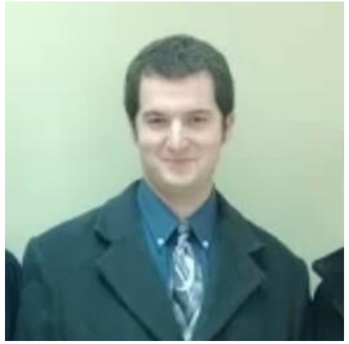
Efficiency of the floating body as a robust measure of dispersion

Luis Rademacher

UC Davis, March 2019

Joint work with Joseph Anderson, Navin Goyal, Anupama Nandi

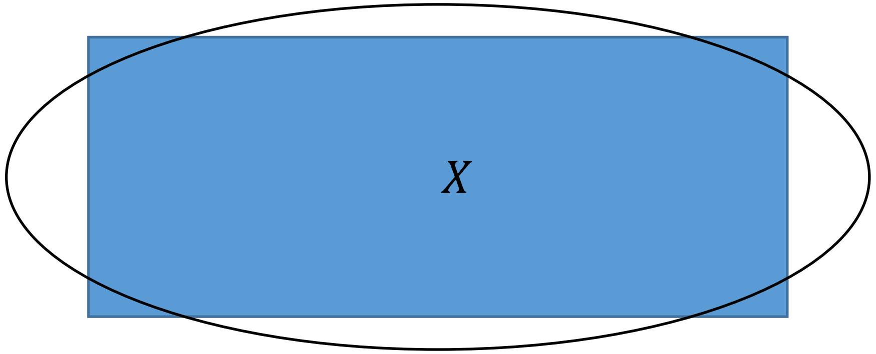


Location/shape of data/distribution

- Data $X^{(1)}, X^{(2)}, \ldots, X^{(n)} \in R^d$.
- Can be interpreted as iid samples from a distribution or random vector (model) X.
- Location:
  - Mean $\mu = E(X)$
- "Shape":
  - Covariance matrix $\Sigma = \operatorname{cov}(X) \ = \ E((X-\mu)(X-\mu)^T)$
- Same for data via empirical distribution (uniform distribution on $\{X^{(1)}, X^{(2)}, \ldots, X^{(n)}\}$).

Legendre’s ellipsoid of inertia

- How is covariance = shape?
- Consider Legendre’s ellipsoid of inertia of X: unique ellipsoid with the same covariance matrix as X.



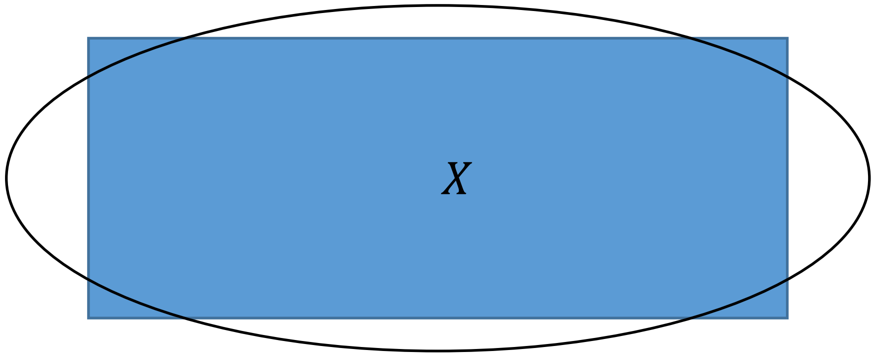
The only surviving portrait of Legendre

Depth, distance of point to data/distribution

- Mahalanobis distance is norm induced by ellipsoid of inertia.

$$d(p, X) = \sqrt{p^T \operatorname{cov}(X)^{-1} p}$$

- Mahalanobis depth: $\operatorname{depth}(p, X) = \frac{1}{1+d(p,X)^2}$



Challenge.

- $\mathrm{cov}(X)$ is frequently used in algorithmic statistical analysis of data.
- What if data/distribution is heavy-tailed
  (some moments are undefined)?
  What if data seems to follow a distribution with infinite second moment?
  (kth moment of r.v. X is $E\left(X^k\right)$)

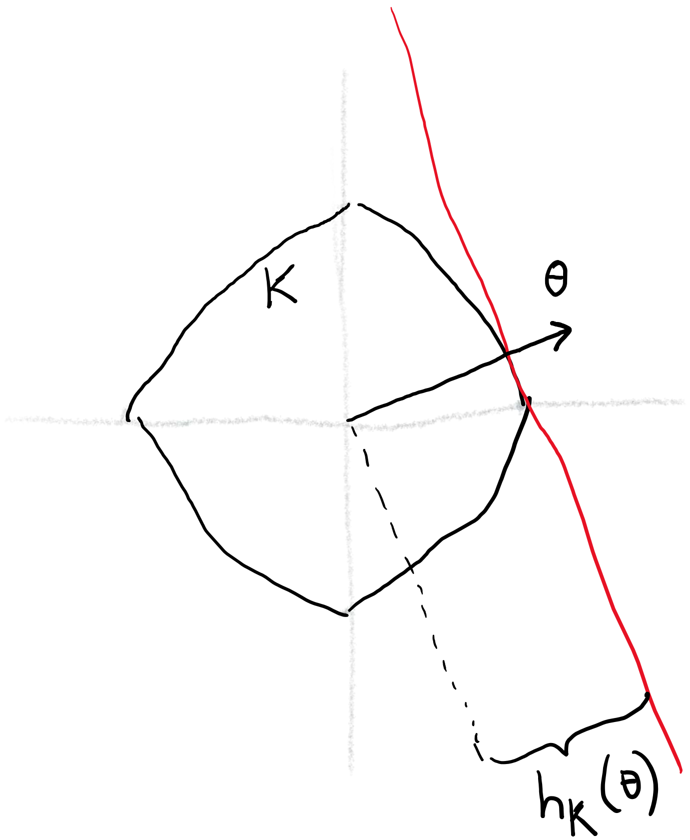
Towards more general shapes

- Support function of a convex body K:

$$h_K(\theta) = \sup_{x \in K} x^T \theta$$

- Support function of Legendre's ellipsoid ($E(X) = 0$ case):

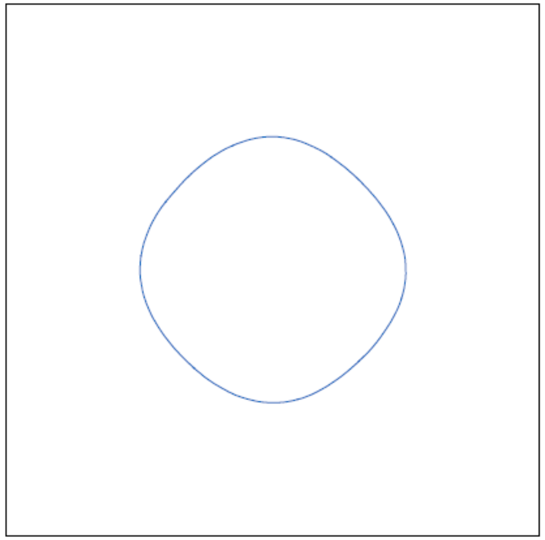
$$h(\theta) = c\sqrt{E((X^T\theta)^2)} = c\sqrt{\theta^T \mathrm{cov}(X)\theta}$$



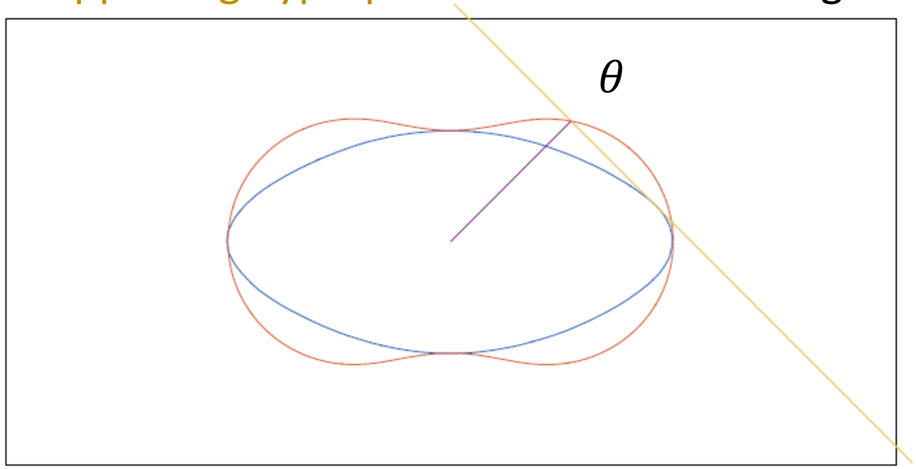
Shape if 2$^{\text{nd}}$ moment is infinite?
Centroid body

- **Definition** (Petty 1961):
  Given random vector X, the centroid body of X, denoted ΓX, is the convex body with support function
  $$h_{\Gamma X}(\theta) = E(|X^T\theta|).$$

centroid body of $[-1,1]^2$



centroid body, support function and supporting hyperplane at 45$^{\text{o}}$ of a rectangle



Shape if 2nd moment is infinite?
Centroid body

- It is not obvious that

$$h_{\Gamma X}(\theta) = E(|X^T \theta|)$$

  is the support function of a convex body. But it is obvious given:
  Thm: $f: R^n \to R$ is a support function iff f is **convex** and **positively homogeneous**.

Without first moment?

- A shape that works for any distribution?

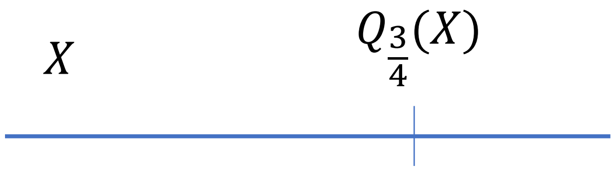
Quantiles

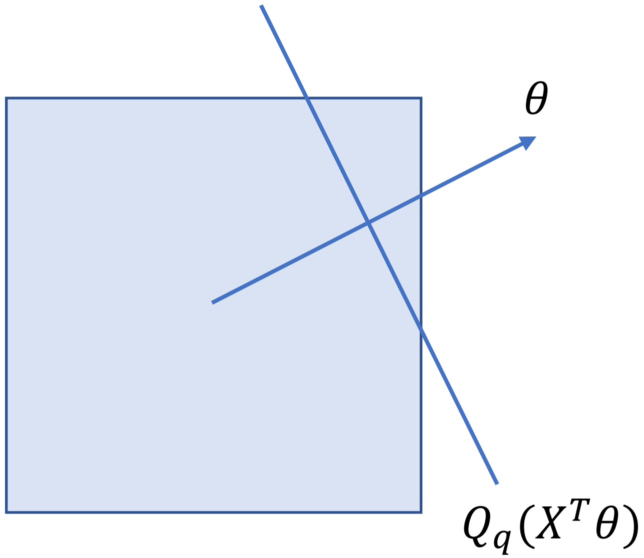
- Median and quantiles are robust against noise, outliers, heavy-tails.
- Multi-dimensional analogue?
- Start with **quantile function**:

$$Q_q(X) = F_X^{-1}(q) \coloneqq \inf\{t\colon P(X \le t) \ge q\}.$$

- **Directional quantile function**:

$$\theta \mapsto Q_q(X^T\theta)$$





Depth region – Convex Floating body

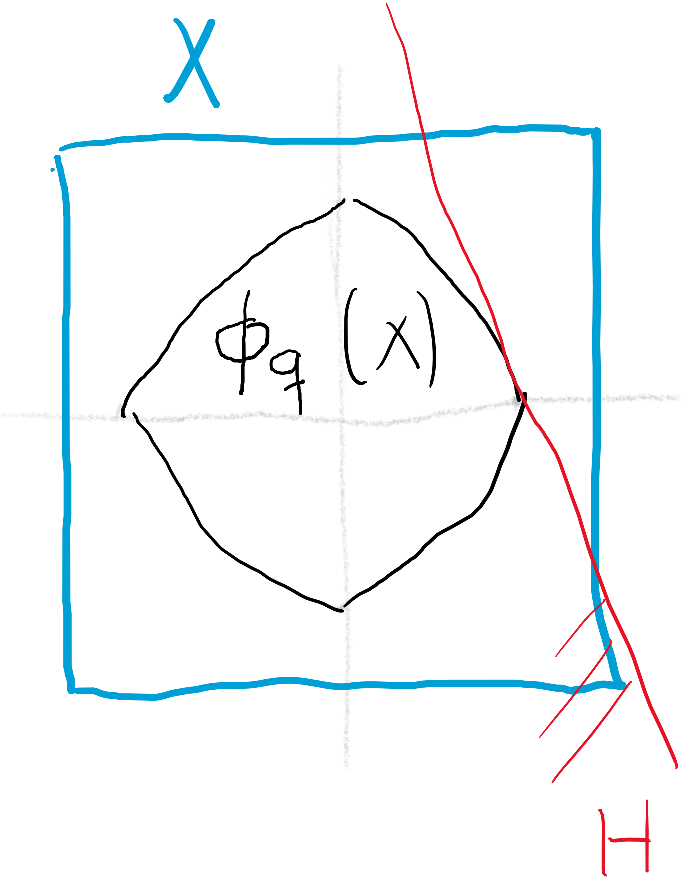
- [Tukey '75] (Halfspace) $\mathrm{depth}(x) = \inf\{P(H): x \in H \text{ halfspace}\}$
- Depth region
  [Tukey '75] [Donoho Gasko '92]

$$\bigcap_{H \text{ halfspace},\, P(H)>q} H$$
$$= \{x: \mathrm{depth}(x) \geq 1 - q\}$$

- (Convex) Floating body
  [Dupin 1822] [Schütt Werner '90] [Bárány Larman '88]

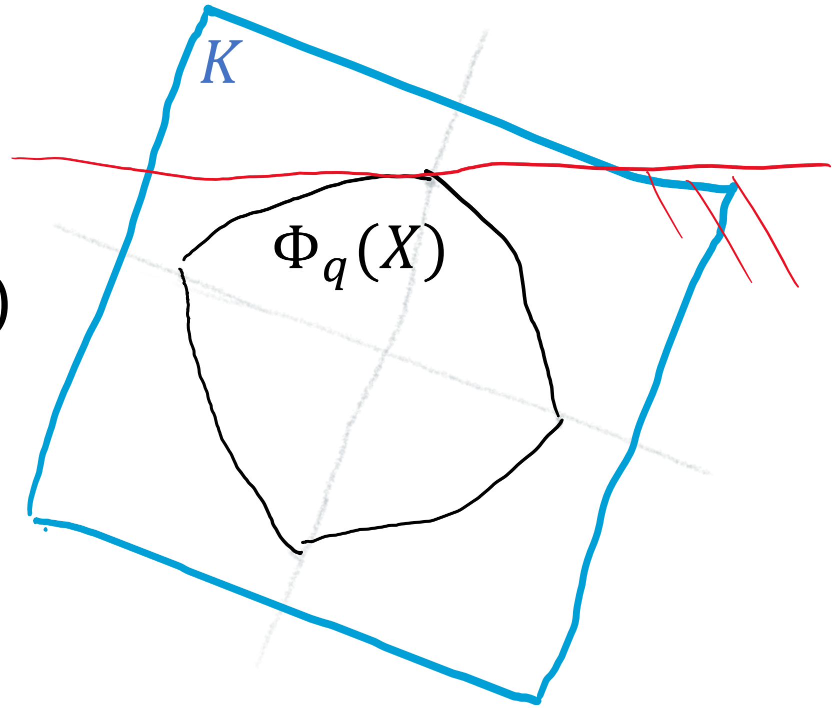
$$\Phi_q X = \bigcap_{H \text{ halfspace},\, P(H)\geq q} H$$
$$= \{x: (\forall\theta) x^T\theta \leq Q_q(X^T\theta)\}$$

- Floating body = depth region when $\mathrm{supp}(X)$ is connected.



Why "floating"?
From mechanics, hydrostatics [Dupin 1822]

- Motivation: **Archimedes principle** implies submerged part of a uniform body floating in a fluid is the same fraction in every orientation.
- For X uniform in K, convex floating body $\Phi_q(X)$ is the set of points that are submerged in every orientation (for q determined by density of K and fluid).



Importance of floating body/depth curves

- Given a dataset or distribution
  - Estimator of :
    - Shape
    - Dispersion
    - Depth of a point
  - Robust
    - Defined for any distribution, even heavy-tailed
    - No moment assumption
  - Difficulty: more complex than an ellipsoid (covariance matrix).

Issue-question: **computational efficiency**

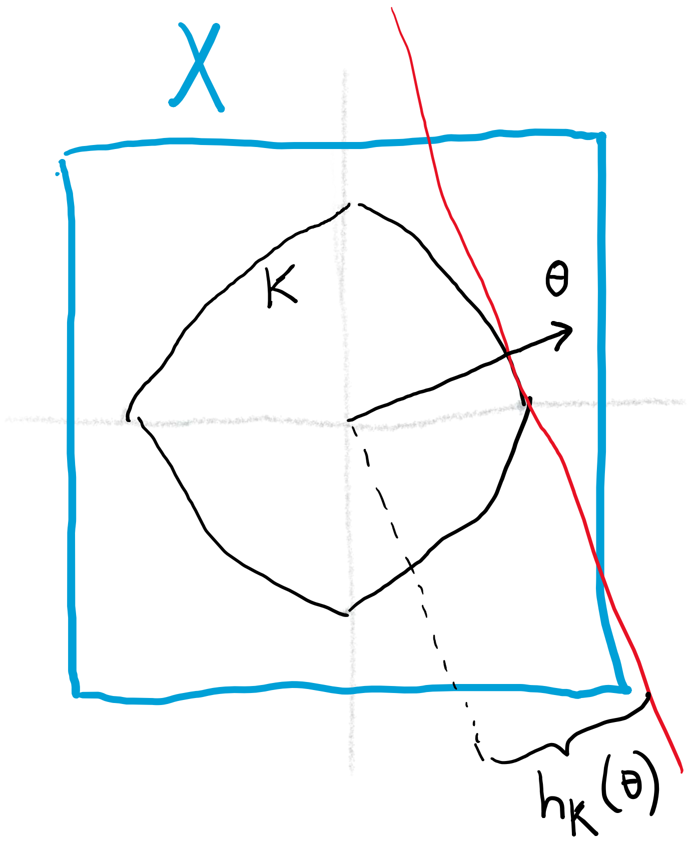
- Can we answer questions about inertia ellipsoid?
  - Yes, given estimate of $\mathrm{cov}(X)$: depth, distance, membership.
- Can we answer questions about centroid body or floating body?
  - Not clear.
  - Basic question (depth and distance reduce to it):
    Membership: given point $x \in R^d$ and level q, is x contained in floating body $\Phi_q X = \bigcap_{H \text{ halfspace}, P(H) \geq q} H$?
- Difficulty: representation as intersection of an **infinite** family of half-spaces.

Issue-question: **computational efficiency**

- Determining membership of a point in depth region (of a finite dataset) is coNP-complete [Johnson Preparata '78].
- Many existing works on exact computation, fixed dimension, hardness of approximation, approximation algorithms.
- How to cope?
  - Support function of a convex body K:

$$h_K(\theta) = \sup_{x \in K} x^T \theta$$

  - If one has efficient evaluation of **support function** of convex body then one can decide membership efficiently.
    - Proof: **ellipsoid algorithm**
    - Also true with approximate support function and approximate membership.



Ellipsoid algorithm

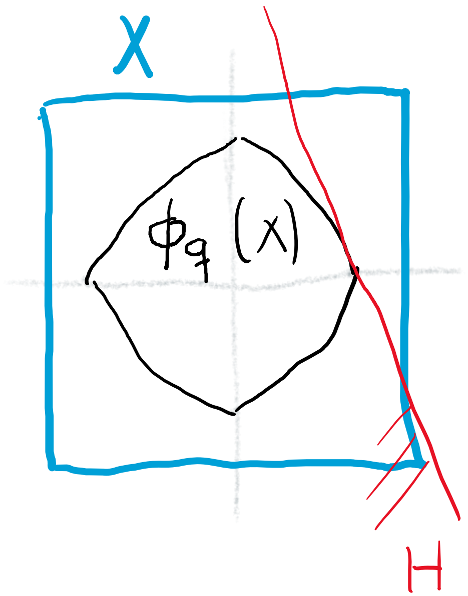
- Efficient algorithm to solve the following problem:
  - Given
    - A point x
    - evaluation access to the support function h_K of a convex body K
  - Determine whether $x \in K$.

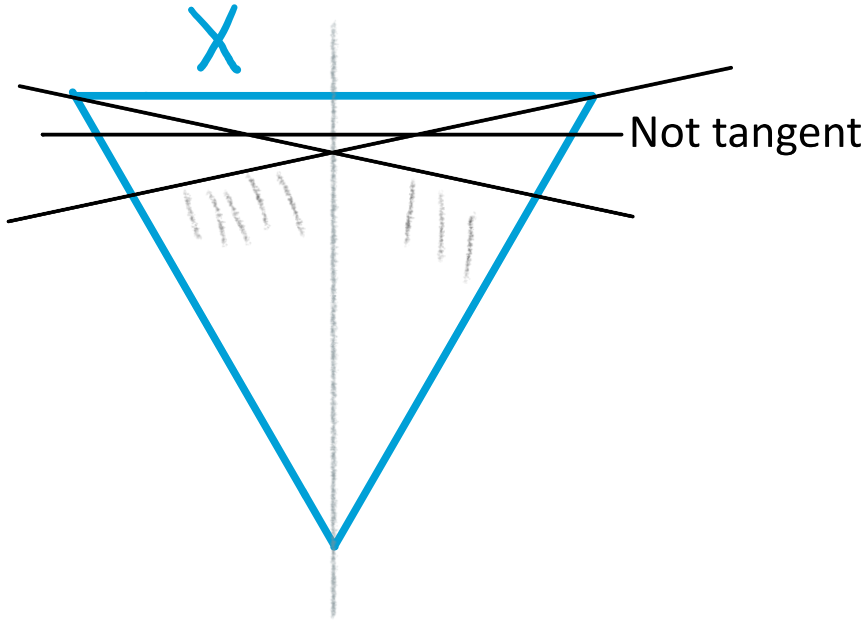
Is quantile function the support?

$\Leftrightarrow$Is every q-quantile hyperplane tangent to $\Phi_q X$?

$\Leftrightarrow h_{\Phi_q X}(\theta) \stackrel{?}{=} Q_q(\theta)$.

- Example: X uniform in square: YES.
- Example: X uniform in triangle: NO.
- Example: X discrete: NO.

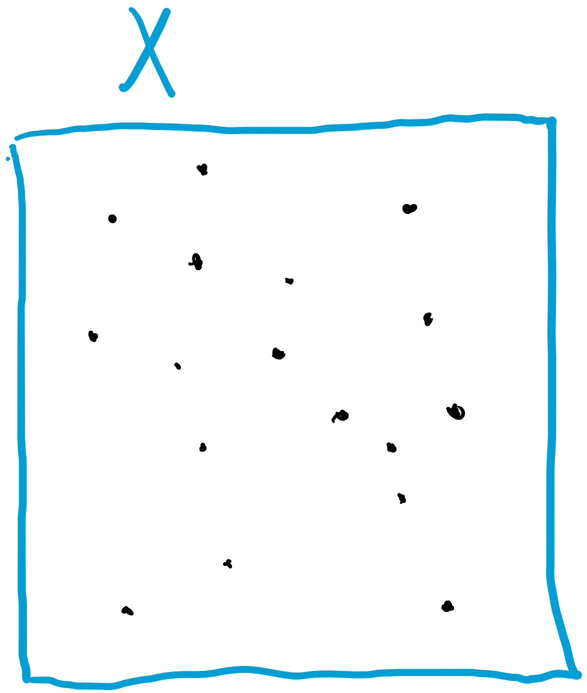




Quantile function = support function sometimes:

1. Symmetric log-concave distribution [Meyer Reisner '91] [Ball]. Generalized by [Bobkov '10]
2. Product distribution with symmetric α-stable coordinates, $\alpha \geq 1$.
   - Floating body $\Phi_q(X)$ is scaled l_β ball, $1 = \frac{1}{\alpha} + \frac{1}{\beta}$.

- Preserved under affine transformation $\Rightarrow$Affine transformations of (2.).
- Datasets that are iid samples of any of the above distributions (approximately, via uniform convergence of empirical distribution).



Log-concave distribution

- A distribution with density f so that $\log f$ is concave.
- Includes:
  - Gaussian
  - Uniform distribution in a convex body.
  - Exponential
  - Dirichlet …

Stable distribution

- CLT (informal): distribution of normalized sum of iid random variables with **finite second moment** converges to **Gaussian distribution**.
- Generalized CLT (informal): distribution of normalized sum of iid random variables converges to some **stable distribution**.
- Formally, symmetric α-stable distribution is distribution with characteristic function $\varphi(t) = e^{-|t|^{\alpha}}$, $\alpha \in (0,2]$.
  - 2-stable is Gaussian
  - 1-stable is Cauchy
  - α-stable with $\alpha < 2$ is heavy-tailed: no moments of order $\geq \alpha$.

Our results

- Sample and time bounds for efficient membership in floating body in cases 1. ("log-concave") and 2. ("stable").
  - Proof idea:
    - Quantile estimation error bounds +
    - ellipsoid algorithm +
    - Vapnik–Chervonenkis theory (uniform convergence of empirical distribution)
- Application: Provably efficient ICA (independent component analysis) with components that are symmetric α-stable for $\alpha \geq 1$.

Our results

- Approximate geometry of product distribution with power-law distributed coordinates via GCLT:
  - **Thm:** X symmetric r.v. with indep. coordinates with tails $1 - F(x) \approx \frac{1}{x}$.
    $S_k =$ sum of k iid copies of X. Then
    - Floating body of S_k is close to floating body of product of Cauchy (=1-stable, hypercube).
    - Proof: Generalized CLT with rate.
- Application: Provably efficient ICA with components that are symmetric power-law distributions (even with infinite first moment).

Proof idea for log-concave case

- **Thm:** X symmetric log-concave r.v. On input x, q, ϵ, δ, can ϵ-weak decide whether $x \in \Phi_q(X)$ in time $poly(d, \frac{1}{1-q}, \frac{1}{\sigma_{min}(\Sigma)}, \sigma_{max}(\Sigma), \frac{1}{\epsilon}, \log\frac{1}{\delta})$.
  - VC-theory implies: Let X be random vector, let Y follow empirical distribution of sample $X^{(1)}, X^{(2)}, \dots, X^{(N)}$. Then for $N \geq \frac{c}{\epsilon^2}\left(d \log\frac{d}{\epsilon} + \log\frac{1}{\delta}\right)$:

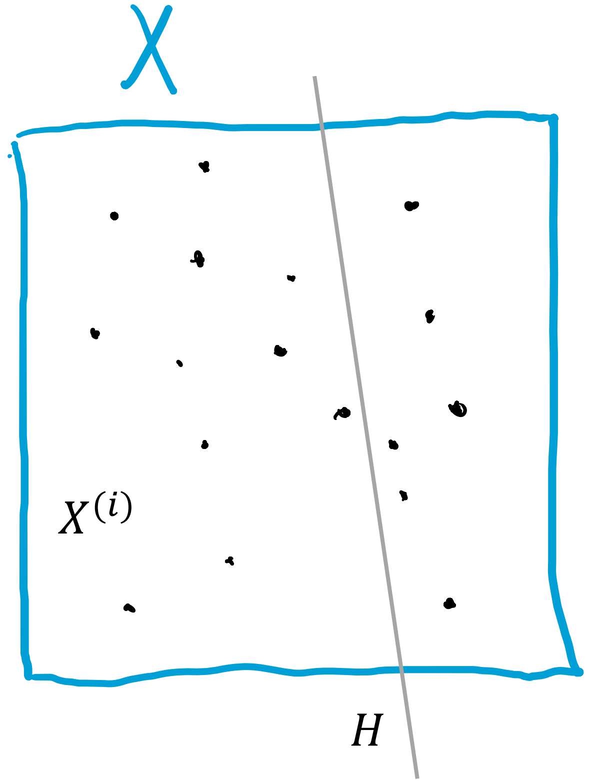
$$P\left(\sup_{H \text{ halfspace}} |\mu_Y(H) - \mu_X(H)| < \epsilon\right) \geq 1 - \delta.$$

  - This implies quantile function $Q_q(Y)$ of sample close to $Q_q(X)$: For

$$N \geq \frac{c}{\epsilon^2(1-q)^2}\left(d \log\frac{d}{\epsilon(1-q)} + \log\frac{1}{\delta}\right)$$

and X symmetric logconcave with $\mathrm{cov}(X) = I$ we have

$$P\left(\sup_{\theta \in S^{d-1}} |Q_q(Y^T\theta) - Q_q(X^T\theta)| \leq \epsilon\right) \geq 1 - \delta.$$



$$\mu_Y(H) = P(Y \in H) = \frac{\#\{X^{(1)}, X^{(2)}, \dots, X^{(N)}\} \cap H}{N}$$

Conclusion

- If dataset follows “good” distribution, then halfspace depth and membership in floating body is efficient.
  - If dataset is just “a set of points”, then depth is NP-hard
  - If dataset is a sample from a symmetric logconcave distribution (say), then approximate depth is efficient (whp).
- Open questions:
  - For which distributions is quantile function $Q_q(\theta)$ the support function of floating body $\Phi_q X$?
  - Ellipsoid algorithm is not practical. Practically efficient algorithm?