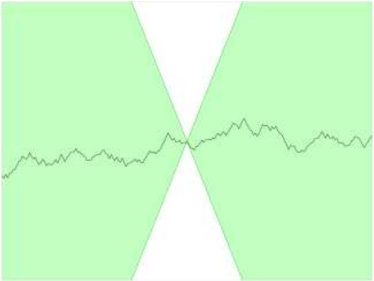
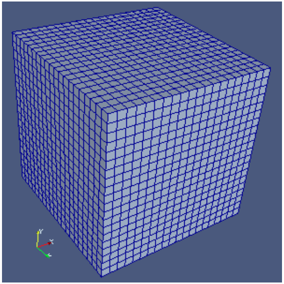
Curse of Dimensionality

Bellman's (1961) phrase concerning exhaustive enumeration on product spaces

Lipschitz-continuous function



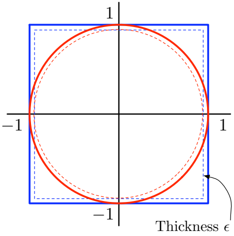
Cartesian grid



Wikipedia

Optimizing, interpolating, or integrating a smooth d-D function to error ϵ requires $O(1/\epsilon^d)$ evaluations.

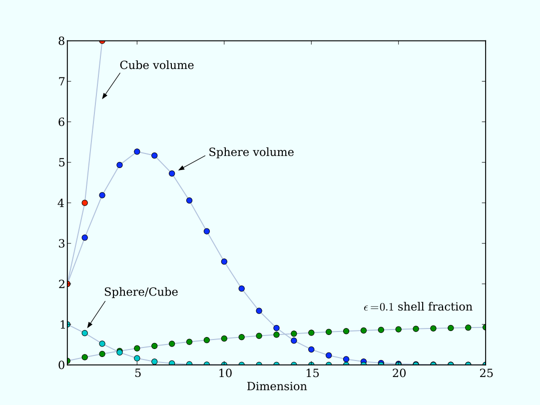
The “Excluded Middle”



$V_\square = 2^d$

$$V_\circ = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$$

$$\text{Shell fraction} = 1 - (1-\epsilon)^d$$



- Hi-D spheres have volume quickly decreasing with *d*
- Spherical core of a hypercube has negligible volume
- Volume in a simple *d*-D region is mostly near the boundary

Uniform Distribution in Hi-D

Consider a large sample of points from $U[0,1]^d$.

$\langle \#$ pts in volume $\delta V \rangle \propto \delta V \rightarrow$ volume effects map over

Empty space phenomenon (Scott & Thompson 1983)

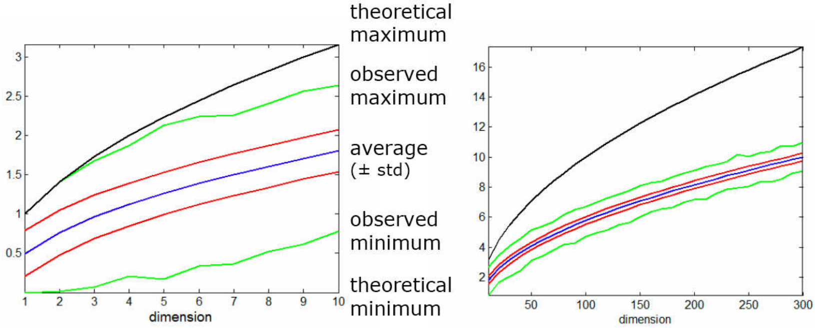
- Most cells in a grid will be empty even for large samples
- Most points are near boundaries: Most points appear extreme/surprising in some respect
- Points are all near a $(d-1)$-D manifold
- Spherical neighborhoods of a point will be nearly empty

Concentration of Euclidean norm

$$
\begin{aligned}
r^2 &= \sum_{i=1}^{d} x_i^2; \qquad x_i^2 \text{ has mean } 1/3, \text{ variance } 4/45 \\
&\approx d \times \text{mean of } d \text{ draws from } N(1/3, 4/45) \\
&\sim d \times \text{draw from } N(1/3, 4/(45 \cdot d))
\end{aligned}
$$

$\rightarrow$ r concentrates near $\sqrt{d/3}$ with *constant* variance

Average norms of 10^4 draws from $U[0,1]^d$



Damien Francois (2005)

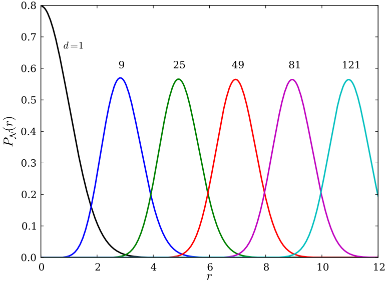
Standard Normal Distribution in Hi-D

Normal distribution has infinite range, with high-density region is localized near origin

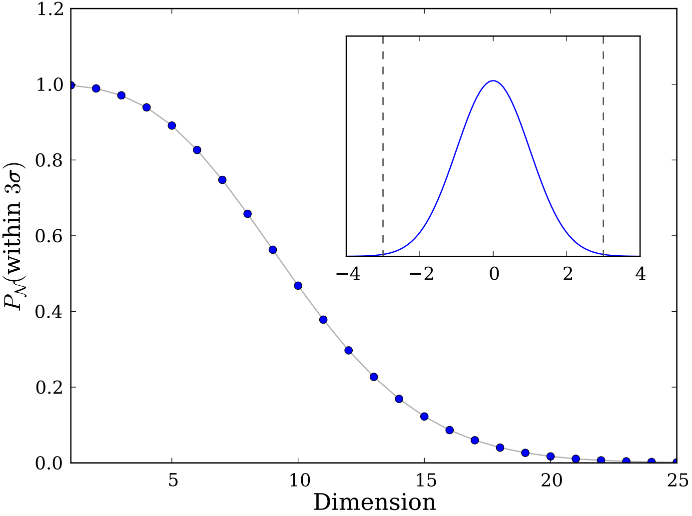
Basic facts:

- Squared radius $\sum_{i=1}^{d} x_i^2$ is χ_d^2
- $\langle \chi_d^2 \rangle = d$; std dev'n $= \sqrt{2d}$

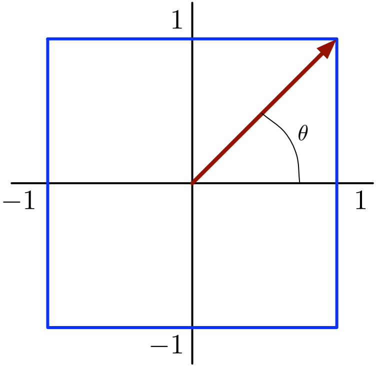
Most points are in thin shells



Most points are in the tails



Null Projection Phenomenon



Diagnoal vector $\vec{v} = [\pm 1, \pm 1, \ldots \pm 1]$

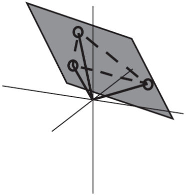
$$\cos\theta = \frac{\vec{v} \cdot \vec{e}_j}{\sqrt{\vec{v} \cdot \vec{v} \; \vec{e}_j \cdot \vec{e}_j}} = \frac{\pm 1}{\sqrt{d}}$$

- Diagonals become nearly orthogonal to all axes
- Clusters near diagonal all get mapped near the origin & may overlap in pairwise scatterplots
- Choice of coordinate system is important for finding structures

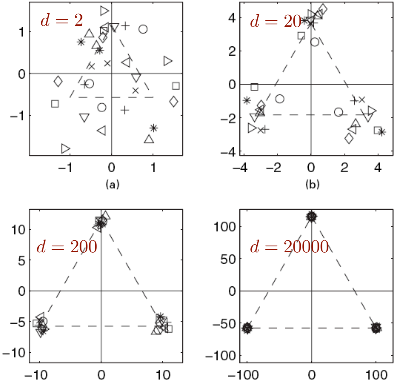
Sets of points lie on unit simplex vertices (Hall+ 2005)

δ-method applied to $\arccos(\mathbf{x}_1 \cdot \mathbf{x}_2) \rightarrow$

$$\theta_{12} = \frac{\pi}{2} + O(1/\sqrt{d}); \qquad \text{pairwise angles} \approx \perp$$



Hall et al. (2005)

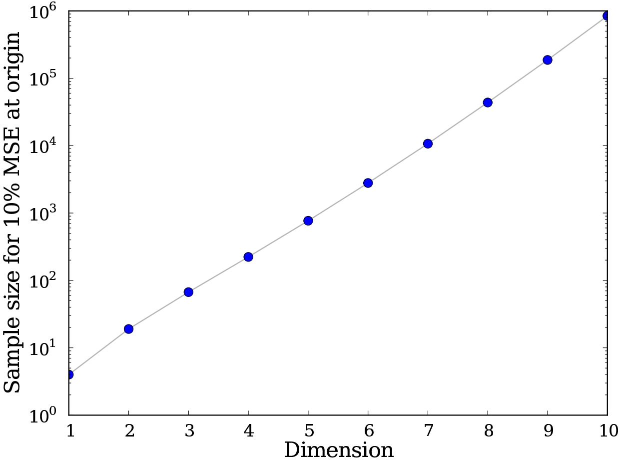


(a) (b) (c) (d)

- Norms and angles between samples become nearly *deterministic*
- Randomness is in rotation of the simplex

Curse of Dimensionality for KDE

Estimate a normal density at the origin to 10% using Gaussian-kernel KDE with optimal smoothing.



Silverman (1986)

Concentration of Measure

Both the uniform and normal settings exhibited Gaussian-like concentration into small volumes.

How generic is this? *Very!*

For random vector with IID components with 8 finite moments:

$$\begin{aligned} E(|\vec{x}|) &= \sqrt{ad-b} + O(1/d) \\ Var(|\vec{x}|) &= b + O(1/\sqrt{d}) \end{aligned}$$

Constants a, b depend on 1st 4 moments

- Norm grows like $\sqrt{d}$ but variance $\approx$ const.
- If you contain a region of the space with a substantial fraction of probability, a small expansion includes nearly all of it
- Smooth functions of d random variables become approximately constant for large d