On a subsequence of random points

Julian Hofstadler

Point Distributions Webinar
August 14, 2020

Outline

The talk is based on the paper *'The power of online thinning in reducing discrepancy'* by R. Dwivedi, O.N. Feldheim, O. Gurel-Gurevich, A. Ramdas.

We will proceed in the following order:

- Discrepancy and old results
- Thinning
- Discrepancy of a thinned sample

Introduction and Notation

- Let $n, d \in \mathbb{N}$ and consider a set of n points in the d-dimensional unit cube, $P_n \subset [0,1]^d$.
- The set of all axis parallel boxes in $[0,1]^d$ is given by $\mathcal{B}$.
- The discrepancy of P_n is defined as

$$D(P_n) = \sup_{B \in \mathcal{B}} |\#(B \cap P_n) - n\lambda(B)|$$

where λ denotes the Lebesgue measure.

Old results

- The Halton-Hammersly point set H_n satisfies

$$D(H_n) \leq C_d \cdot (\log n)^{d-1}.$$

  It is an example for a *low-discrepancy set*.

- K. Roth showed that

$$D(P_n) \geq c_d (\log n)^{\frac{d-1}{2}}.$$

Bounds in the high dimensional setting

Let $\mathcal{B}^\star$ be the set of axis-parallel boxes anchored in 0. The star-discrepancy is given by

$$D^\star(P_n) = \sup_{B \in \mathcal{B}^\star} |\#(B \cap P_n) - n\lambda(B)|.$$

Theorem (HNWW, 2001)

With high probability a set of n iid uniformly distributed points satisfies

$$D^\star(P_n) \leq C \cdot \sqrt{d} \cdot \sqrt{n}.$$

The constant C is absolute and it was shown by Aistleitner that $C \leq 10$ is possible.

Random points

A result due to Doerr shows that $\mathbb{E}(D^\star(P_n)) \geq K \cdot \sqrt{d} \cdot \sqrt{n}$, where $K > 0$ is an absolute constant.

So random points are good w.r.t. d, but still there are some questions:

- Can we achieve bounds where the order of n is $\log n$?
- Is it possible to reduce the probability of clustering?
- What if we are allowed to reject some points?

Thinning

Thinning may help to avoid clustering. We need:

- A sequence of iid random points $(X_k)_{k\in\mathbb{N}}$, uniformly distributed in $[0,1]^d$.
- A sequence iid random variables $(U_k)_{k\in\mathbb{N}}$, uniformly distributed in $[0,1]$.
- A sequence of measurable functions $(f_k)_{k\in\mathbb{N}}$, $f_k\colon ([0,1]^d)^k \times [0,1]^d \to [0,1]$.

The above collection is called a *thinning strategy*.

Thinning II

If the first k points $Z_1, \ldots, Z_k$ are already chosen and r is the number of rejections that were made so far we

- keep X_{k+r+1} if $f_k(Z_1, \ldots, Z_k, X_{k+r+1}) \geq U_k$,
- reject X_{k+r+1} if $f_k(Z_1, \ldots, Z_k, X_{k+r+1}) < U_k$.

In the case that X_{k+r+1} is rejected we have to keep X_{k+r+2}.

If all thinning functions f_k satisfy $f_k \geq 1 - \beta$ for some $0 < \beta < 1$, we say that the points $Z_1, \ldots, Z_n$ are a $(1 + \beta)$-thinned sample.

$(1+\beta)$-thinning

Lemma

Let μ be an absolutely continuous probability measure on $[0,1]^d$ whose density satisfies

$$1 - \frac{\beta}{2} \leq \rho(x) \leq 1 + \frac{\beta}{2}.$$

Then $\rho(x) - \frac{\beta}{2}$ defines a thinning function whose $(1+\beta)$ thinned samples are distributed w.r.t. μ.

Idea: Define a thinning strategy via densities.

Conditional densities - Notation

Assume we already know $\{Z_1, Z_2, \dots, Z_t\} = P_t$.

- The *counting measure* is denoted by $\nu_t(A) = \#(A \cap P_t)$.
- The density for Z_{t+1} is given by ρ_t and may depend on P_t.
- The probability of $Z_{t+1} \in A$ is then given by

$$\mathbb{P}(Z_{t+1} \in A \,|\, P_t) = \mu_t(A) = \int_A \rho_t(x) dx.$$

- Additionally we would like to have $1 - \frac{\beta}{2} \leq \rho_t \leq 1 + \frac{\beta}{2}$ for all t, to have the thinning property.

Dyadic Intervals and Haar functions

Let $\ell \in \mathbb{N}_0$ and $k \in \{0, 1, \dots, 2^\ell - 1\}$.

- We call $I = [k2^{-\ell}, (k+1)2^{-\ell})$ *dyadic interval* with length $2^{-\ell}$.
- Splitting I into its left and right half I_{left}, I_{right} we define

$$h(x) = \begin{cases} 1 & \text{if } x \in I_{left} \\ -1 & \text{if } x \in I_{right} \\ 0 & \text{else} \end{cases}$$

  as Haar function w.r.t. I.
- The order of h is $\mathcal{O}(h) = -\log_2 |I_{left}| = -\log_2 |I_{right}| = \ell + 1$. The constant function $\chi_{[0,1]}$ has order 0.

Haar functions - multidimensional case

Let $I_1, I_2, \dots, I_d$ be dyadic intervals of length $2^{-\ell_1}, 2^{-\ell_2}, \dots, 2^{-\ell_d}$ with associated Haar functions $h_1, h_2, \dots, h_d$.

- The rectangle $R = I_1 \times I_2 \times \cdots \times I_d$ is called a dyadic rectangle of size $|R| = 2^{-\ell}$, where $\ell = \sum_{i=1}^{d} \ell_i$.
- The Haar function w.r.t. R is given by

$$H(x) = h_1(x_1) \cdot h_2(x_2) \cdots h_d(x_d),$$

  and its order is given by $\mathcal{O}(H) = \sum_{i=1}^{d} \mathcal{O}(h_i)$. Again, the constant function is the only function of order 0.
- We define the set $\mathcal{H}_m^\ell = \{H\colon\, m \le \mathcal{O}(H) \le \ell\}$ as set of Haar functions which have order between m and ℓ.

Connecting Haar functions and conditional densities

For any Haar function H we define $H^+ = \{x \in [0,1]^d \colon H(x) = 1\}$ and $H^- = \{x \in [0,1]^d \colon H(x) = -1\}$. (Note that $|H^+| = |H^-|$, if H has at least order 1.)

So we can use

$$\nu_t(H) := \nu_t(H^+) - \nu_t(H^-)$$

to see if there are more points in H^+ or H^-.

We want to use ν_t to see how well distributed the already given points are w.r.t. $\operatorname{supp} H = H^+ \cup H^-$.

The process

We define the densities $(\rho_t)_{t\in\mathbb{N}_0}$ as

$$\rho_t(x) = 1 - \frac{\beta}{2W(\ell)} \sum_{H\in\mathcal{H}_1^\ell} \mathrm{sgn}(\nu_t(H))H(x),$$

where $\ell = \ell(t) = \lfloor \log_2 t \rfloor$, $\beta \in (0,1)$ and $W(\ell) = \sum_{i=1}^{\ell} \binom{d-1+i}{d-1}$.
We have

$$1 - \frac{\beta}{2} \leq \rho_t(x) \leq 1 + \frac{\beta}{2},$$

i.e. the requirements of the 'thinning lemma' are fulfilled.
We will denote this process by $Z = (Z_t)_{t\in\mathbb{N}}$ form here on.

The process - a closer look

Inside the sum we see

$$\operatorname{sgn}(\nu_t(H))H(x) = \begin{cases} -1 & \text{if } x \in H^+ \text{ and } \nu_t(H^+) < \nu_t(H^-) \\ 1 & \text{if } x \in H^- \text{ and } \nu_t(H^+) < \nu_t(H^-) \\ -1 & \text{if } x \in H^- \text{ and } \nu_t(H^-) < \nu_t(H^+) \\ 1 & \text{if } x \in H^+ \text{ and } \nu_t(H^-) < \nu_t(H^+) \\ 0 & \text{else.} \end{cases}$$

Set $\theta = \theta_\ell = \frac{\beta}{W(\ell)}$.

Lemma

The process defined by the conditional densities $(\rho_t)_{t\in\mathbb{N}}$ is θ-balancing for $\mathcal{H}_1^\ell$. In particular this means that for any $H \in \mathcal{H}_1^\ell$ and any t there holds

1) $\kappa \leq \mu_t(H^+ \cup H^-) \leq 3\kappa$
2) $\mu_t(H^+) \geq \mu_t(H^-) + \theta_\ell \kappa \qquad$ *if* $sgn(\nu_t(H)) = -1$
3) $\mu_t(H^-) \geq \mu_t(H^+) + \theta_\ell \kappa \qquad$ *if* $sgn(\nu_t(H)) = 1$,

where $\kappa = |H^+| = |H^-|$.

Lemma

Let Z be a process which is θ-balancing for $\mathcal{H}_1^\ell$. If for some Haar function $H \in \mathcal{H}_1^\ell$ and for some $0 \leq s$ we have that

$$\mathbb{E}\left(\exp\left(\theta \frac{|\nu_s(H^+) - \nu_s(H^-)|}{2}\right)\right) < \frac{150}{\theta^2},$$

then for all $s \leq t$ it holds

$$\mathbb{E}\left(\exp\left(\theta \frac{|\nu_t(H^+) - \nu_t(H^-)|}{2}\right)\right) < \frac{150}{\theta^2}.$$

A first concentration bound for our process

Lemma

For the process Z and any Haar function H we have

$$\mathbb{E}\left(\exp\left(\frac{\beta|\nu_n(H)|}{2W(\ell)}\right)\right) < \frac{C\,W(\ell)^2}{\beta^2},$$

where $C > 0$ is an absolute constant.

Sketch of proof: Use the balancing property for Haar functions with 'large' support. For those with 'small' support we use 'domination' by a binomially distributed random variable.

Lattice rectangles

We say that R is a $2^{-\ell}$-lattice rectangle if R is an axis-parallel rectangle with corners on $2^{-\ell}\mathbb{Z}^d \cap [0,1]^d$, i.e. R has the form

$$R = [a_1 2^{-\ell}, b_1 2^{-\ell}) \times [a_2 2^{-\ell}, b_2 2^{-\ell}) \times \cdots \times [a_d 2^{-\ell}, b_d 2^{-\ell}),$$

where $a_i, b_i \in \{0, 1, \ldots, 2^{\ell}\}$ for $1 \leq i \leq d$.

Lemma

If R is a $2^{-\ell}$-lattice rectangle, then there exist disjoint dyadic rectangles $D_1, \ldots, D_k$, where $k \leq (2\ell)^d$, of size at least $2^{-\ell d}$ such that

$$\bigcup_{i=1}^{k} D_i = R.$$

A concentration bound for lattice rectangles

If we combine everything we saw so far we can proof a more general concentration bound.

Lemma

For any $2^{-\ell}$-lattice rectangle R we have

$$\mathbb{E}\left(\exp\left(\frac{\beta|D(R,P_n)|}{2W(\ell)(2\ell)^d}\right)\right) < \frac{C\,W(\ell)^2}{\beta^2},$$

where $C > 0$ is an absolute constant.

Sketch of proof

First we estimate the discrepancy in terms of Haar functions, i.e. for some 'suitable' coefficients $a_j(H)$ we have

$$|D(R, P_n)| \leq \sum_{j=1}^{k} \sum_{H \in \mathcal{H}_1^{d\ell}} |a_j(H)| \cdot |\nu_n(H)|,$$

such that $\sum_{j=1}^{k} \sum_{H \in \mathcal{H}_1^{d\ell}} |a_j(H)| \leq 1$.

Due to Jensen's inequality we have

$$e^{\sum_{j=1}^{k} \sum_{H \in \mathcal{H}_1^{d\ell}} \frac{|a_j(H)|\beta}{2W(\ell)} \cdot |\nu_n(H)|} \leq \sum_{j=1}^{k} \sum_{H \in \mathcal{H}_1^{d\ell}} |a_j(H)| \cdot e^{\frac{\beta}{2W(\ell)} |\nu_n(H)|},$$

which allows to use the previous concentration bound.

Main results

Theorem

The process Z generates a sequence such that for any $n \in \mathbb{N}$ and any $q > 0$ it holds

$$\mathbb{P}\left(D(P_n) \geq \frac{(\log_2 n)^{2d}}{\beta} \left(q + 1000 + 100 d^2 \log_2 n\right) \right) \leq \beta^{-2} e^{\frac{-q}{50}}.$$

Corollary

For the process Z we almost surely have

$$\limsup_{n \to \infty} \frac{D(P_n)}{(\log_2(n))^{2d+1}} \leq \frac{100(d^2+1)}{\beta}.$$

Sketch of Proof: Use Markov's inequality, approximation of axis-parallel rectangles by lattice rectangles and the concentration bound.

Open questions

Due to numerical simulations Dwivedi et al. conjecture that the process Z satisfies

$$\limsup_{n\to\infty} \frac{D(P_n)}{(\log_2(n))^{\frac{3d}{2}+1}} < \infty.$$

In the same paper they also describe another technique, called the *greedy Haar strategy*, for which they conjecture that this produces a sequence Z such that

$$\limsup_{n\to\infty} \frac{D(P_n)}{(\log_2(n))^{d+1}} < \infty.$$

However, for this strategy there are only results from simulations.

More open questions

There is even more work to do

- At some points the authors think one could do better.
- Are there other thinning strategies which perform better?
- Is it possible to achieve a bound of the form $C\sqrt{n \cdot d}$?
- Thinned samples w.r.t. other discrepancies, e.g. on the sphere.
- Can we use thinned samples in approximation methods which work with random points?

Thank you for your attention!

Literature I

Main paper:

R. Dwivedi, O. N. Feldheim, O. Gurel-Gurevich, and A. Ramdas. The power of online thinning in reducing discrepancy. *Probability Theory and Related Fields*, 174(1-2):103-131, 2019.

More on thinning and random choices:

Y. Azar, A. Z. Broder, A. R. Karlin, and E. Upfal. Balanced allocations. *SIAM journal on computing*, 29(1):180-200, 1999.

Y. Peres, K. Talwar, and U. Wieder. Graphical balanced allocations and the $(1+\beta)$-choice process. *Random Structures & Algorithms*, 47(4):760-775, 2015.

Discrepancy theory:

C. Aistleitner. Covering numbers, dyadic chaining and discrepancy. *Journal of Complexity*, 27(6):531-540, 2011.

Literature II

D. Bilyk. On roth's orthogonal function method in discrepancy theory. *Unif. Distrib. Theory*, 6(1):143-184, 2011.

J. Dick and F. Pillichshammer. Digital nets and sequences: discrepancy theory and quasi-Monte Carlo integration. *Cambridge University Press*, 2010.

B. Doerr. A lower bound for the discrepancy of a random point set. Journal of Complexity, 30(1):16-20, 2014.

S. Heinrich, E. Novak, G. Wasilkowski, and H. Wozniakowski. The inverse of the star-discrepancy depends linearly on the dimension. *Acta Arithmetica*, 96:279-302, 01 2001.

J. Matousek. Geometric discrepancy: An illustrated guide, volume 18. *Springer Science & Business Media*, 2009.