Dispersion of a Point Set

Enhanced Bounds and Practical Applications

Matěj Trödler

FNSPE CTU in Prague

June 21, 2024

Contents

Dispersion of a Point Set

- Let $f\colon [0,1]^d \to \mathbb{R}$ be a real continuous function
- Sequence of points in the cube $\left(x_n\right)_{n\in\mathbb{N}} \subset [0,1]^d$
- Define $m_1 = f(x_1)$ and subsequently $m_{i+1} = \max(m_i, f(x_{i+1})),\ \forall i \in \mathbb{N}$
- Niederreiter [1, 2]: $m_n \xrightarrow{n\to\infty} M \iff f$ "sufficiently continuous" and points well distributed

$$M - \omega(d_N) \leq m_N \leq M, \quad \omega(t) := \sup_{||x-y||\leq t} |f(x) - f(y)|$$

- By the dispersion of the point set $\left(x_n\right)_{n=1}^{N}$ we mean

$$d_N = \max_{x\in[0,1]^d} \min_{1\leq n\leq N} ||x - x_n||$$

Discrepancy and Integration

- Approximation of the integral

$$I_N := \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

- If points are uniformly distributed, then

$$I_N \xrightarrow{N\to\infty} \int\limits_{[0,1]^d} f(x)dx$$

- Error in approximation proportional to discrepancy

$$D_N = \sup_B \left| \frac{\#(X \cap B)}{\#X} - \mu(B) \right|,$$

where B are boxes with axes parallel to the cube

Dispersion

Definition 1

Let $X \subset [0,1]^d$, $d \in \mathbb{N}$ be a set of points in the space $\mathbb{R}^d$. By the dispersion of the set X we mean

$$\operatorname{disp}(X) := \sup_{B\colon B \cap X = \emptyset} |B|,$$

where $B = I_1 \times \cdots \times I_d$, $\forall j \in \hat{d}\colon I_j \subset [0,1]$ is a box with axes parallel to the cube and the symbol $|B|$ denotes its volume.

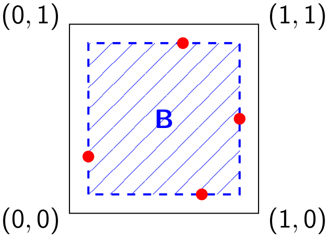

Figure: Box B forming the dispersion of $X = \{(0.7, 0.1), (0.9, 0.5), (0.1, 0.3), (0.6, 0.9)\}$ in $\mathbb{R}^2$

n-th Min. Dispersion and Its Inverse

Definition 2

Let $n, d \in \mathbb{N}$. Then the n-th minimal dispersion of the cube $[0,1]^d$ is defined as

$$\mathrm{disp}(n,d) := \inf_{\substack{X \subset [0,1]^d \\ \#X = n}} \mathrm{disp}(X)$$

and its inverse function as

$$N(\varepsilon, d) := \min\{n \colon \mathrm{disp}(n,d) \leq \varepsilon\}.$$

- Clearly, $N(\varepsilon, d) = 1$ for every $\varepsilon \in [\frac{1}{2}, 1]$ and $d \in \mathbb{N}$
- Intuitively, $\mathrm{disp}(n,d) \approx n^{-1} \;\; \forall d \in \mathbb{N}$

Illustration of Variable Behavior

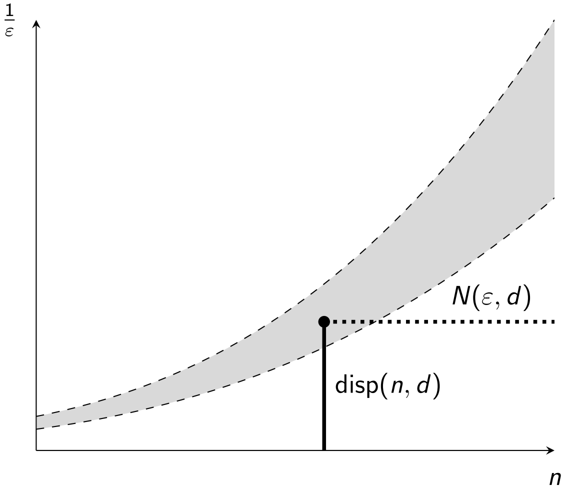

Figure: The relationship between $\mathrm{disp}(n, d)$ and $N(\varepsilon, d)$.

Estimation Development

- Lower bounds are much more difficult
- Upper bounds often use probabilistic methods
- Elementary estimate using the *pigeonhole principle*

$$\frac{1}{n+1} \leq \mathrm{disp}(n, d) \implies \frac{1}{\varepsilon} - 1 \leq N(\varepsilon, d)$$

- C. Aistleitner et al. [3]: $\exists C > 0$, $\forall d \in \mathbb{N}$ and $\varepsilon \in (0, \frac{1}{4})$:

$$C \frac{\log d}{\varepsilon} \leq N(\varepsilon, d). \tag{1}$$

- A. Litvak and G. V. Livshyts [4]: for any $d \geq 2$ and $\varepsilon \in (0, \frac{1}{2}]$:

$$N(\varepsilon, d) \leq 12e \frac{4d \ln \ln \left(\frac{8}{\varepsilon}\right) + \ln \left(\frac{1}{\varepsilon}\right)}{\varepsilon}. \tag{2}$$

Applications of Dispersion

- Data-mining [5]
  - Some attributes never occur together
  - Identification and quantification of empty spaces in data
  - Useful for outlier analysis, anomaly detection, clustering, ...
- Quasi Monte-Carlo methods [6]
  - Points with low dispersion are better than random sampling
  - The difference is notably larger with increasing dimension
- Cutting undamaged parts of iron from a damaged block [7]
- Anywhere uniform point distribution is needed:
  - Optimization
  - Genetic algorithms
  - Computer graphics, ...

Frequent Proof Principle

- Transition from infinitely many boxes to a discrete plane.
- Use of a testing set of boxes of volume greater than ε.
- Each box must intersect.
- Size of such a set $\rightarrow N(\varepsilon, d) \rightarrow \text{disp}(n, d)$.

Estimation via Restriction Set

- Use of cubes with one short and the remaining sides long.

Definition 3 (Testing Cubes)

Let $k, d \in \mathbb{N}$, $A \subset \{1, \dots, d\}$, and $j \in \{1, \dots, d\} \setminus A$. Define a testing cube $B_{j,A} = I_1 \times \cdots \times I_d \subset [0,1]^d$ as:

- $I_j = (0, 2\varepsilon)$,
- $\forall i \in A : I_i = (2\varepsilon, 1)$,
- $\forall l \in \{1, \dots, d\} \setminus (A \cup j) : I_l = (0, 1)$.

Finally, construct the set $\mathcal{B} = \{B_{j,A} : A \subsetneq \{1, \dots, d\}, j \in \{1, \dots, d\} \setminus A\}$.

- $|A| \approx \frac{1}{\varepsilon} \implies |B_{j,A}| > \varepsilon$.
- If $X = \{x^1, \dots, x^n\} \cap \mathcal{B} \neq \emptyset$, then

$$\forall A, \forall j, \exists u \in \hat{n} : (x^u)_j \in (0, 2\varepsilon) \quad \text{and} \quad (x^u)\big|_A \in \prod_{i=1}^{|A|} (2\varepsilon, 1)$$

$$\iff \phi(x^u)_j = 0 \wedge \phi(x^u)\big|_A = 1$$

Restriction Set and Dispersion Estimation

Definition 4

Let $N, l, d \in \mathbb{N}$ such that $1 \leq l \leq d$. We say a set of points $x^1, \ldots, x^N \in \{0,1\}^d$ is (l,d)-restriction set if $\forall A \subset \{1, \ldots, d\} : |A| = l - 1$ and $\forall j \in \{1, \ldots, d\} \setminus A$, there is a point x^u with

$$(x^u)_j = 0 \;\wedge\; (x^u)\big|_A = 1.$$

Subsequently, define the size of the smallest (l,d)-restriction set as

$$R(l,d) = \min\{N \in \mathbb{N} : \exists\{x^1, \ldots, x^N\} \subset \{0,1\}^d \text{ that is } (l,d)\text{-restriction set}\}.$$

- Probabilistic and combinatorial estimations can be constructed on $R(l,d)$.
- It can be shown that $R\left(2^{k-2}, d\right) \leq N\left(2^{-k}, d\right)$.

Corollary 5

There exists a constant $C > 0$ such that for any $\varepsilon \in \left(0, \frac{1}{2}\right)$ and $d \in \mathbb{N}, d \geq 2$,

$$C \frac{\log d}{\varepsilon} \leq N(\varepsilon, d).$$

r-cover-free families

- It can be shown that the concept of a restriction set is equivalent to that of an r-cover-free system.

Definition 6

Let $d, r \in \mathbb{N}$ with $r < d$, and $\mathcal{F} = \{F_1, \ldots, F_d\}$ be a system of subsets of set X. We say $\mathcal{F}$ is *r-cover-free* if

$$\forall A \subset \{1, \ldots, d\}, |A| = r, \ \forall j \in \{1, \ldots, d\} \setminus A : \ F_j \not\subset \bigcup_{i \in A} F_i.$$

Finally, define the smallest size of set X as

$$C(1, r, d) = \min\{n \in \mathbb{N} : \{F_1, \ldots, F_d\} \subset X^d, |X| = n \text{ is } r\text{-cover-free}\}.$$

Estimation of Dispersion using r-cover-free families

- N. Alon, V. Asodi [8]: $\exists c > 0, \forall r, d \in \mathbb{N}$: $r \leq 2\sqrt{d}$ such that $c\frac{r^2 \log d}{\log r} < C(1, r, d)$.
- Proof principle analogous to the previous one.

Theorem 7

There exists $c > 0$ such that for any $d \geq 2$ and ε satisfying $\frac{1}{4} \geq \varepsilon \geq \frac{1}{4\sqrt{d}}$, the following holds:

$$N(\varepsilon, d) > \frac{c \log d}{\varepsilon^2 \cdot \log \frac{1}{\varepsilon}}. \quad (3)$$

- Limitation that estimation holds only for limited $\varepsilon \rightarrow$ generalization and extension.

Generalization of (w, r)-cover-free concept

- Non-coverage of a single set can be generalized to non-coverage of intersections of multiple sets.
- $\{F_1, \dots, F_d\}$ is (w, r)-cover-free if $\bigcap_{j \in W} F_j \not\subset \bigcup_{i \in A} F_i$.
- Allows considering cubes with more than one short edge.
- Using estimation for such sets, dispersion can again be estimated.
- Resulting estimation will be valid even for smaller ε.
- Also utilizing the recurent relationship

$$N(\xi, d) \geq k \cdot N(k\varepsilon, d) \quad \forall k \in \mathbb{N}, k\xi = \varepsilon$$

- Currently working on a rigorous mathematical proof.

- H. Niederreiter, P. Peart, *Localization of search in quasi-Monte Carlo methods for global optimization*. SIAM J. Sci. Stat. Comput. 7, 1986, 660-664.
- H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM, Philadelphia, 1992.
- C. Aistleitner, A. Hinrichs, D. Rudolf, *On the size of the largest empty box amidst a point set*. Discrete Applied Mathematics 230, 2017, 146–150.
- A. E. Litvak, G. V. Livshyts, *New bounds on the minimal dispersion*. Journal of Complexity 72, 2022, 101648.
- J. Edmonds, J. Gryz, D. Liang, R. J. Miller, *Mining for empty spaces in large data sets*. Theoretical Computer Science 296(3), 2003, 435–452.
- G. Rote, R. F. Tichy, *Quasi-Monte-Carlo methods and the dispersion of point sequences*. Mathematical and Computer Modelling 23, 1996, 9–23.
- A. Naamad, D. T. Lee, W.-L. Hsu, *On the maximum empty rectangle problem*. Discrete Applied Mathematics 8(3), 1984, 267–277.
- N. Alon, V. Asodi, *Learning a hidden subgraph*. SIAM Journal on Discrete Mathematics, 18(4), 2005, 697-712.

Conclusion

Thank you for your attention!