervals have length $\leq X$ and we may deduce

$$\sum_{\ell_k > X} \ell_k \geq 1 - (S - F)X.$$

Note that the $(S - F)X$ short interval cannot exceed length 1 and thus this lower bound is nonnegative: $1 - (S - F)X \geq 0$. The Cauchy-Schwarz inequality implies

$$\begin{aligned} 0 &\leq \left(\frac{1}{F} \sum_{\ell_k > X} \ell_k \right) - X = \frac{1}{F} \sum_{\ell_k > X} (\ell_k - X) \\ &\leq \frac{1}{F} \left(\sum_{\ell_k > X} (\ell_k - X)^2 \right)^{1/2} \sqrt{F} \end{aligned}$$

Therefore

$$2X \geq \sum_{\ell_k > X} (\ell_k - X)^2 \geq \frac{1}{F} \left[\sum_{\ell_k > X} (\ell_k - X) \right]^2.$$

We have

$$\sum_{\ell_k > X} (\ell_k - X) \geq 1 - SX.$$

If $X \leq 1/(2S)$, then

$$2X \geq \frac{1}{F} \left[\sum_{\ell_k > X} (\ell_k - X) \right]^2 \geq \frac{1}{4F} \geq \frac{1}{4S}.$$

This implies the desired result. $\square$

5. A RELATED SEQUENCE

5.1. A periodic discrepancy energy and numerics. We point out a related sequence that seems to be roughly equally good at carrying out the desired task of updating an existing set of points in $[0, 1]$ to approximate the uniform distribution as quickly and uniformly as possible. It is remarkable that this sequence seems to arise independently in several different papers in the literature. The method is defined as follows: introduce the 1-periodic function $p : \mathbb{R} \rightarrow \mathbb{R}$ which satisfies

$$p(x) = x^2 - x + \frac{1}{6} \quad \text{for } 0 \leq x \leq 1$$

and is then extended periodically. This is the second Bernoulli polynomial. Given $0 \leq x_1, \dots, x_n \leq 1$, the new point is defined via

$$x_{n+1} = \arg \min_{0 \leq x \leq 1} \sum_{k=1}^n p(x - x_k).$$

This type of construction arises in at least four different ways (which, ultimately, are all equivalent in one form or another)

- as a way to greedily minimize Zinterhof's diaphony [51] which is originally defined as a Fourier multiplier (see also Lev [32])
- in a 2014 paper of Hinrichs and Oettershagen [25] where it arises as the reproducing kernel for functions in the Sobolev space $H^1(\mathbb{T}^2)$. They show that if $\{x_1, \dots, x_n\} \subset [0, 1]$ is given, then the periodic L^2 -discrepancy can be written as

$$L_2^{\text{per}} = -\frac{1}{3} + \frac{1}{n^2} \sum_{i,j=1}^n \left(\frac{1}{3} + p(x_i - x_j) \right).$$

- in a 2020 paper of Brown and Steinerberger [7] where it arises as a linearization of the Wasserstein distance W_2 or, alternatively, from the identity

$$2\pi^2 \sum_{0 \neq h \in \mathbb{Z}} \frac{e^{2\pi i h x}}{h^2} = \{x\}^2 - \{x\} + \frac{1}{6}$$

- and in a completely different way, from a regularity measure proposed by Wagner [50], see below for a discussion.

We were motivated by a Lemma of Wagner [50]: consider the 1-periodic function $h(x) = 1/2 - \{x\}$ and introduce the corresponding function

$$\Delta(x) = \sum_{i=1}^n h(x - x_i).$$

Lemma 5 (Wagner [50]). *The function Δ is piecewise linear with slope $-N$ and jumps of height 1 at each point. It has mean value 0*

$$\int_0^1 \Delta(x) dx = 0.$$

It is connected to the (extreme) discrepancy D_N via

$$\frac{1}{2} D_N \leq \max_{0 \leq x \leq 1} |\Delta(x)| \leq D_N.$$

$\Delta(x)$ is also shown to have a number of other structural properties (that will not be of importance here, we refer to [50]). Motivated by this nice equivalence, one may be tempted to add the next point x_{n+1} so as to minimize the expression

$$\int_0^1 \Delta_{n+1}(x)^2 dx \rightarrow \min.$$

An explicit computation shows that

$$\int_0^1 \left(\sum_{i=1}^{n+1} h(x - x_i) \right)^2 = \int_0^1 \Delta_n(x)^2 + \int_0^1 h(x - x_{n+1})^2 dx + 2 \int_0^1 \Delta_n(x) h(x - x_{n+1})$$

implying that one should set the next point x_{n+1} so that

$$\int_0^1 \Delta_n(x) h(x - x_{n+1}) dx \rightarrow \min.$$

This can be further simplified. One has

$$\begin{aligned} \int_0^1 \Delta_n(x) h(x - x_{n+1}) dx &= \int_0^1 \sum_{k=1}^n \left(\frac{1}{2} - \{x - x_k\} \right) \left(\frac{1}{2} - \{x - x_{n+1}\} \right) dx \\ &= \sum_{k=1}^n \int_0^1 \left(\frac{1}{2} - \{x - x_k\} \right) \left(\frac{1}{2} - \{x - x_{n+1}\} \right) dx. \end{aligned}$$

A short computation shows

$$\int_0^1 \left(\frac{1}{2} - \{x - x_k\} \right) \left(\frac{1}{2} - \{x - x_{n+1}\} \right) dx = p(x_k - x_{n+1})$$

where $p(x)$ is the second Bernoulli polynomial. This is exactly the sequence above.

5.2. Numerical results. We return to the special case of trying to find a sequence of points in $[0, 1]$ such that $\{x_1, \dots, x_N\}$ is as uniformly distributed as possible *uniformly* in N . This is a very difficult problem in its own right. The two metrics most commonly used in the area are the star-discrepancy D_N^*

$$D_N^* = \max_{0 \leq x \leq 1} \left| \frac{\#\{1 \leq i \leq N : x_i \leq x\}}{N} - x \right|$$

and the periodic discrepancy (also known as extreme discrepancy in one dimension)

$$D_N = \max_{\substack{J \subset [0,1] \\ J \text{ interval}}} \left| \frac{\#\{1 \leq i \leq N : x_i \in J\}}{N} - |J| \right|.$$

These two quantities are related and satisfy $D_N^* \leq D_N \leq 2D_N^*$. It is known that the best uniform estimate one can hope for asymptotically is [42]

$$D_N^* \leq c \frac{\log N}{N}$$

where the constant c can be chosen as small as ~ 0.22 [18, 36] and cannot be smaller than ~ 0.065 [31, 30]. As is customary in such situations, the upper bound is probably closer to the truth. This scaling suggests that the appropriate scale-invariant quantities are $ND_N^*/\log N$ and $ND_N/\log N$. Classic examples that are known to attain a very good rate are the van der Corput sequence [48] and the Kronecker sequence with golden ratio. We refer to the Kronecker sequence as the Fibonacci sequence, it is given by $(\{\varphi n\})_{n \in \mathbb{N}}$, where $\{\varphi n\}$ is the fractional part

of golden ratio times n . A greedy contender is the much more recent Kritzinger sequence [27] for which one of the authors [9] has shown numerically that it performs exceedingly well. We start by comparing with the ‘periodic energy’ construction introduced in §5.1 with initial values $\{1/3, 1/2\}$. The results are shown in Figure 9. We see that the sequence from §5.1 performs well and is, within a small constant,

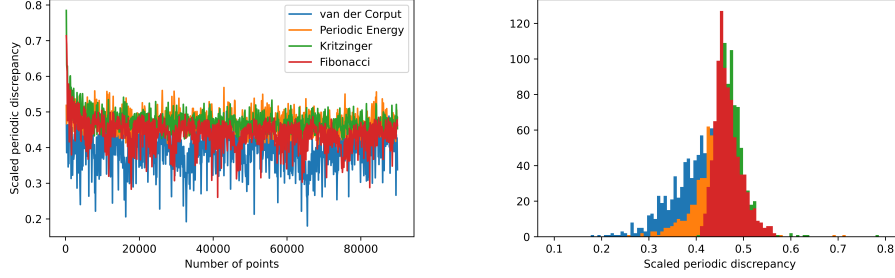


FIGURE 9. Left: $ND_N/\log N$ for three excellent sequences and the new sequences from §5.1. Right: a histogram of the values.

As a second experiment, we consider the main object in this paper. For the purpose of excellent numerics, we replace the energy

$$\sum_{i=1}^n \left| x_i - \frac{i}{n} \right| \quad \text{by the closely related energy} \quad \sum_{i=1}^n \left| x_i - \frac{2i-1}{2n} \right|.$$

From a mathematical perspective, these two objects are very nearly indistinguishable and the first one is a bit more convenient to work with. This is motivated

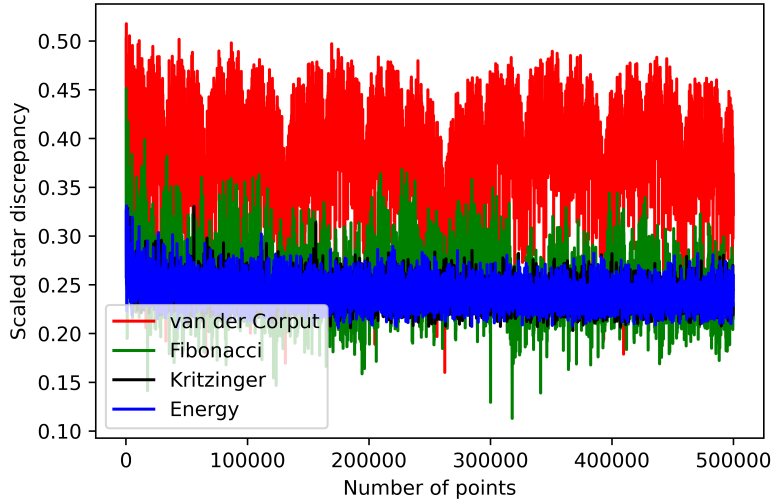


FIGURE 10. The values of $ND_N^*/\log N$ Fibonacci, van der Corput, Kritzinger and Energy sequences. Energy and Kritzinger are extremely similar and, on average, the best performers.

by the fact that the set $\{(2i-1)/(2n) : i \in \{1, \dots, n\}\}$ is known to be optimal for the star discrepancy in one dimension [28]. We use a very close variant of the algorithm described in §2.2 to obtain the sequence. Figure 10 compares the star discrepancy of this sequence with state-of-the-art constructions. It performs as well as the Kritzing sequence, which was shown to empirically outperform the Fibonacci sequence on average [9]. We believe these results to be fascinating for a number of different reasons, including

- (1) the empirical performance of greedy energy minimization is as good (if not better) than the known classical constructions
- (2) our construction and the Kritzing construction have comparable behavior and appear to still be (albeit very slowly decaying); this suggests that there is a common underlying theory or, alternatively,
- (3) that possibly both of these sequences are either asymptotically optimal or close to optimal in some suitable sense.

All of this suggests we are possibly reaching the limit of what can be obtained via greedy constructions, and possibly for one-dimensional sequences in general. Needless to say, a theoretical explanation at that level of regularity is still missing.

REFERENCES

- [1] T. van Aardenne-Ehrenfest, Proof of the Impossibility of a Just Distribution of an Infinite Sequence Over an Interval, *Proc. Kon. Ned. Akad. Wetensch.* 48, 3-8, 1945.
- [2] J. Beck and W. Chen, *Irregularities of Distribution*, Cambridge Tracts in Mathematics (No. 89), Cambridge University Press, 1987.
- [3] D. Bilyk, On Roth's orthogonal function method in discrepancy theory, *Unif. Distrib. Theory* 6, no. 1 (2011), 143-184.
- [4] D. Bilyk, M. Lacey and A. Vagharshakyan, On the small ball inequality in all dimensions, *J. Funct. Anal.* 254 (2008), no. 9, 2470-2502.
- [5] O. Bousquet and S. Gelly and K. Kurach and O. Teytaud and D. Vincent, Critical Hyper-Parameters: No Random, No Cry, *CoRR* abs/1706.03200 (2017).
- [6] L. Brown and S. Steinerberger, On the Wasserstein distance between classical sequences and the Lebesgue measure, *Transactions of the American Mathematical Society* 373, no. 12 (2020): 8943-8962.
- [7] L. Brown and S. Steinerberger, Positive-definite functions, exponential sums and the greedy algorithm: a curious phenomenon. *Journal of Complexity*, 61, 101485.
- [8] B. Chazelle, *The discrepancy method. Randomness and complexity*. Cambridge University Press, Cambridge, 2000.
- [9] F. Clément, Outperforming the Best 1D Low-Discrepancy Constructions with a Greedy Algorithm. *arXiv preprint arXiv:2406.18132*.
- [10] F. Clément, C. Doerr, and L. Paquete, Heuristic approaches to obtain low-discrepancy point sets via subset selection, *Journal of Complexity* 83 (2024): 101852.
- [11] F. Clément, C. Doerr, K. Klamroth, and L. Paquete. "Constructing optimal star discrepancy sets." *Proceedings of the American Mathematical Society, Series B* 12, no. 07 (2025): 78-90.
- [12] J. van der Corput, Verteilungsfunktionen I, *Proc. Akad. Wetensch.* 38 (1935), 813-821.
- [13] J. van der Corput, Verteilungsfunktionen II, *Akad. Wetensch., Proc.* 38 (1935), 1058-1068
- [14] J. Dick and F. Pillichshammer, *Digital nets and sequences. Discrepancy theory and quasi-Monte Carlo integration*. Cambridge University Press, Cambridge, 2010.
- [15] H. Don, On the distribution of the distances of multiples of an irrational number to the nearest integer, *Acta Arithmetica* 139 (2009), page 253-264
- [16] M. Drmota, R. Tichy, *Sequences, discrepancies and applications*. Lecture Notes in Mathematics, 1651. Springer-Verlag, Berlin, 1997.
- [17] R. El Nmeir, H. Luschgy and G. Pagès, New approach to greedy vector quantization. *Bernoulli*, 28 (2022), 424-452.
- [18] H. Faure, Good permutations for extreme discrepancy, *J. Number Theory* 42 (1992), no. 1, 47-56.

- [19] R. Garnett, Bayesian Optimization, Cambridge University Press, 2023.
- [20] A. Gómez, L. Pronzato and M. J. Rendas, Incremental space-filling design based on coverings and spacings: improving upon low discrepancy sequences. *Journal of Statistical Theory and Practice*, 15 (2021), 77.
- [21] J. González and Z. Dai and P. Hennig and N. Lawrence, Batch Bayesian Optimization via Local Penalization, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, PMLR 51:648-657, 2016.
- [22] S. Graf and H. Luschgy, Foundations of quantization for probability distributions. Springer Science & Business Media, 2000.
- [23] C. Graham, Irregularity of distribution in Wasserstein distance. *Journal of Fourier Analysis and Applications*, 26 (2020), p. 1-21
- [24] G. Halász, On Roth's method in the theory of irregularities of point distributions, *Recent progress in analytic number theory*, Vol. 2 (Durham, 1979), Academic Press, London, pp. 79-94.
- [25] A. Hinrichs and J. Oettershagen, Optimal point sets for quasi-Monte Carlo integration of bivariate periodic functions with bounded mixed derivatives. In *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC*, Leuven, Belgium, April 2016 (pp. 385-405). Cham: Springer International Publishing.
- [26] N. Kirk, On Proinov's lower bound for the diaphony. *Unif. Distrib. Theory*, 15 (2020), 39-72.
- [27] R. Kritzing, Uniformly distributed sequences generated by a greedy minimization of the L^2 discrepancy. *Moscow Journal of Combinatorics and Number Theory*, 11 (2022), 215-236.
- [28] L. Kuipers, R. Niederreiter, *Uniform Distribution of Sequences*, Dover 2012
- [29] G. Larcher, On the star discrepancy of sequences in the unit interval. *Journal of Complexity*, 31 (2015), 474-485.
- [30] G. Larcher and F. Puchhammer, An improved bound for the star discrepancy of sequences in the unit interval, *Uniform Distribution Theory* 11 (2016), no. 1.
- [31] G. Larcher and F. Puchhammer, An improved bound for the star discrepancy of sequences in the unit interval, *Unif. Distrib. Theory* 11 (2016), p. 1-14
- [32] V. F. Lev, On two versions of L^2 -discrepancy and geometrical interpretation of diaphony. *Acta Mathematica Hungarica*, 69(4), 281-300.
- [33] L. Li and K. Jamieson and G. DeSalvo and A. Rostamizadeh and A. Talwalkar, Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization, *Journal of Machine Learning Research* 18 (2018) 1-52.
- [34] H. Luschgy and G. Pagš, Greedy vector quantization. *Journal of Approximation Theory* 198 (2015), 111-131.
- [35] J. de Nobel and D. Vermetten and T. Bäck and A. Kononova, Sampling in CMA-ES: Low Numbers of Low Discrepancy Points, *Proc. of IJCCI 2024*, (2024), 120-128
- [36] V. Ostromoukhov. Recent progress in improvement of extreme discrepancy and star discrepancy of one-dimensional sequences. *MCQMC 2008*, pages 561-572.
- [37] F. Pausinger, Greedy energy minimization can count in binary: point charges and the van der Corput sequence, *Annali di Matematica Pura ed Applicata* (1923-) 200, no. 1 (2021): 165-186.
- [38] P. D. Proinov. On irregularities of distribution. *C. R. Acad. Bulgare Sci.* 39 (1986) 31-34.
- [39] L. Pronzato and A. Zhigljavsky, Bayesian quadrature, energy minimization, and space-filling design. *SIAM/ASA Journal on Uncertainty Quantification*, 8 (2020), 959-1011.
- [40] L. Pronzato and A. Zhigljavsky, Quasi-uniform designs with optimal and near-optimal uniformity constant. *Journal of Approximation Theory* 294 (2023), 105931.
- [41] K. F. Roth, On irregularities of distribution. *Mathematika* 1, 73-79 (1954).
- [42] W. Schmidt, Irregularities of distribution. VII. *Acta Arith.* 21 (1972), 45-50.
- [43] S. Steinerberger, A nonlocal functional promoting low-discrepancy point sets, *Journal of Complexity* 54 (2019): 101410.
- [44] S. Steinerberger, Dynamically defined sequences with small discrepancy, *Monatshefte für Mathematik* 191, no. 3 (2020): 639-655.
- [45] S. Steinerberger, Wasserstein distance, Fourier series and applications. *Monatshefte für Mathematik*, 194 (2021), 305-338.
- [46] S. Steinerberger, On combinatorial properties of greedy Wasserstein minimization, *Journal of Mathematical Analysis and Applications* 532, no. 1 (2024): 127940.

- [47] S. Steinerberger, Polynomials with Zeros on the Unit Circle: Regularity of Leja Sequences, *Mathematika* 67 (2021), p. 553-568
- [48] J.G. van der Corput, Verteilungsfunktionen II, *Akad. Wetensch. Amsterdam Proc.*, 38 (1935), 1058-1066
- [49] C. Villani, *Topics in Optimal Transportation* (American Mathematical Society, Providence, 2003)
- [50] G. Wagner, Erdős-Turán inequalities for distance functions on spheres. *Michigan Mathematical Journal*, 39 (1992), 17-34.
- [51] P. Zinterhof, Über einige Abschätzungen bei der Approximation von Funktionen mit Gleichverteilungsmethoden (German). *Osterr. Akad. Wiss. Math.-Naturwiss. Kl. S.-B. II*(185), 121–132 (1976)

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF WASHINGTON, SEATTLE

Email address: `fclement@uw.edu`

Email address: `steinerb@uw.edu`