Stochastic Global Optimization: promises and limitations

Anatoly Zhigljavsky

Cardiff School of Mathematics, Cardiff University, WALES

EUROPT 2016, Warsaw, July 2, 2016

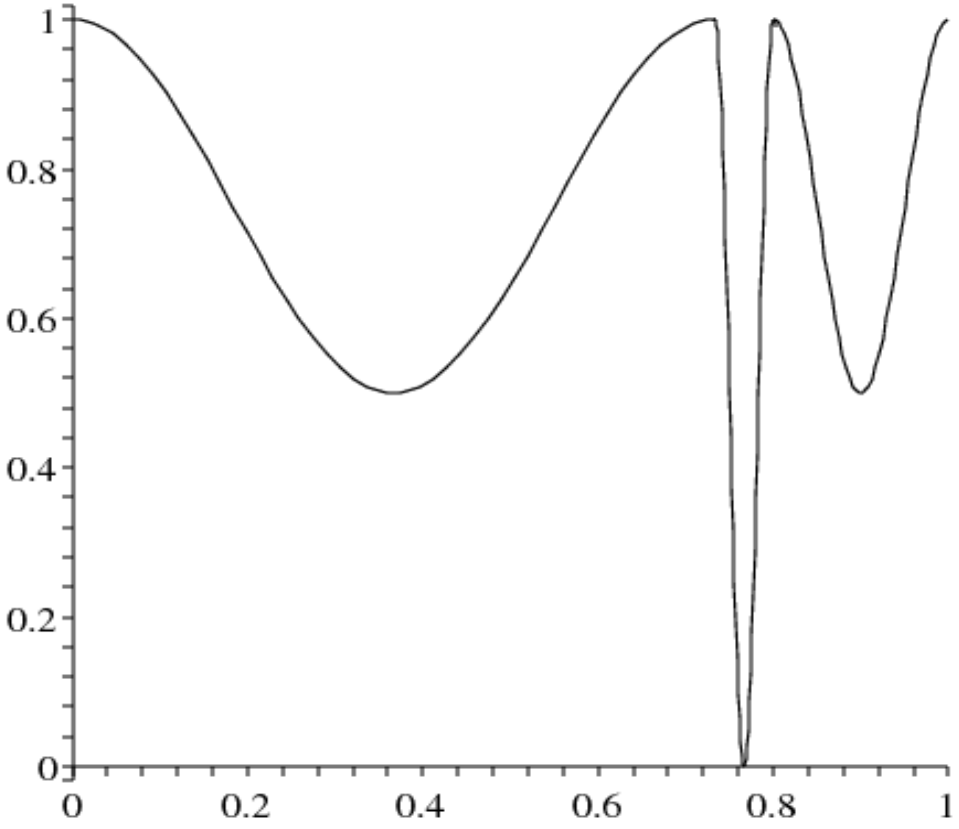
WALES CHAMPION !

CONTENTS

Part I. Introduction

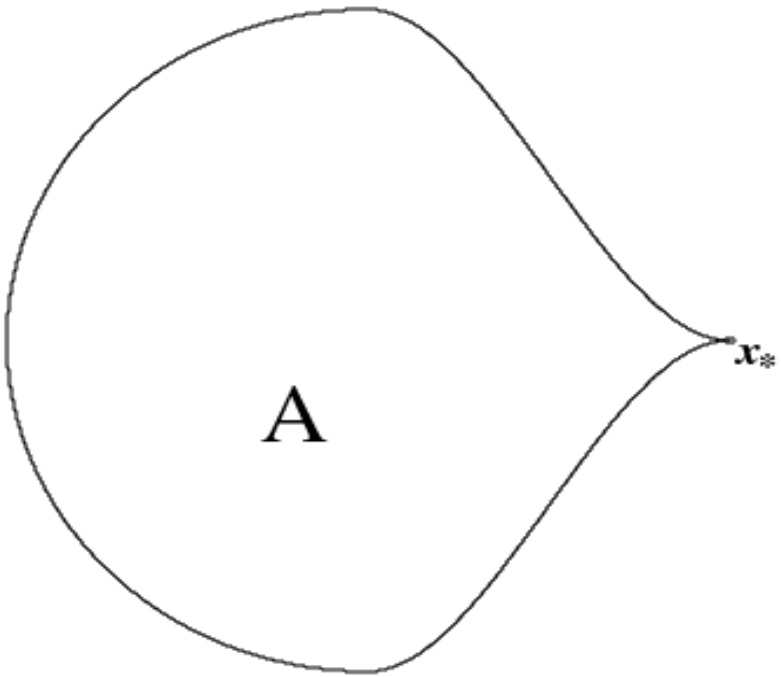
Global optimization: Statement of the problem

$$f(x) \to \min_{x \in A}; \quad x_* = \arg\min_{x \in A} f(x)$$



Typical assumptions on A

- simple structure
- situation like below is dissalowed



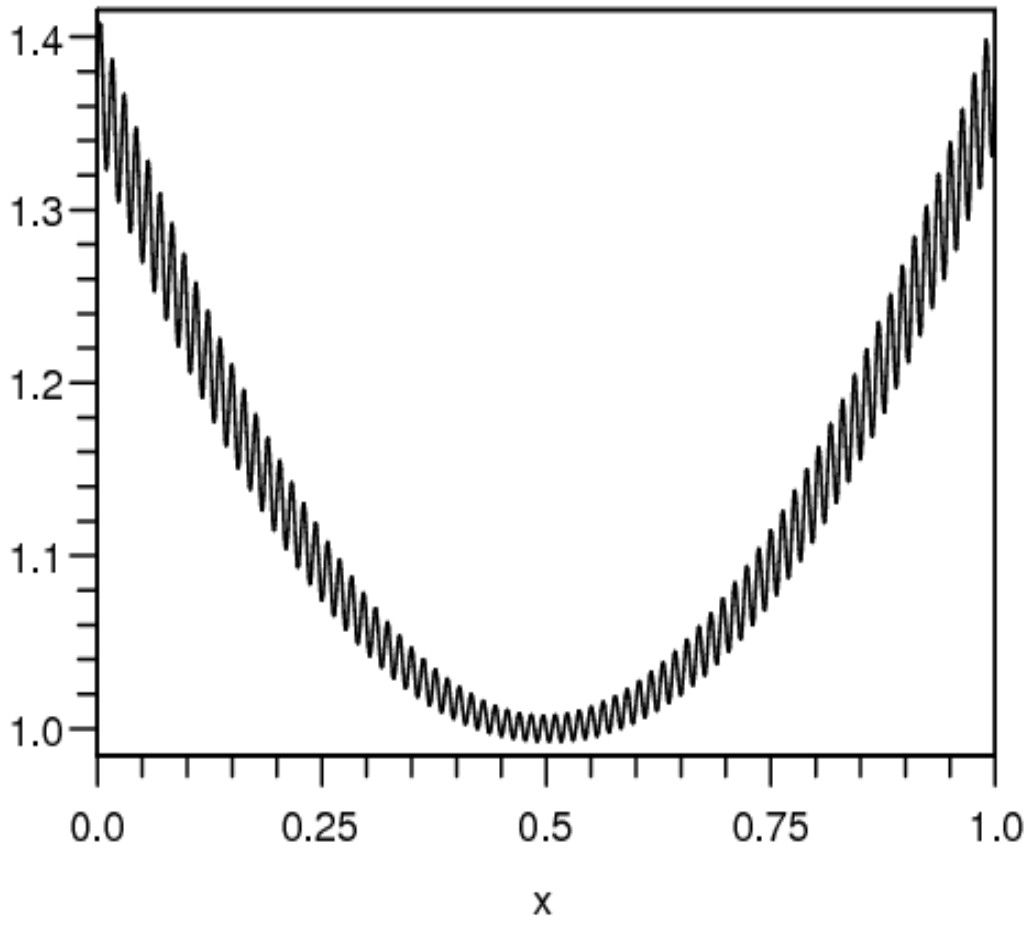
Typical assumptions on f

- Lipschitz condition

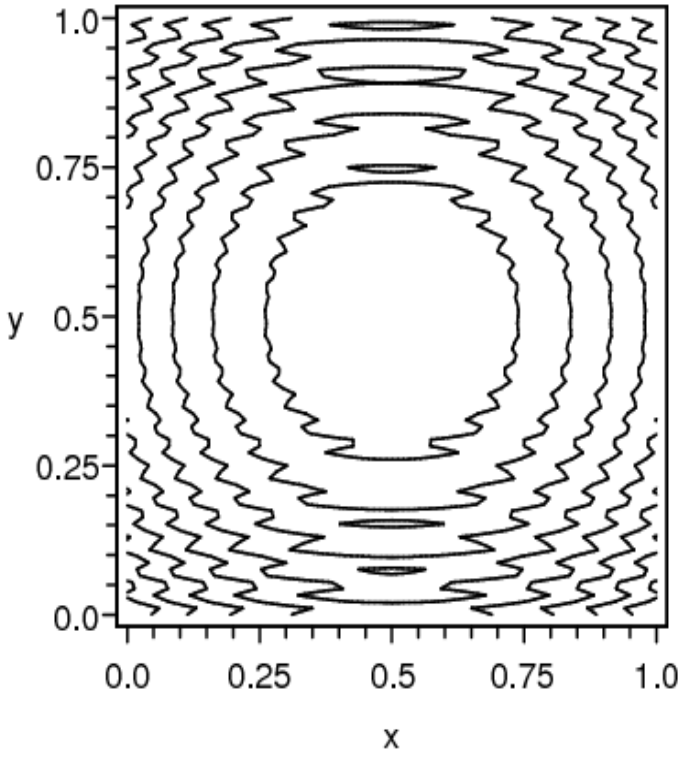
$$|f(x) - f(x')| \leq L||x - x'||$$

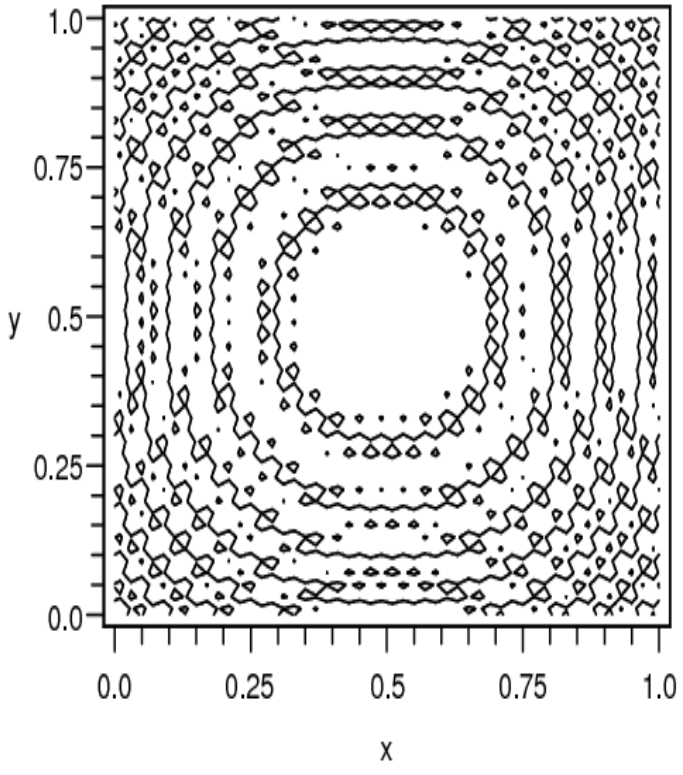
- smoothness
- restrictions on the number of local minima or on the volumes of the domains of attractions of local minimizers (or just the global one)
- special assumptions like $f = f_1 - f_2$ with f_1 and f_2 convex (so-called DC programming)
- explicit (like polynomial) expressions allowing the use of interval methods

Is it a multiextremal function?



The same but in two dimensions





Part II. Elements of multivariate geometry

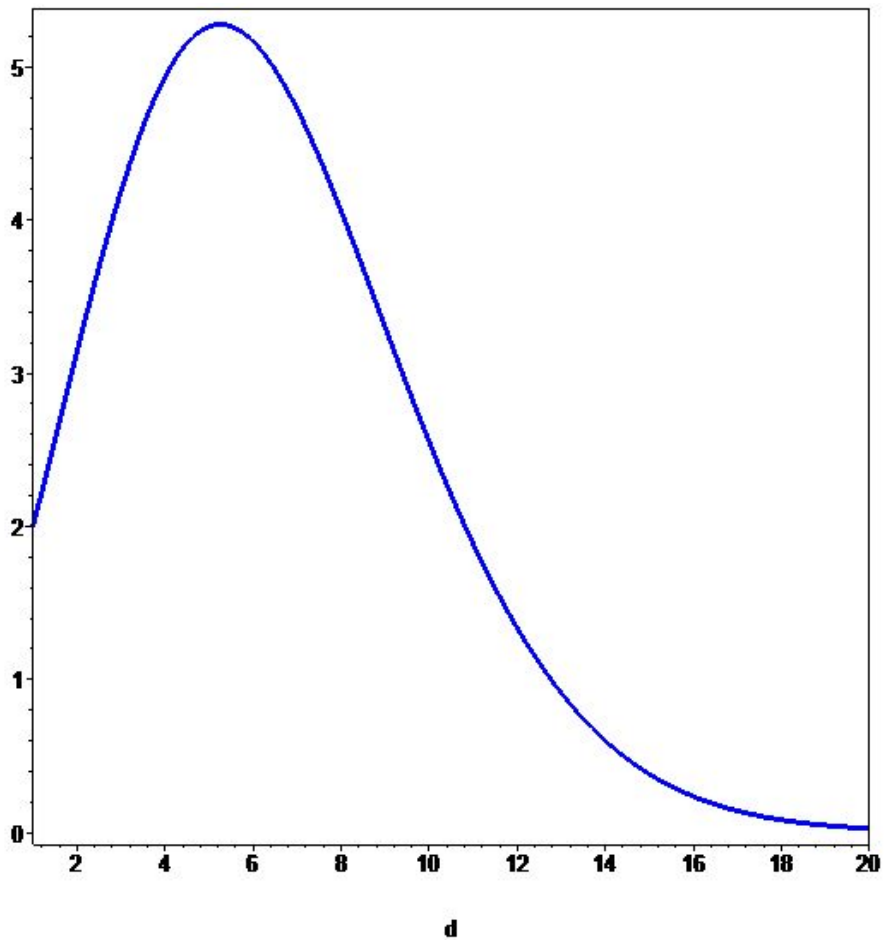
References

K. Ball (1997) An elementary introduction to modern convex geometry

J. Hopcroft, R. Kannan (2015) Foundations of Data Science

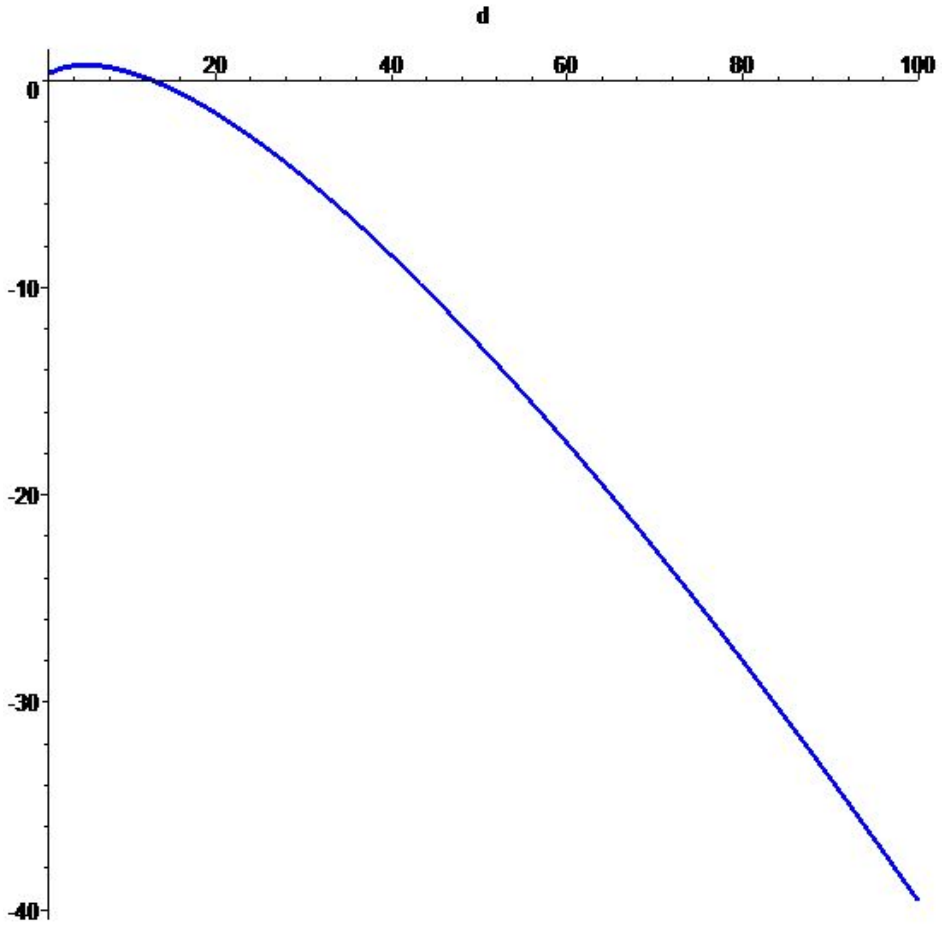
Volume of the d-dimensional unit ball $B(0,1) = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$

$$V_d = \mathrm{vol}(B(0,1)) = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$$



Volume of the d-dimensional unit ball

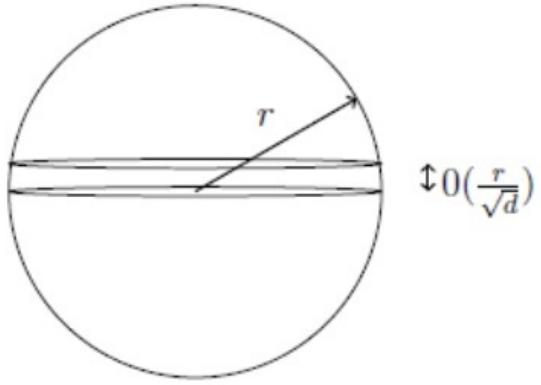
$\log_{10} V_d$ as a function of d:



F.e., $V_{100} \simeq 2.368 \cdot 10^{-40}$

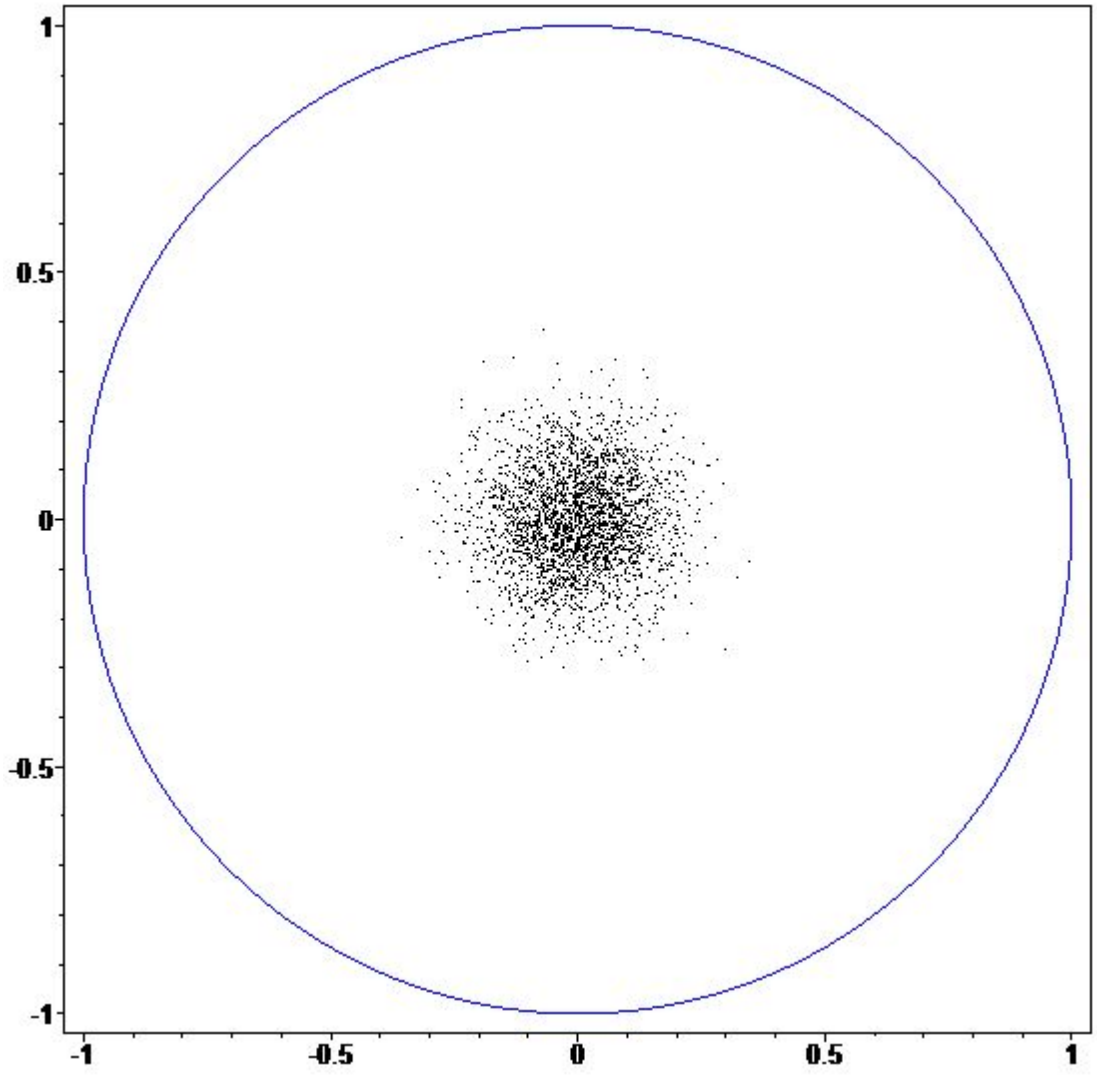
d-dimensional ball

Almost all the volume is near the equator:



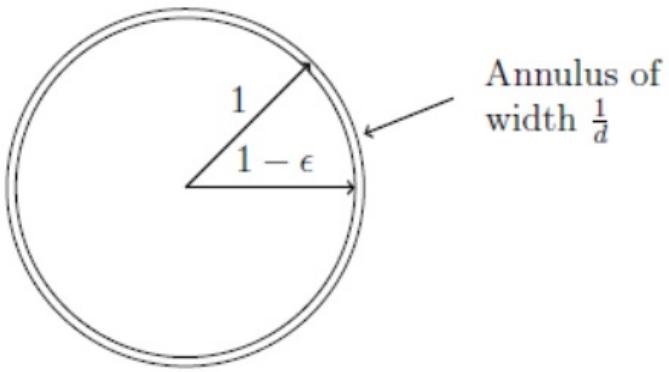
Th. *For any $c > 0$, the fraction of the volume of the unit ball above the plane $x_1 = c/\sqrt{d-1}$ is less than $\frac{2}{c}\exp\{-c^2/2\}$.*

Random points in a ball; projection to 2 dimensions



d-dimensional ball

Almost all the volume is also there (in $B(0,1) \setminus B(0,1-\epsilon)$ with $\epsilon = c/d$):

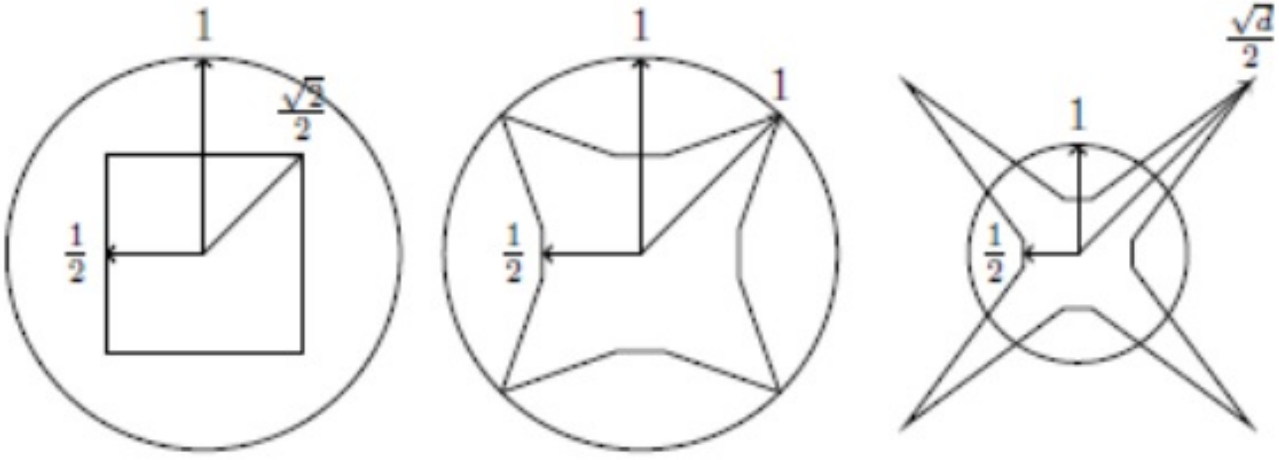


Indeed, $\text{vol}(B(0,1-\epsilon))/\text{vol}(B(0,1)) = (1-\epsilon)^d \simeq 0$ for $\epsilon = c/d$, large d and c fixed but large enough.
Radius of a uniform random point has density $p_d(r) = dr^{d-1}, 0 \leq r \leq 1$.

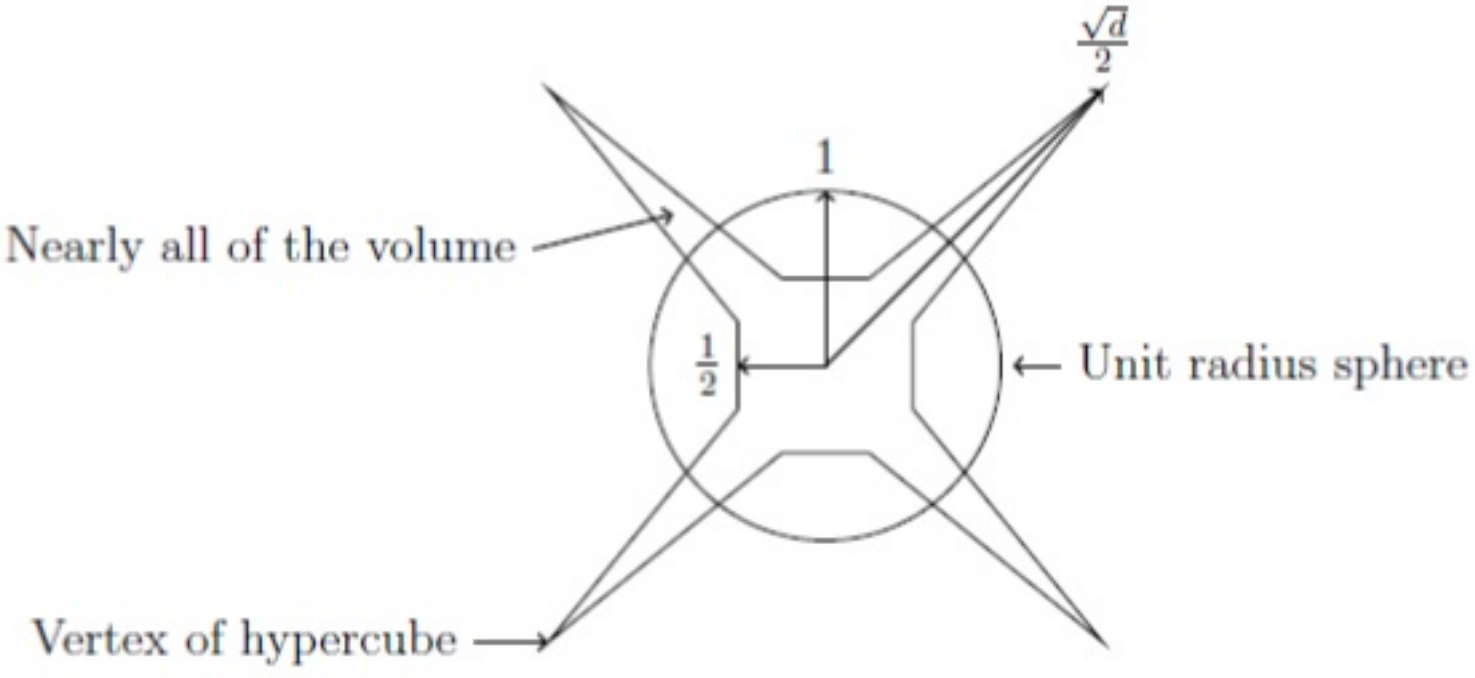
d-dimensional cube and ball

Unit cube: $\{x = (x_1, \ldots, x_d) \in \mathbb{R}^d : |x_i| \leq 1/2\}$
Unit ball: $B(0,1) = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$
Length of the cube's half-diagonal:

$$\sqrt{\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \ldots + \left(\frac{1}{2}\right)^2} = \frac{\sqrt{d}}{2}$$

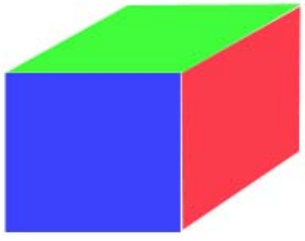


d-dimensional cube

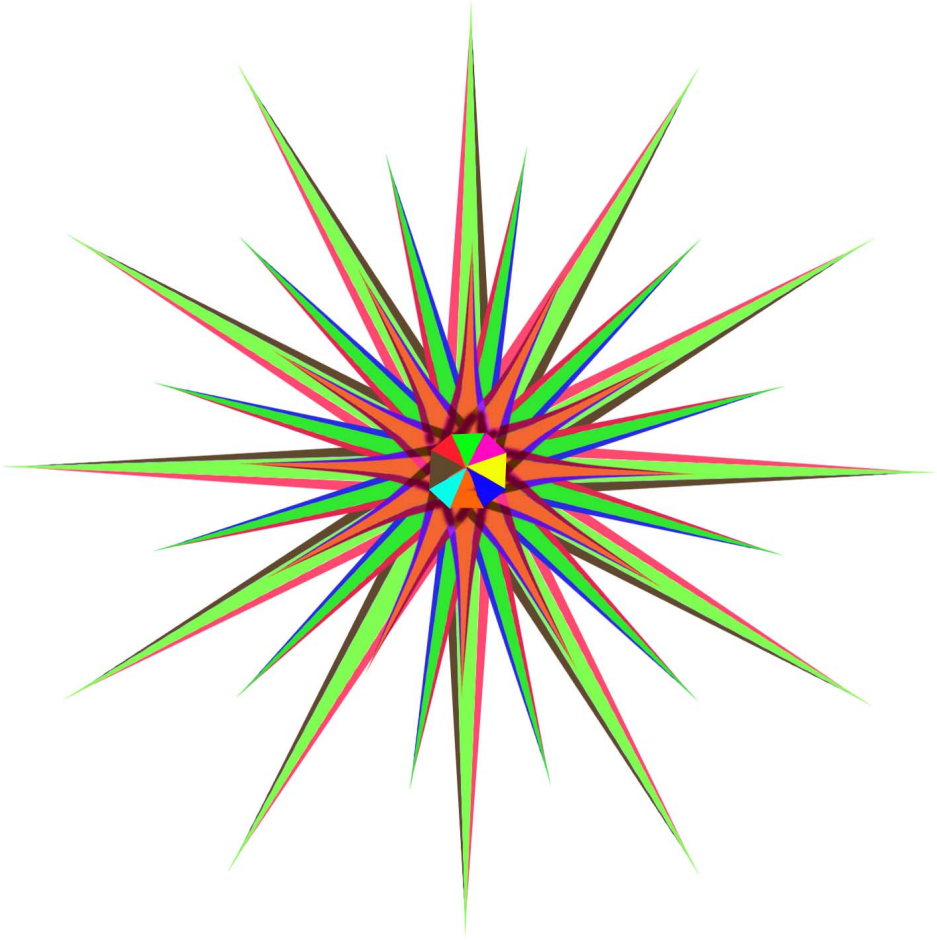


What is a shape of the d-dimensional cube?

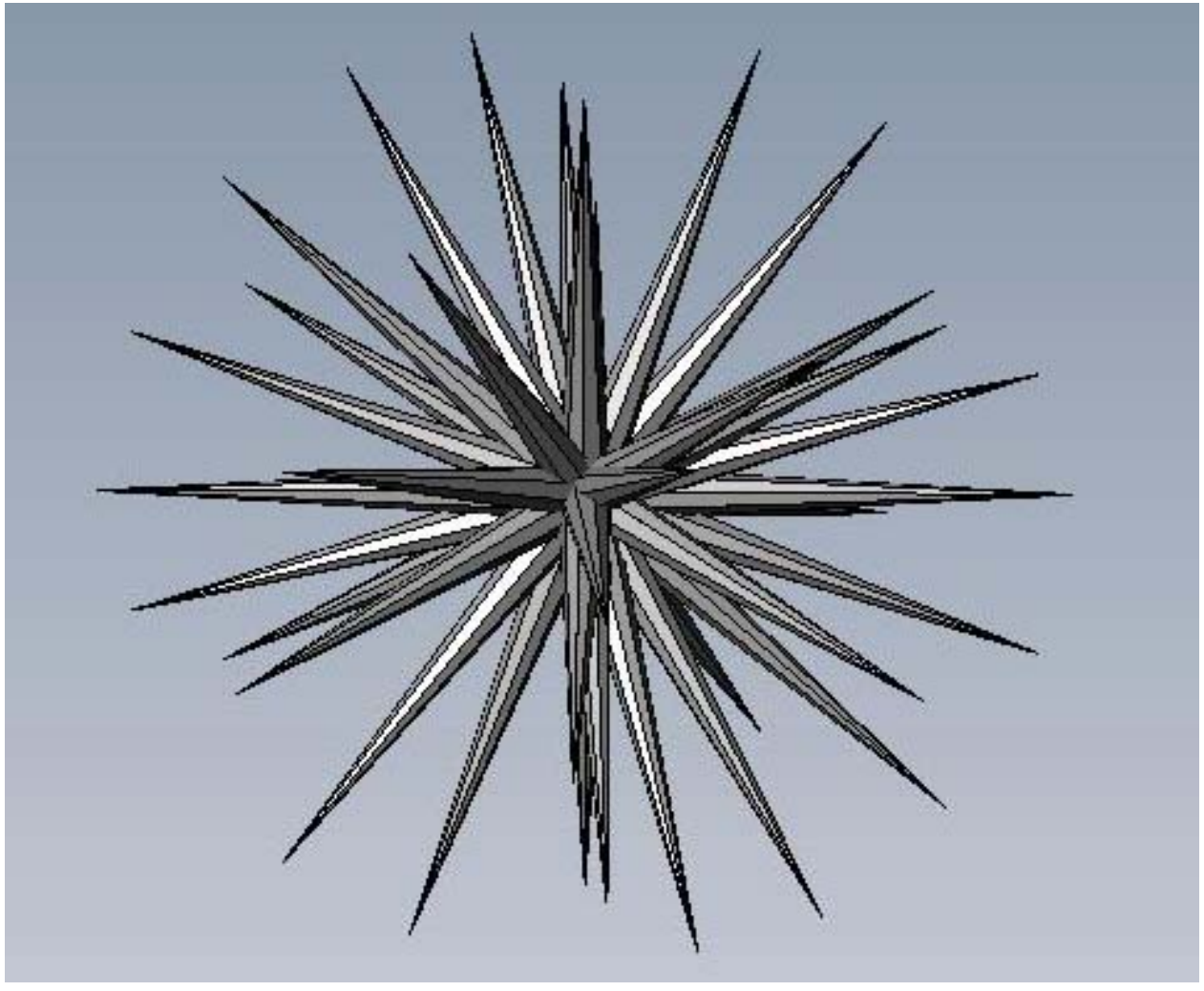
$[-\frac{1}{2}, \frac{1}{2}]^2$

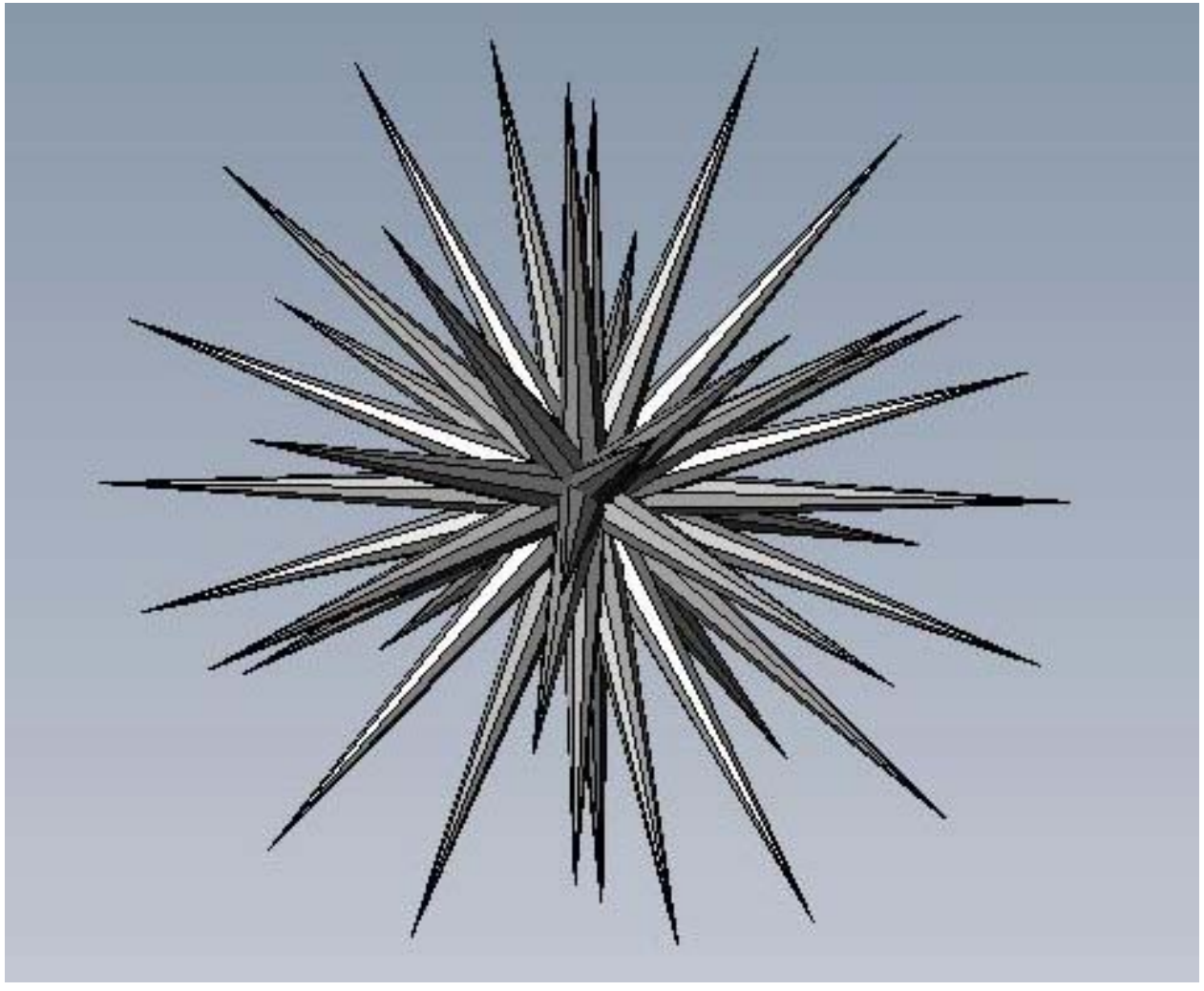


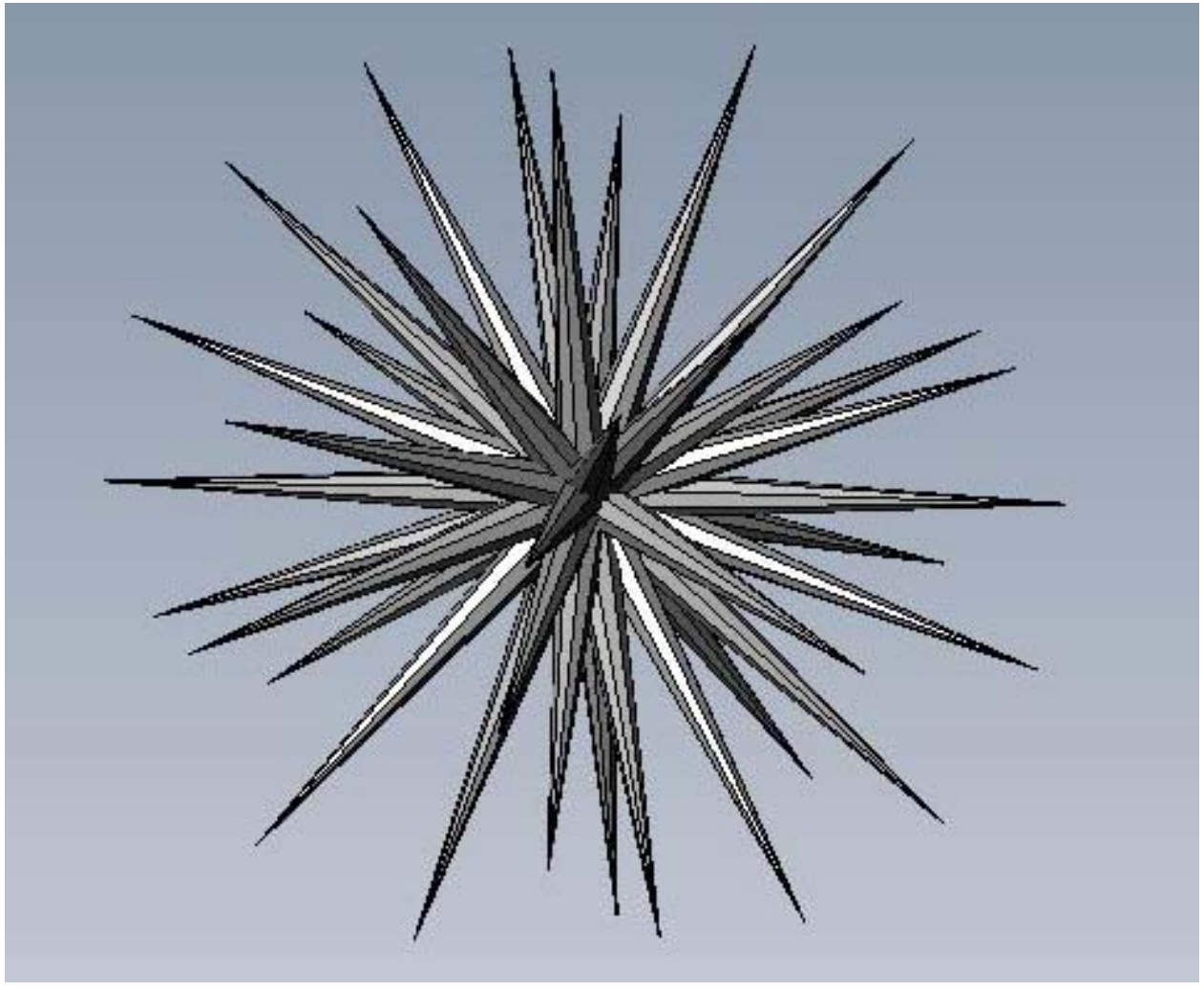
$[-\frac{1}{2}, \frac{1}{2}]^3$

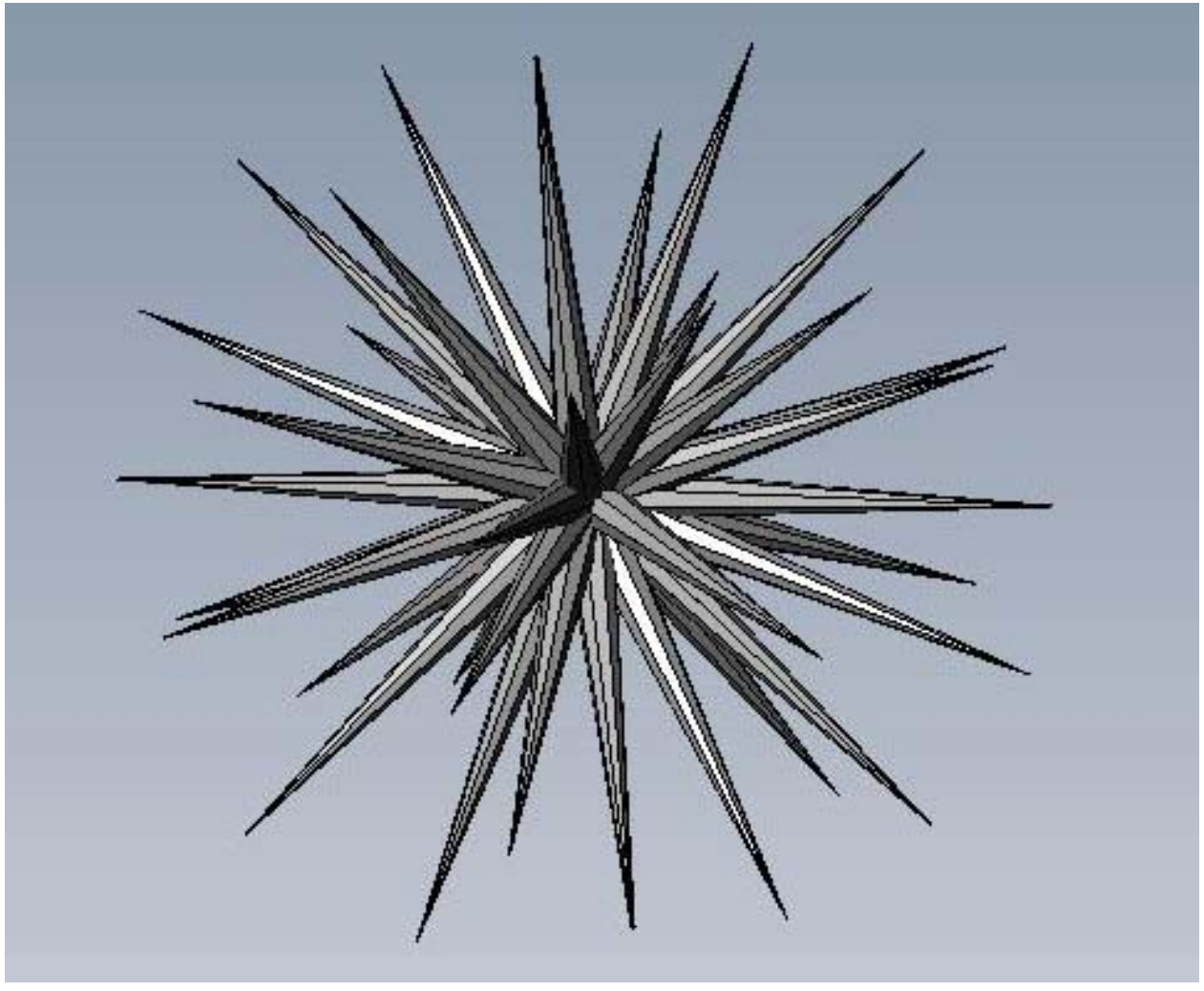


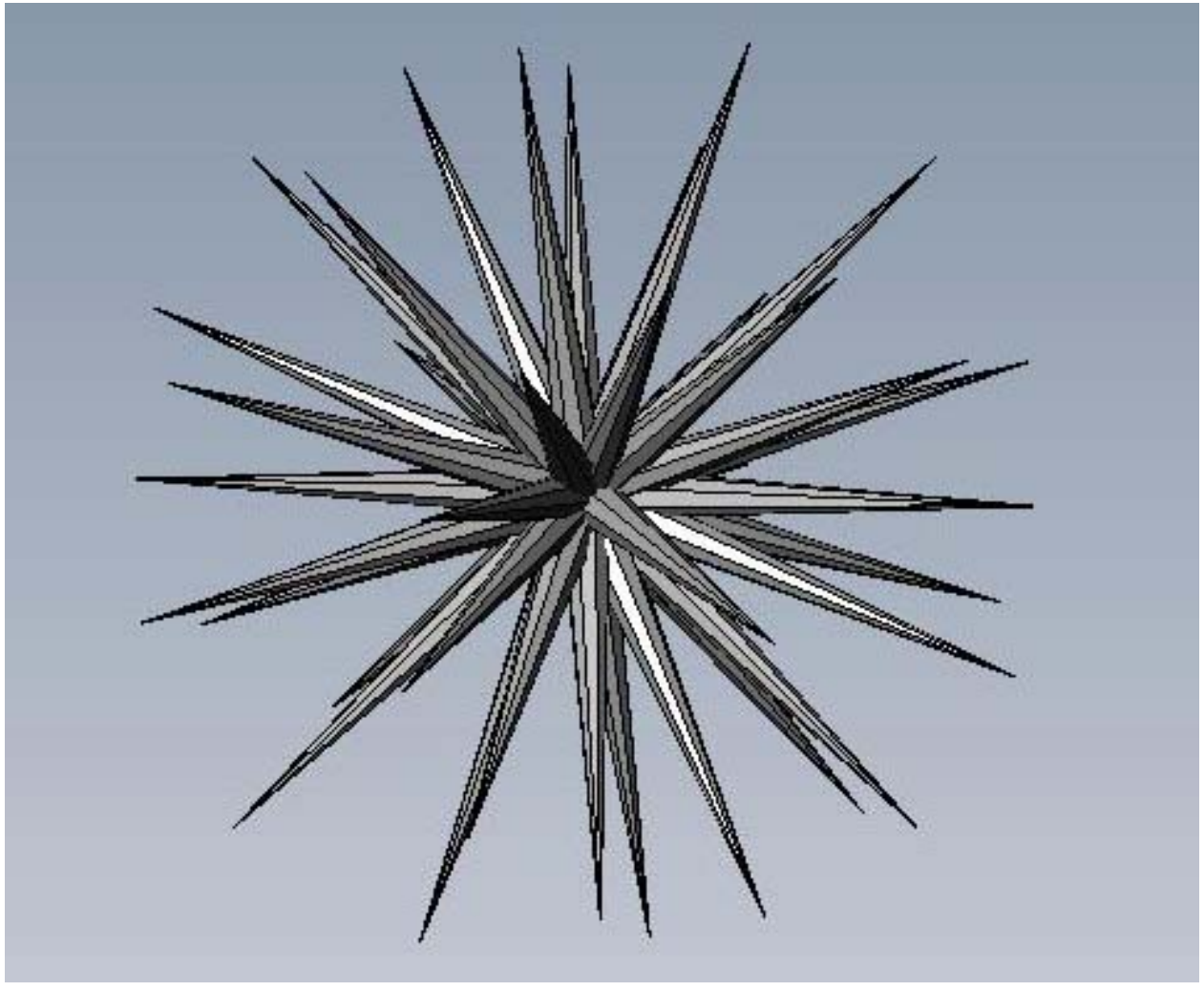
$[-\frac{1}{2}, \frac{1}{2}]^8$





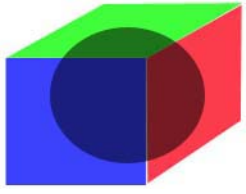


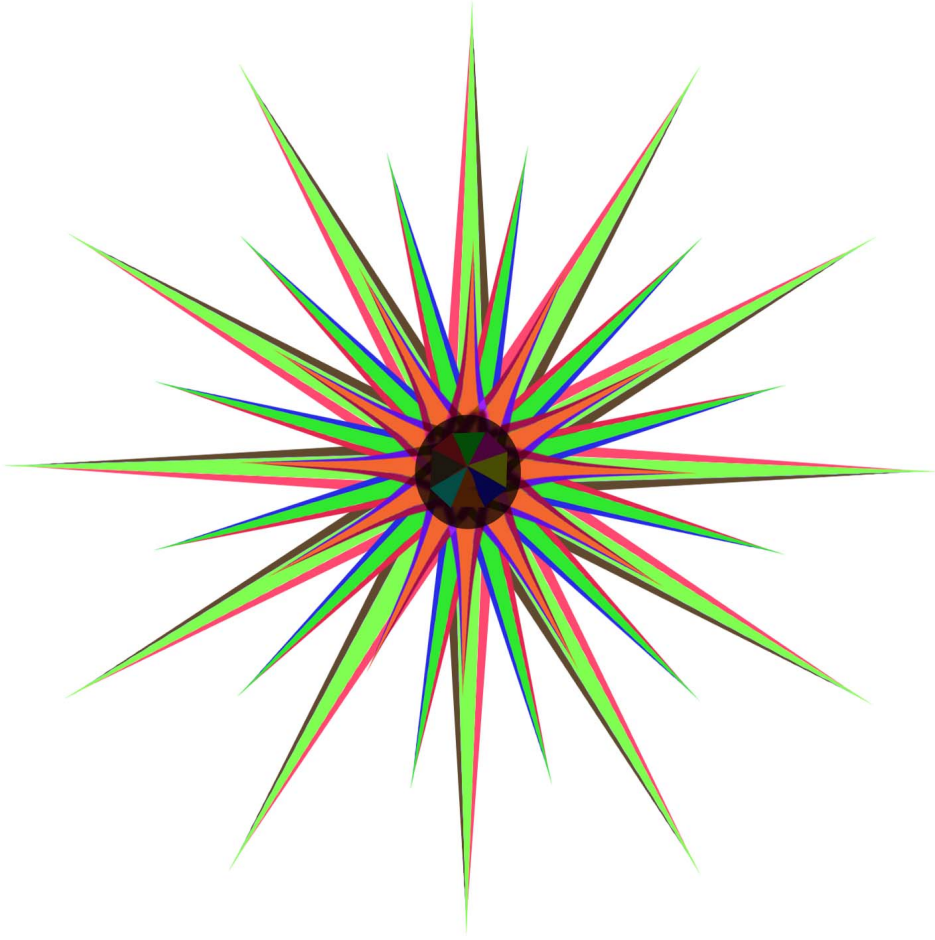




Volume of the largest inscribed ball into the unit cube

Volume of the cube $=1$, $v_d = \frac{\pi^{d/2}}{2^d \Gamma(1+d/2)}$ (volume of the ball of radius $1/2$)

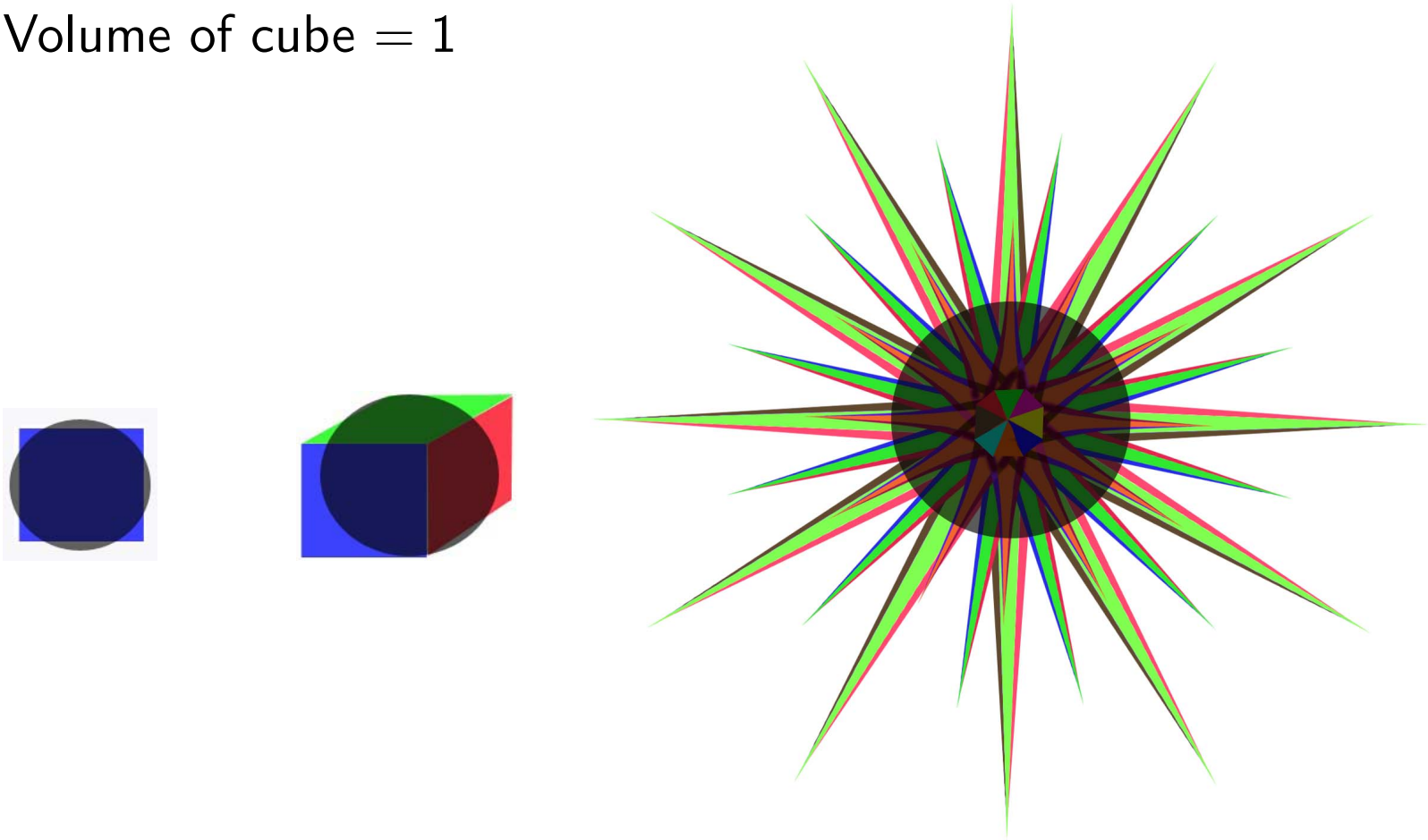




$v_2 \simeq 0.78$ $v_3 \simeq 0.52$ $v_{10} \simeq 0.0025$ $v_{20} \simeq 0.25 \cdot 10^{-7}$ $v_{100} \simeq 10^{-70}$

Radius of the ball of volume 1

Volume of cube $= 1$



$r_2 \approx 0.56$; $r_3 \approx 0.62$; $r_8 \approx \sqrt{\frac{d}{2\pi e}}\Big|_{d=8} \approx 0.84$; $r_d \sim c\sqrt{d}$, $d \to \infty$

If we project the mass distribution of the ball of volume 1 onto a single direction, we get a distribution that is approximately Gaussian with variance $1/(2\pi e)$, which does not depend on d.

Gaussian random vectors

If x is Gaussian $N(0, I_d)$ then distance from the origin

$$r = \sqrt{\sum_{i=1}^{d} x_i^2}$$

is very close to $\sqrt{d}$: for any $0 < \beta < \sqrt{d}$,

$$Pr\{\sqrt{d} - \beta \leq r \leq \sqrt{d} + \beta\} \geq 1 - 3\beta^2/64$$

Two i.i.d. Gaussian vectors are almost orthogonal to each other. Similar for uniform r.v. in a ball and in a cube.

Uniform points in a cube are at almost the same distance from each other

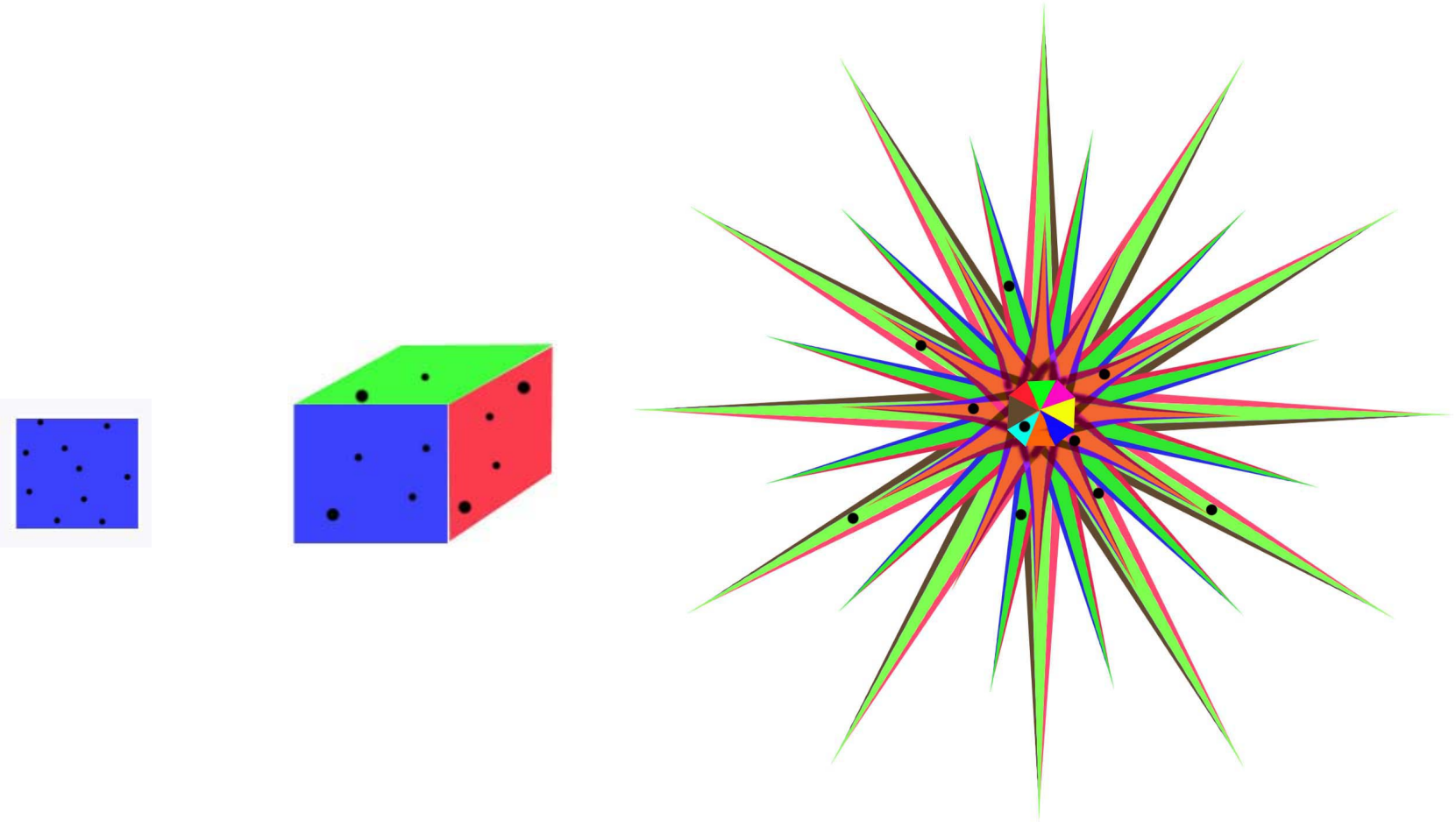
The distribution of the distances

$$\|x - y\| = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$$

is concentrated around its expected value which is approximately $\sqrt{d/6}$.

Similar result holds for the ball.

10 uniform points

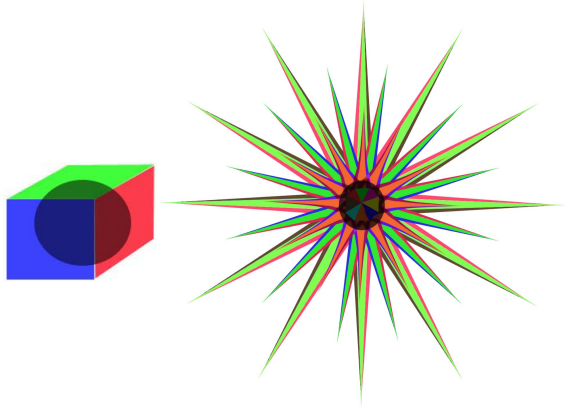


Distance between points grows very fast with dimension d; all points are at approximately the same distance from each other.

Summary: Features of the d-dimensional unit cube for large d

In high-dimensional space

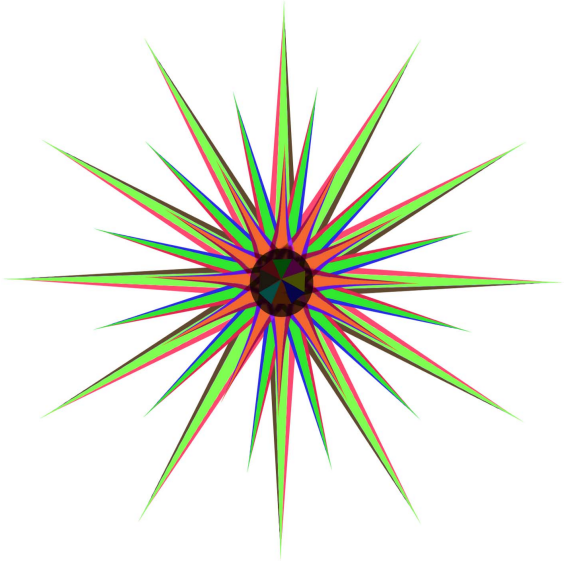
- The 'middle' of the cube is virtually empty.
- The cube is a 'union' of its corners.
- The 'average' radius of the unit cube is about $\sqrt{\frac{d}{2\pi e}}$. At the same time, the distance from the center to the middle of cube's facets is 0.5 for any dimension d.



Consequences for the global optimization

If the dimension of the feasible domain is high then:

- our 2D and 3D intuition may be totally misleading
  (so that we can devise bad algorithms and make conceptual mistakes in their analysis);
- the test functions (like Rastrigin and Schekel) we use in dimensions 20 or more do not reflect reality;
- the tables below will not look surprising.



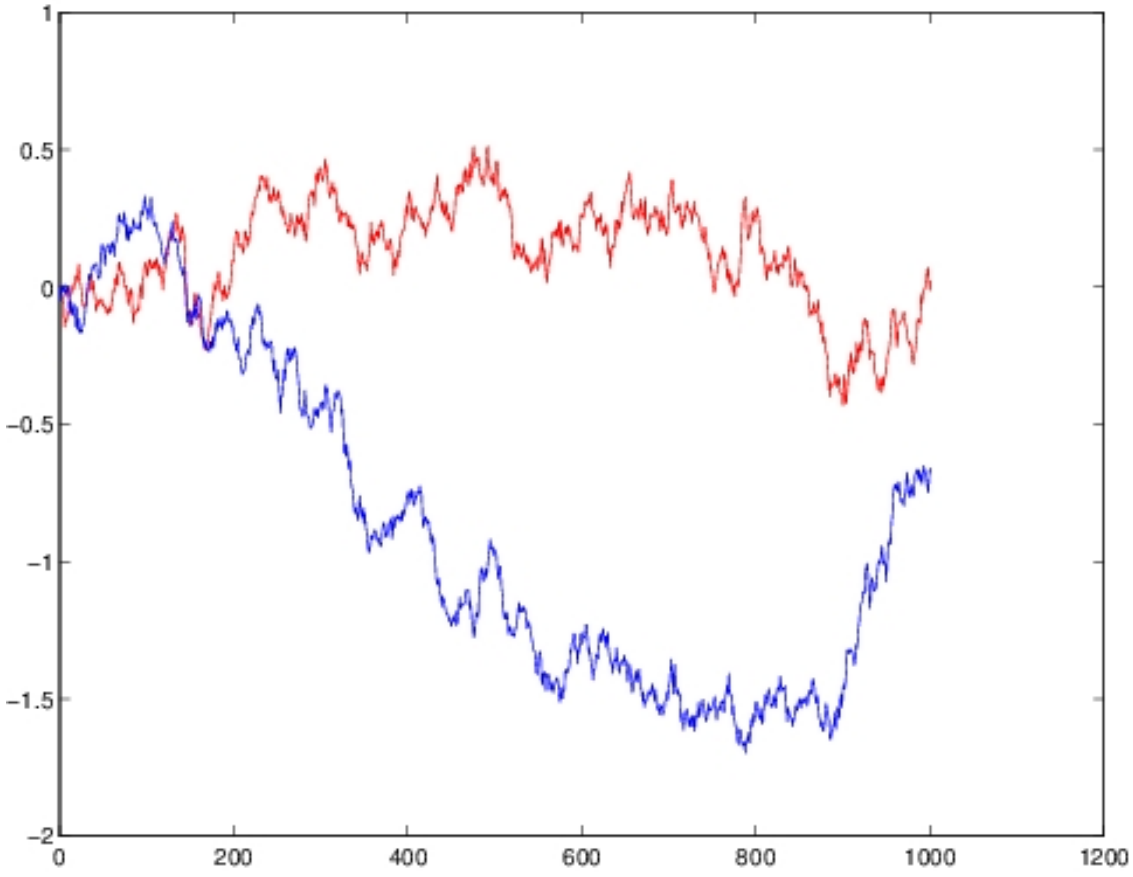
Main topic:
Stochastic global optimization

Stochastic global optimization

- Global random search (methodology, theory)
- Stochastic models about the objective function (kriging/zilinskasing)
- Heuristics
- Applications

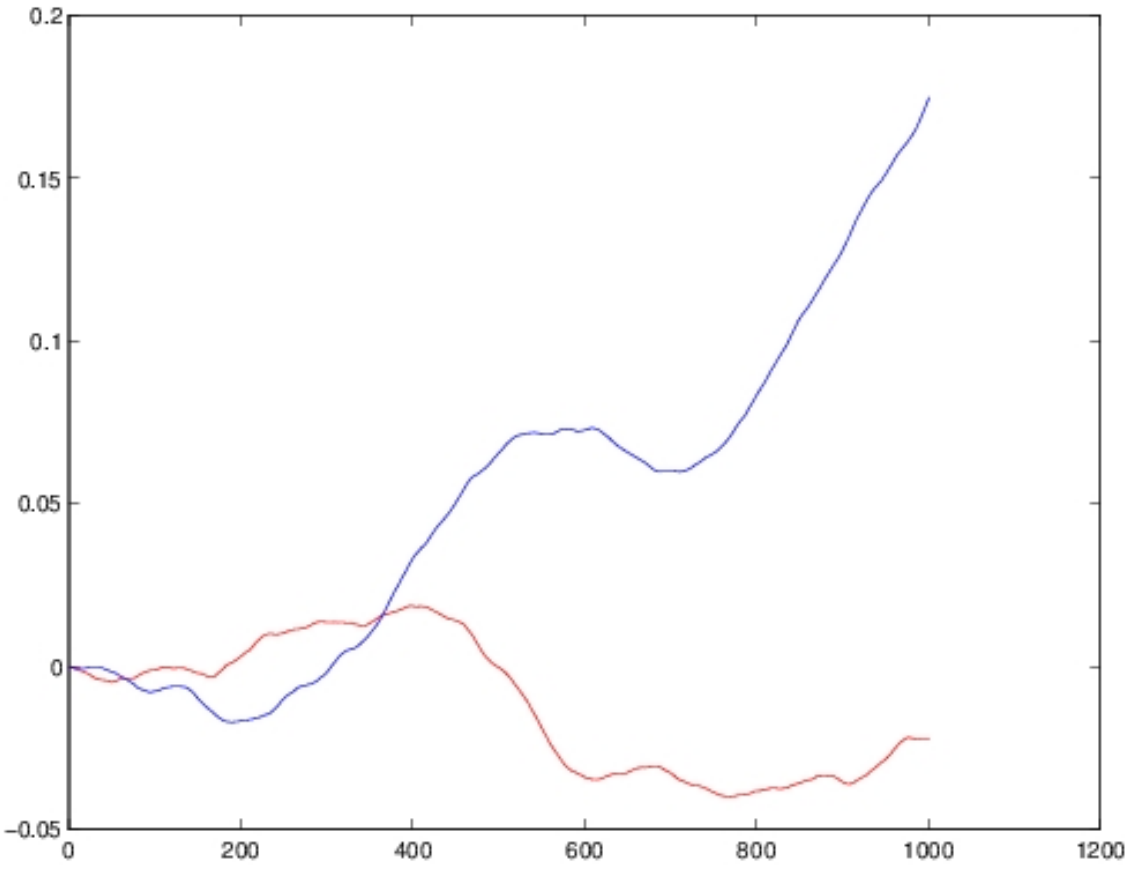
Stochastic models about the objective function

Sample paths of the Wiener process:



Stochastic models about the objective function

Sample paths of the integrated Wiener process:



Global random search (theory)

- Convergence and rate of convergence
- Statistical inference
- Clever choice of updating rules using probabilistic considerations
- Decrease of randomness in choosing the points and making the decisions

Convergence: Borel-Cantelli lemma

Global random search algorithm converges if

$$\sum_{j=1}^{\infty} \inf P_j(B(x_*, \varepsilon)) = \infty \tag{1}$$

for any $\varepsilon > 0$, where $B(x_*, \varepsilon) = \{x \in A : ||x - x_*|| \leq \varepsilon\}$; the infimum in (1) is taken over all possible previous points and the results of the objective function evaluations at them.
Standard choice of probability distributions to guarantee convergence:

$$P_{j+1} = \alpha_{j+1} P_U + (1 - \alpha_{j+1}) Q_j \,, \quad \sum_j \alpha_j = \infty \,.$$

Rate of convergence

The number of points required to get precision ε with probability $\geq 1 - \gamma$, for different dimensions d:

d	$\gamma = 0.1$			$\gamma = 0.05$		
	$\varepsilon = 0.5$	$\varepsilon = 0.2$	$\varepsilon = 0.1$	$\varepsilon = 0.5$	$\varepsilon = 0.2$	$\varepsilon = 0.1$
1	0	5	11	0	6	14
2	2	18	73	2	23	94
3	4	68	549	5	88	714
5	13	1366	43743	17	1788	56911
10	924	$8.8 \cdot 10^{6}$	$9.0 \cdot 10^{9}$	1202	$1.1 \cdot 10^{7}$	$1.2 \cdot 10^{10}$
20	$9.4 \cdot 10^{7}$	$8.5 \cdot 10^{15}$	$8.9 \cdot 10^{21}$	$1.2 \cdot 10^{8}$	$1.1 \cdot 10^{16}$	$1.2 \cdot 10^{22}$
50	$1.5 \cdot 10^{28}$	$1.2 \cdot 10^{48}$	$1.3 \cdot 10^{63}$	$1.9 \cdot 10^{28}$	$1.5 \cdot 10^{48}$	$1.7 \cdot 10^{63}$
100	$1.2 \cdot 10^{70}$	$7.7 \cdot 10^{109}$	$9.7 \cdot 10^{139}$	$1.6 \cdot 10^{70}$	$1.0 \cdot 10^{110}$	$1.3 \cdot 10^{140}$

Example: $P_{j+1} = \alpha_{j+1} P_U + (1 - \alpha_{j+1}) Q_j\,, \quad \alpha_j = 1/j\,.$

Using the approximation $\sum_{j=1}^{n} \alpha_j \simeq \ln n$, we obtain
$n(\gamma) \simeq \exp\{-\ln \gamma / P_U(B)\}$.
If $A = [0,1]^d$ this gives $n(\gamma) \simeq \exp\{-\ln \gamma / P_U(B)\}$.
Assuming further $B = B(x_*, \varepsilon)$ we obtain $n(\gamma) \simeq \exp\{\mathrm{const} \cdot \varepsilon^{-d}\}$, where $\mathrm{const} = (-\ln \gamma)/V_d$ (if x_* lies closer to the boundary of A than ε then $n(\gamma)$ is even larger).
For example, for $\gamma = 0.1$, $d = 10$ and $\varepsilon = 0.1$, $n(\gamma)$ is a number larger than $10^{1000000000}$
Even for $d = 3$, $\gamma = 0.1$ and $\varepsilon = 0.1$, the value of $n(\gamma)$ is huge:
$n(\gamma) \simeq 10^{238}$.

Statistical inference about $M = \min f$

$$F(t) = \Pr\{x \in A : \ f(x) \leqslant t\} = \int_{f(x) \leqslant t} P(dx)$$

is the c.d.f. of the sample $\{y_j = f(x_j), j = 1, \ldots, N\}$ with M the lower end-point of the interval where F is concentrated. Here $x_j \sim P$.
The main assumption:

$$F(t) = c_0(t - M)^\alpha + \mathrm{o}((t - M)^\alpha) \quad \text{as } t \downarrow M\,.$$

c_0 is unknown but it is not important; α is the tail index (either known or unknown).
Statistical inference about $M = \min f$ are based on several smallest values extracted from the sample $\{y_j = f(x_j),\ j = 1 \ldots, n\}$.
$M = \operatorname{ess\,inf} \eta$, where η has c.d.f. $F(t)$.

Optimal linear estimator of $M = \min f$ based on k order statistics

$$\widehat{M}_{N,k} = c \sum_{i=1}^{k} \left[u_i / \Gamma(i + 2/\alpha) \right] y_{i,N} \,,$$

where $\Gamma(\cdot)$ is the Gamma-function,

$$u_i = \begin{cases} (\alpha + 1) \,, & \text{for } i = 1 \,, \\ (\alpha - 1)\Gamma(i), & \text{for } i = 1, \ldots, k-1 \,, \\ (\alpha - \alpha k - 1)\Gamma(k) \,, & \text{for } i = k \,, \end{cases}$$

$$1/c = \begin{cases} \sum_{i=1}^{k} 1/i \,, & \text{for } \alpha = 2, \\ \frac{1}{\alpha - 2} \left(\alpha \Gamma(k+1)/\Gamma(k + 2/\alpha) - 2/\Gamma(1 + 2/\alpha) \right) , & \text{for } \alpha \neq 2 \,. \end{cases}$$

Confidence intervals for $M = \min f$ and testing statistical hypotheses about M

The following confidence interval for M has asymptotic (as $N \to \infty$) confidence level $1 - \delta$:

$$[y_{1,N} - (y_{k,N} - y_{1,N})/c_{k,\delta}, y_{1,N}], \quad \text{where} \quad c_{k,\delta} = \left[1 - (1-\delta)^{1/k}\right]^{-1/\alpha} - 1\,.$$

Procedures of testing hypotheses about M are based on constructing confidence intervals for M. If we want to test the hypothesis $H : M \leqq f_*$ then we construct a c.i. and if f_* belongs to this c.i., then the hypothesis H gets accepted.
These procedures can be used for:

- devising stopping rules in global random search
- branch and probability bound methods

Branch and (probability) bound methods

- branching of the optimization set into a tree of subsets (e.g by a triangulation),
- making (probabilistic) decisions about the prospectiveness of the subsets for further search, and
- selection of the subsets that are recognized as prospective for further branching.

Tail index α

$$F(t) = \Pr\{x \in A : \ f(x) \leqslant t\} = c_0(t - M)^{\alpha} + \mathrm{o}((t - M)^{\alpha}) \quad \text{as } t \downarrow M\,.$$

- Estimation of α is difficult and non-conclusive.
- If $f(x)$ is locally linear close to x^* (when x^* lies on the boundary of A) then $\alpha = d$.
- If $f(x)$ is locally quadratic around x^* then $\alpha = d/2$.

Curse of dimensionality: Precision of statistical inferences as a function of α

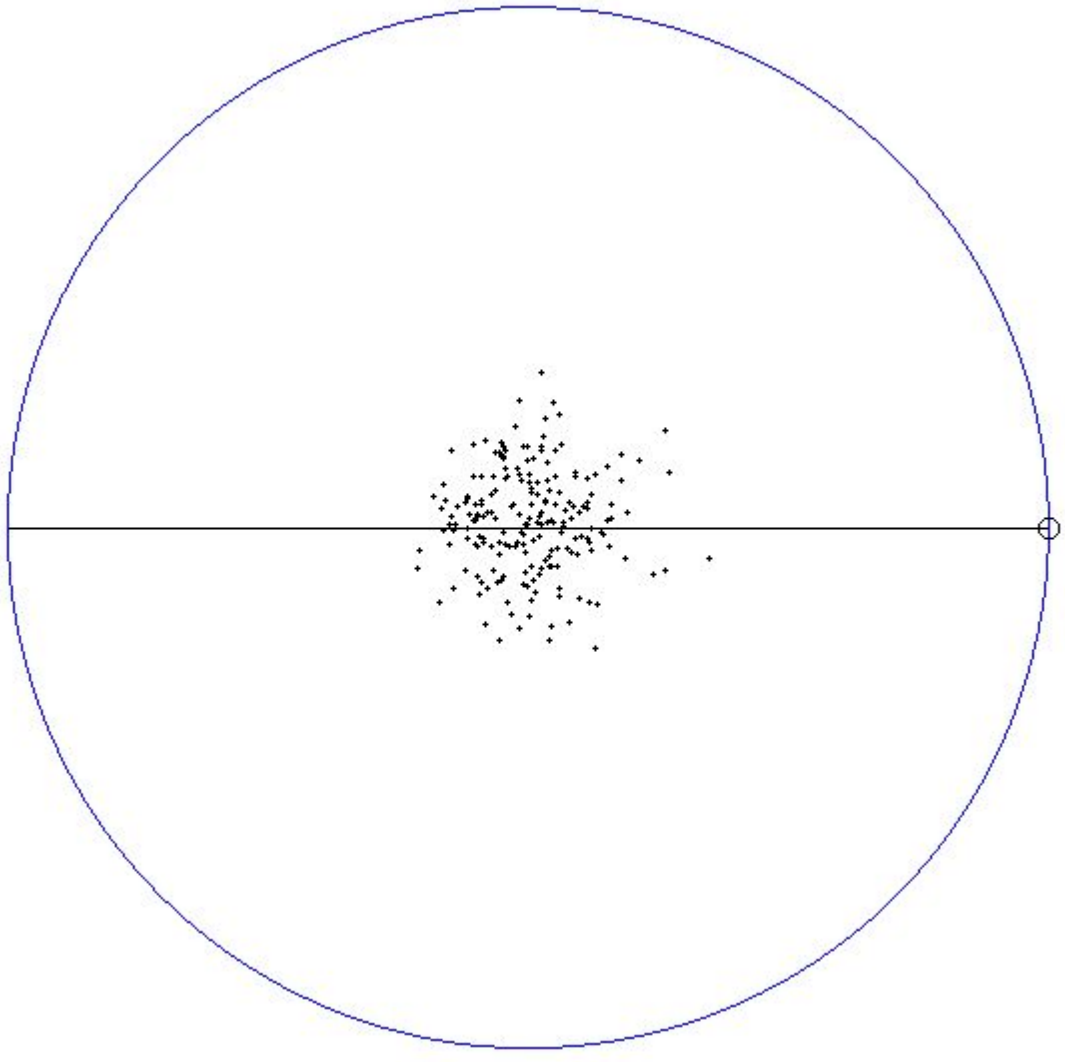
MSE of the (optimal) estimators has order

$$cN^{-2/\alpha} \quad as \quad N \to \infty\,.$$

Example: $c = 1$, $\alpha = d$. To achieve precision (value of MSE) 0.01 the sample size N should be

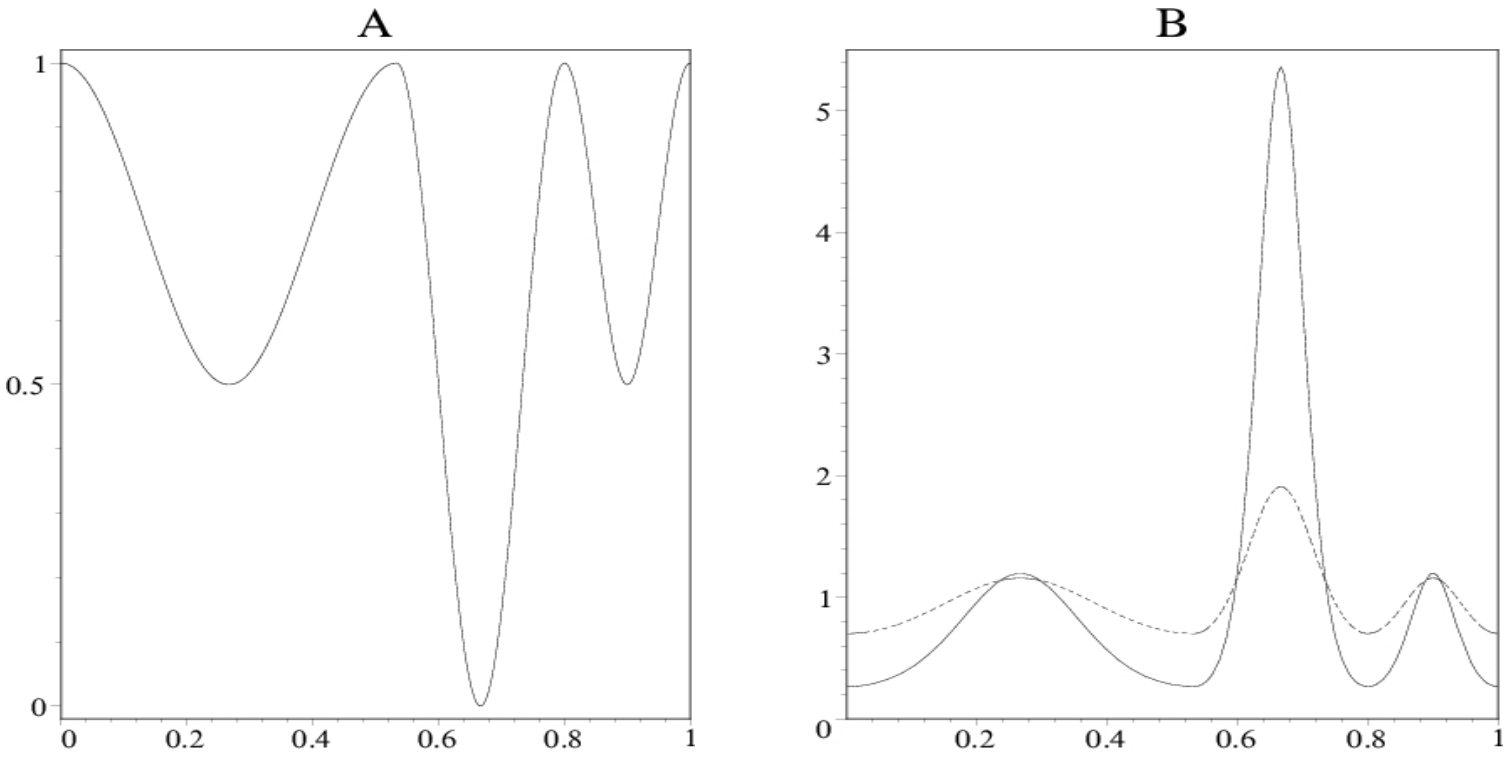
$$N = 10^d\,.$$

Random points in a ball; projection to 2 and 1 dimensions



Probabilistic models: Gibbs densities

$$\pi_\beta(x) = \exp\{-\beta f(x)\} \Big/ \int_A \exp\{-\beta f(z)\} dz \ .$$



(A) Graph of the objective function f;
(B) Gibbs densities with $\beta = 1$ (dotted line) and $\beta = 3$ (solid line)

Evolutionary (population-based) methods

Parent generation:

$$x_1^{(j)}, \ldots, x_{n_j}^{(j)}$$

Generation of descendants (children):

$$x_1^{(j+1)}, \ldots, x_{n_{j+1}}^{(j+1)}$$

Population-based methods are defined by:

(a) the stopping rule,

(b) the rules for computing the numbers n_j (population sizes), and

(c) the rules for obtaining the population of descendants from the population of parents.

Stopping rules

1. total number of points generated
2. Using Liptchitz-type conditions
3. Using statistical procedures

Population sizes

1. n_j depend on the statistical information gathered during the search;
2. the sequence of n_j is non-increasing: $n_1 \geqslant n_2 \geqslant \ldots \geqslant n_j \geqslant \ldots$;
3. $n_j = n$ for all j;
4. the sequence of n_j is non-decreasing: $n_1 \leqslant n_2 \leqslant \ldots \leqslant n_j \leqslant \ldots$

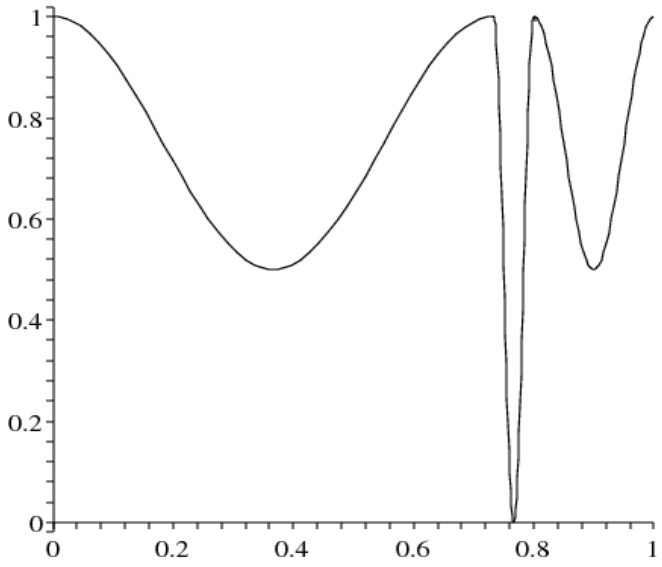
Choice of updating probability laws

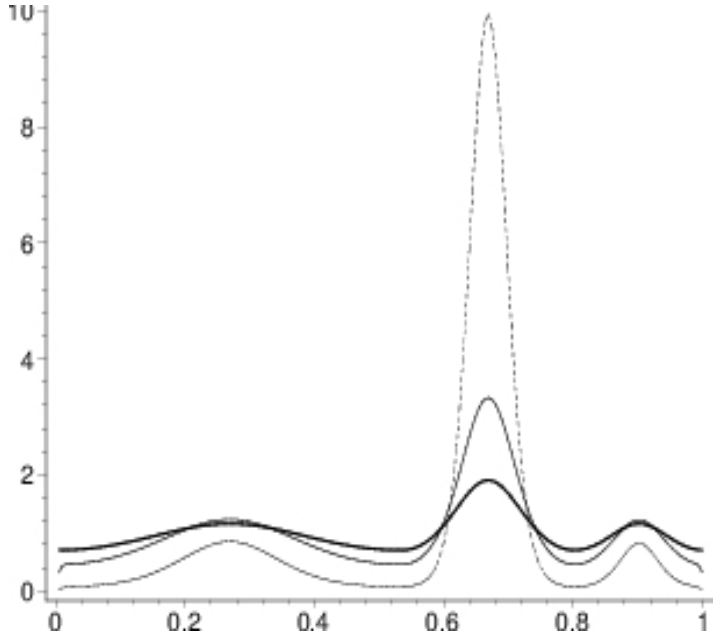
$$R_{j+1}(dx) = \left[\int R_j(dz) f_j(z)\right]^{-1} \int R_j(dz) f_j(z) Q_j(z, dx).$$

$$R(dx) = \left[\int g(z) R(dz)\right]^{-1} \int R(dz) g(z) Q(z, dx),$$

where, for example, $g(x) = \exp\{-\beta f(x)\}$

Limiting behaviour of the population distributions





Similar to Gibb's densities but the limiting measures do not have to attract to the neighbourhood of the global minimizer

Thank you and references

Thank you for attention

References

1. Zhigljavsky, A. and Zilinskas A. (2008). Stochastic Global Optimization, Springer
2. Zhigljavsky, A. (1991) Theory of Global Random Search, Kluwer
3. Many other books