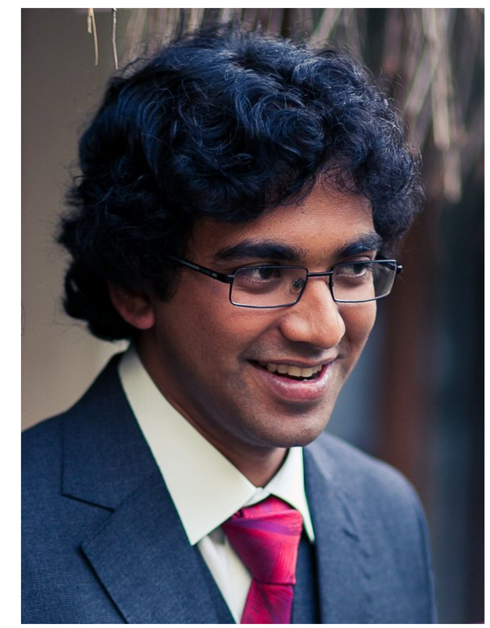
# Estimating High-dimensional Confidence Sets: A Robust Estimation Perspective


**Vaidehi Srinivas**

Northwestern University, Computer Science

**Chao Gao**

University of Chicago, Statistics

**Liren Shan**

Toyota Technological Institute, Chicago

**Aravindan Vijayaraghavan**

Northwestern University, Computer Science


COLT 2025 Workshop on Predictions and Uncertainty

# High-dimensional Confidence Sets

**Goal:** Find a high-density region of an **arbitrary** distribution

**Problem:** Given samples drawn i.i.d. from an unknown distribution $\mathscr{D}$ over $\mathbb{R}^d$, and target **coverage rate** $\delta$, fit the **smallest volume** confidence set $S$ such that

$$\mathbb{P}_{y\sim\mathscr{D}}(y \in S) \geq \delta\,.$$

(Think of $\delta = 0.9$)

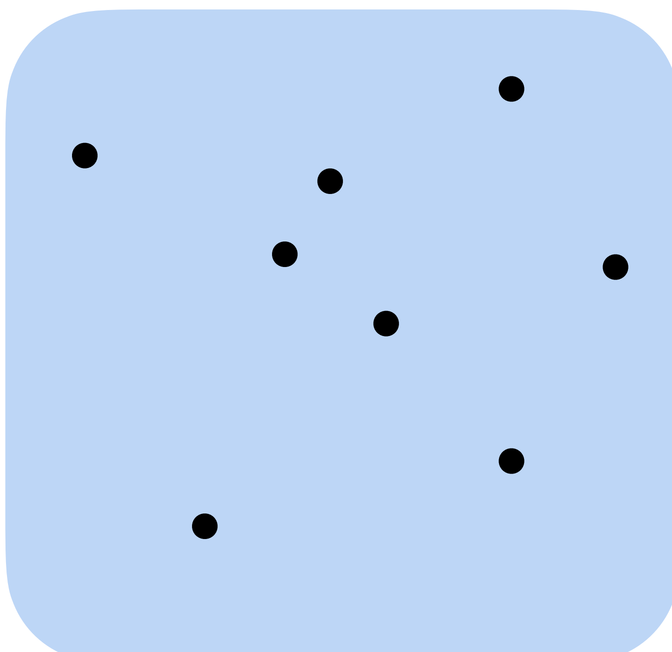
**"Volume optimality"**

**Applications:** Central problem in statistics

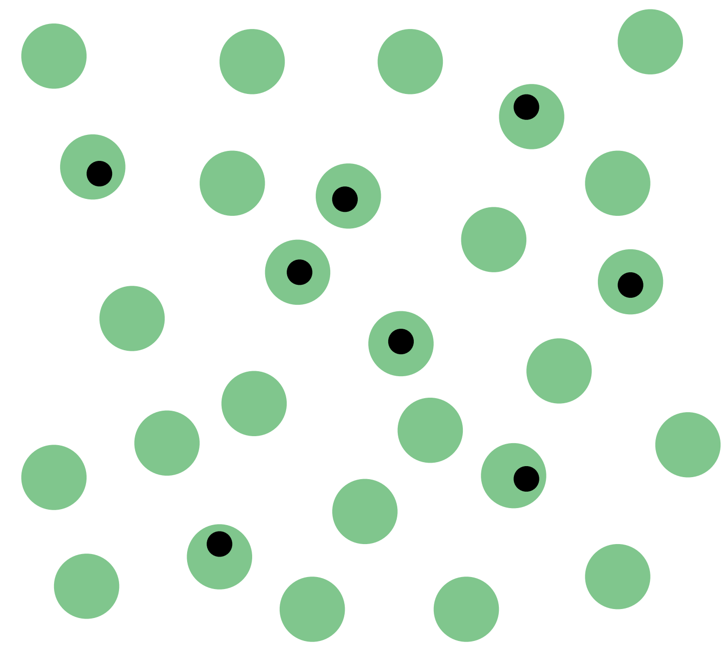
- Conformal prediction [Gao Shan **S.** Vijayaraghavan '25],
- Estimating density level sets [Garcia Kutalik Cho Wolkenhauer '03],
- Support estimation [Schölkopf Platt Shawe-Taylor Smola Williamson '01],
- Robust estimation [Rousseeuw '84, 85],
- and many more!

# Trouble with Volume Optimality

**Bad news:** Impossible to achieve volume optimality in general!



$\mathscr{D}_1$
does not admit a small confidence set



$\mathscr{D}_2$
admits a small confidence set

Cannot distinguish between $\mathscr{D}_1$ and $\mathscr{D}_2$

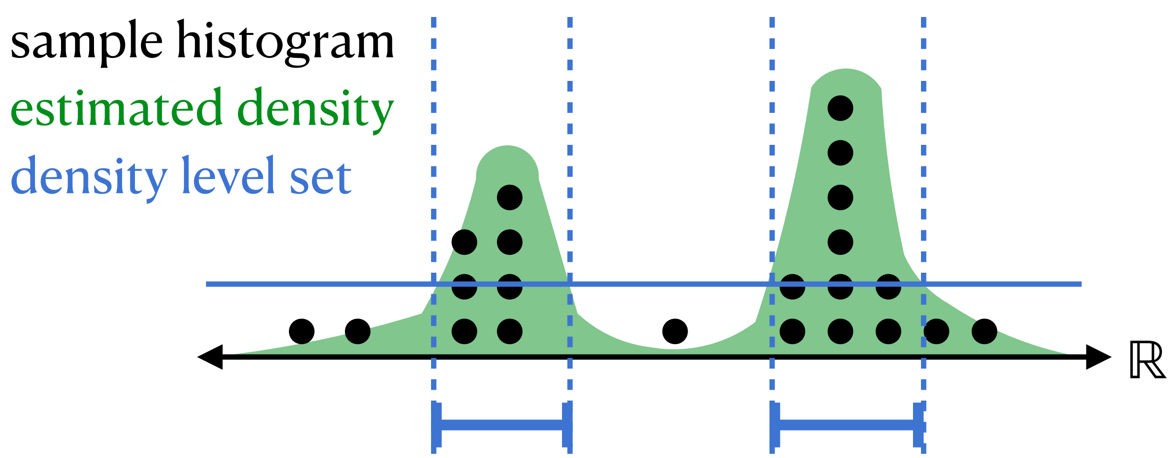
Forced to lose in either **coverage** or **volume**

(even in one dimension!)

# Restricted Volume Optimality

**Strategy 1:** Restrict the class of distributions

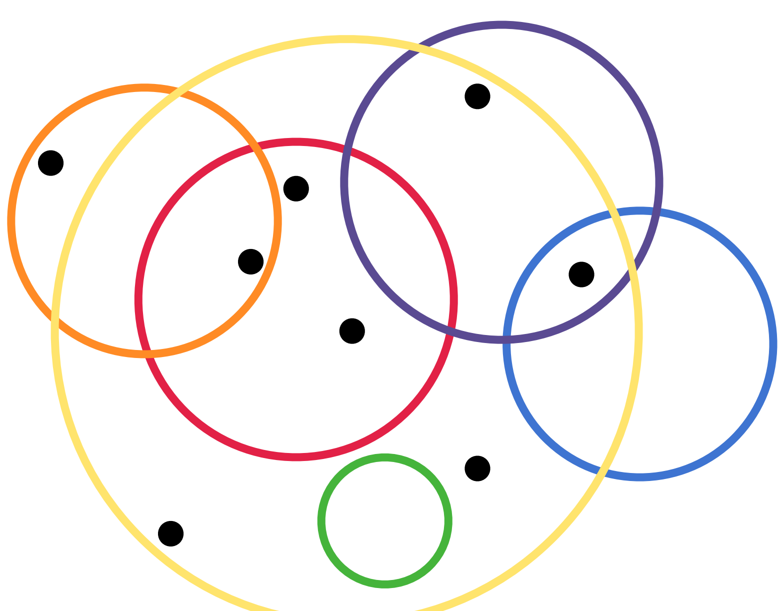
**Example:** density estimation [Lei, Robins, Wasserman '13], [Izbicki, Shimizu, Stern '22]




- Works for nicely-behaved distributions in one dimension
- Statistically intractable in high-dimensions

**Strategy 2:** Restrict the class of confidence sets [Scott Nowak '05][Gao, Shan, **S.**, Vijayaraghavan '25]

**Examples:** confidence intervals over $\mathbb{R}$, Euclidean balls over $\mathbb{R}^d$



- Set families of bounded VC-dimension exhibit **uniform convergence**
- Coverage of all balls simultaneously converge in poly($d$) samples $\Rightarrow$ **statistically tractable**!
- Aim to compete with best set in the family

# Learning a Confidence Set

**Focus:** $\mathscr{S} = \mathscr{B}$ the set of Euclidean balls in $\mathbb{R}^d$.

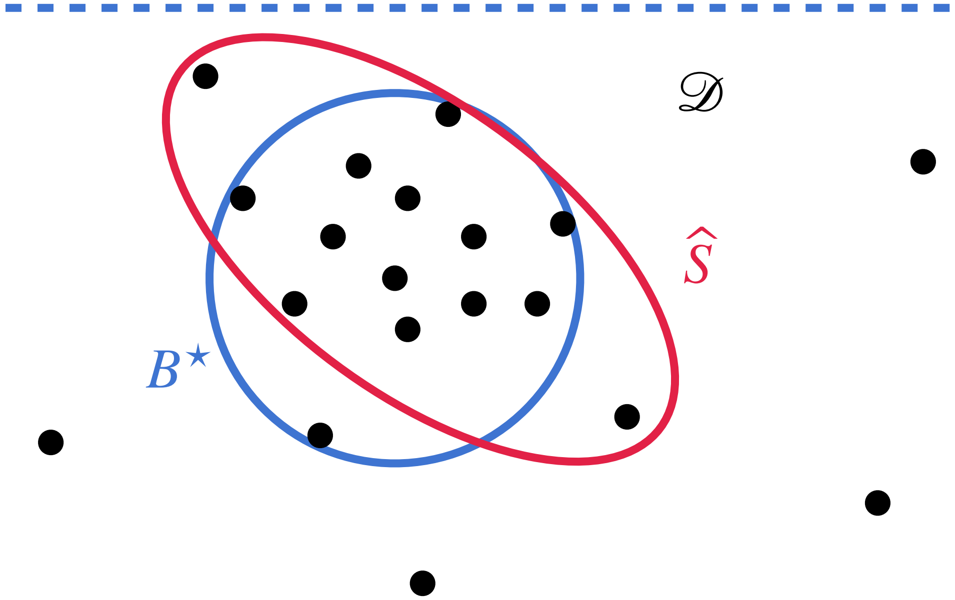
**Updated problem:** Let $B^\star \in \mathscr{B}$ be the minimum volume set with coverage $\geq \delta$ over $\mathscr{D}$. Find a set $\widehat{S}$ with coverage $\geq \delta$ over $\mathscr{D}$, and

$$\mathrm{vol}(\widehat{S}) \lesssim \mathrm{vol}(B^\star).$$

(also natural to compare "radii": $\mathrm{vol}(\widehat{S})^{1/d} \lesssim \mathrm{vol}(B^\star)^{1/d}$)

→ $\widehat{S}$ does not necessarily need to be in $\mathscr{B}$, similar to PAC learning ($\widehat{S} \in \mathscr{B}$ is **proper learning**)

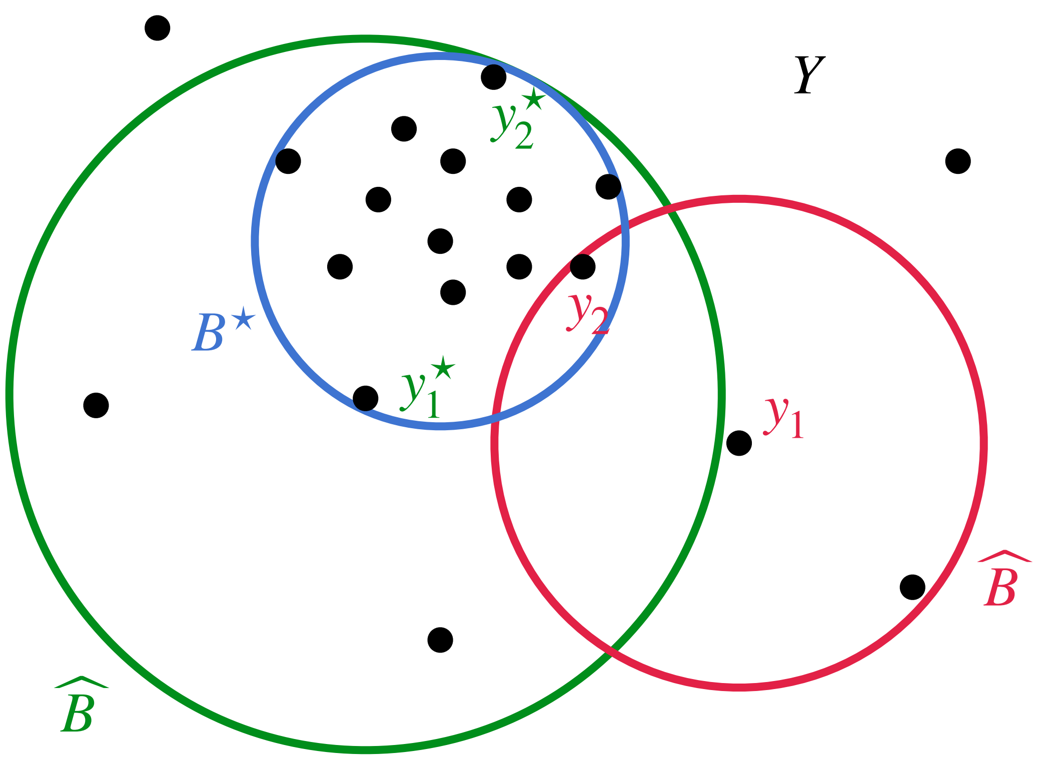
Arbitrary nature of $\mathscr{D}$ makes this is a **worst-case** problem (rather than average-case)




How to compute such an $\widehat{S}$ in polynomial time?

- Can find the min.-volume ball enclosing **all** samples in polynomial time (SDP)
- Capturing $\delta$-fraction is more challenging, hope to achieve volume **approximation**

# Simple Approximation




Sample set of points $Y$ from $\mathscr{D}$

**Algorithm:**

- For every pair of points $y_1, y_2 \in Y$, construct a ball $\widehat{B}$ with center $y_1$ and farthest point $y_2$
- Search over all $\leq n^2$ possibilities, output smallest vol. such ball containing at least $\delta$ of $Y$

**Analysis:** Choose $y_1^\star, y_2^\star$ to be maximally distant points in the optimal solution $B^\star$ to cover all of $B^\star$ and get

$$2\text{-approx. in radius} \iff (2^d)\text{-approx. in volume.}$$

Strategies based on coresets can get volume approximations [Badoiu Har-Peled Indyk '02]

$$\exp\left(O(d/\mathrm{polylog}(d))\right).$$

**Proper learning!**

# Informal Result

Can we get a better approximation?

For any constant $\varepsilon > 0$, **NP-hard** to **properly** approximate min.-volume ball containing $\delta$ of $\mathscr{D}$ up to factor

$$(1 + 1/d^{\varepsilon}) \text{ in radius} \iff \exp(d^{1-\varepsilon}) \text{ in volume}$$

Can get much better approximation via improper learning!

**Improper learning:** Find confidence set that is an ellipsoid

**Informal result:** Given a polynomial number of samples from $\mathscr{D}$, in polynomial time it is possible to find an ellipsoid $\widehat{E}$ that achieves coverage $\approx \delta$, and has volume at most

$$\mathrm{vol}(\widehat{E}) \leq \exp(\widetilde{O}(d^{1/2})) \cdot \mathrm{vol}(B^{\star}),$$

where $B^{\star}$ is the optimal ball achieving coverage $\delta$ over $\mathscr{D}$.

**via robust estimation!**

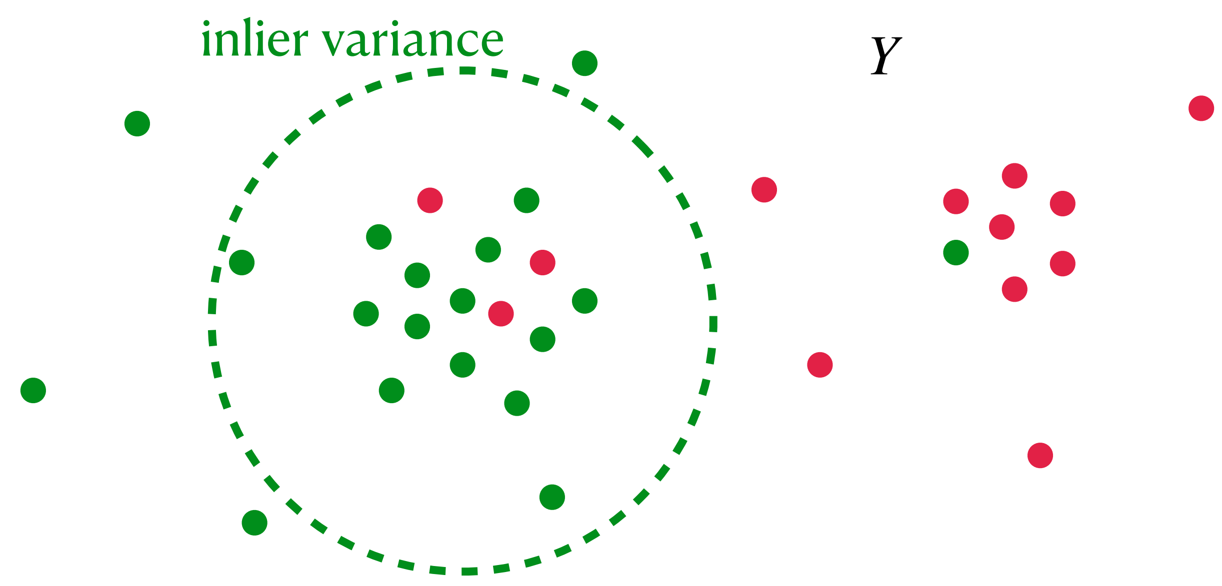
# Robust High-dimensional Estimation

**Goal:** Estimate statistics of adversarially corrupted data

**Example:** Given samples $Y$ from a distribution

$$\mathscr{D} = \delta\mathscr{D}_{\text{inlier}} + (1-\delta)\mathscr{D}_{\text{outlier}},$$

estimate the mean of $\mathscr{D}_{\text{inlier}}$.




Need to take advantage of structure in inlier distribution

→ for example: inliers have **bounded variance** in every direction

In polynomial time, have techniques to robustly estimate

- **median**: minimizes sum of distances to samples
- **mean**: minimizes sum of squared distances to samples

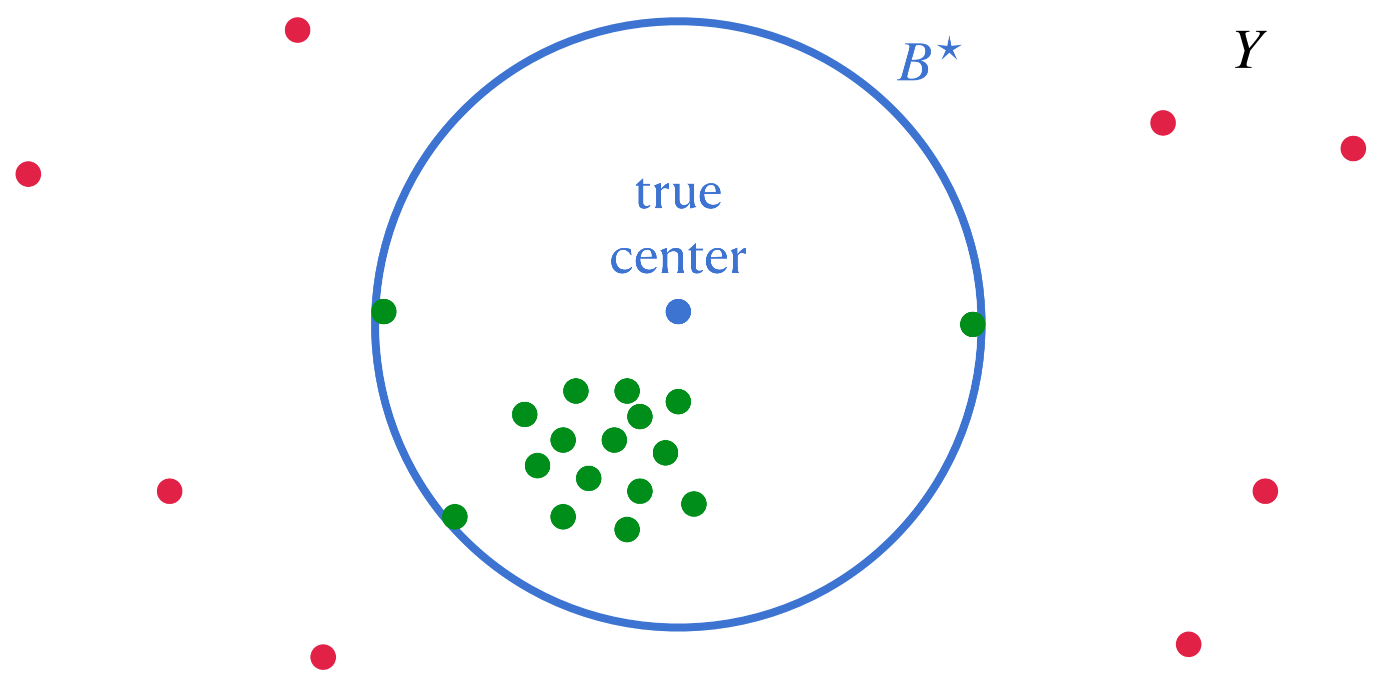
Can we robustly estimate the **center**?
(minimizes maximum distance to samples)

# Connection to Confidence Sets

**Volume optimality:** Compete with $B^\star$, the optimal ball capturing $\delta$-fraction of $\mathscr{D}$

Let $\mathscr{D}_{\text{inlier}}$ be the distribution in $B^\star$, $\mathscr{D}_{outlier}$ be the distribution outside $B^\star$, thus

$$\mathscr{D} = \delta \mathscr{D}_{\text{inlier}} + (1-\delta)\mathscr{D}_{\text{outlier}}.$$




Would like to estimate the **center** (minimize the maximum distance to points) of the inlier distribution $\mathscr{D}_{\text{inlier}}$ .

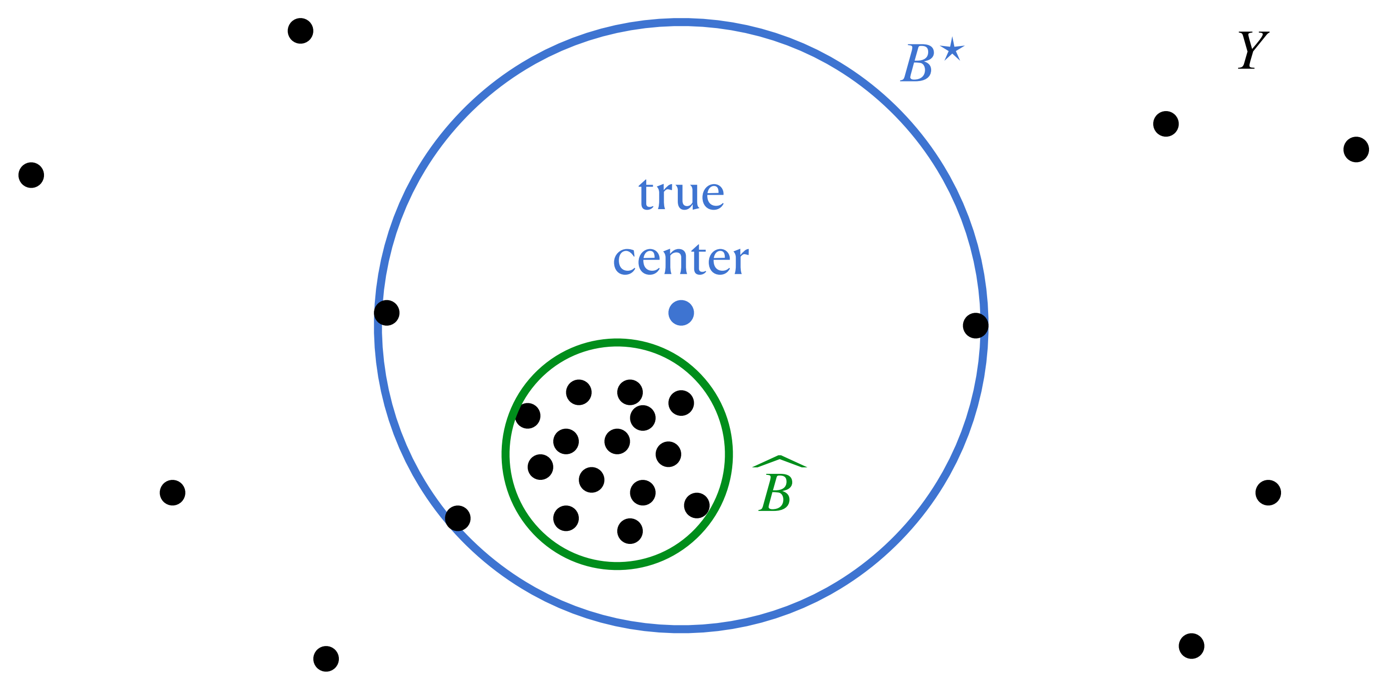
$\rightarrow$ would give the center of $B^\star$, easy to guess radius once we have the center

**Any method that produces small confidence sets must be robust to outliers!**

# Robust Center Estimation?

In $\mathbb{R}^d$, the **center** only depends on $d + 1$ points, so it is **not robustly estimatable**.

(cannot hope for robust estimation, even statistically)




**Relaxation:** compete with best ball that covers $\delta$-fraction of $Y$, but only cover $(\delta - \gamma)$-fraction of $Y$, for small **coverage slack factor** $\gamma > 0$.

(Think of $\delta = 0.9, \gamma = 0.01, \delta - \gamma = 0.89$)

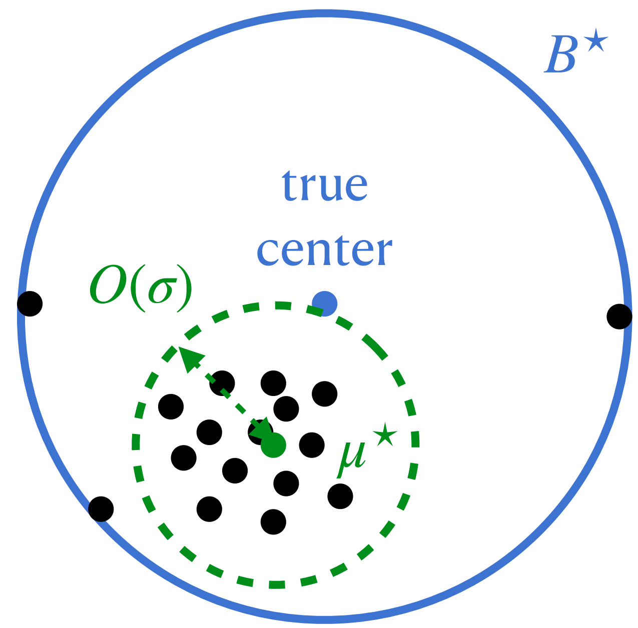
Coverage slack is relatively benign, and gives us a foothold in the algorithmic problem

# Mean as Robust Proxy for Center

Chebyshev's inequality says most points in $B^\star$ (inliers) must be within a few **standard deviations**, $\sigma$, of the mean $\mu^\star$ of $B^\star$

$\implies \mu^\star$ is a proxy for the center of **most** of the points!

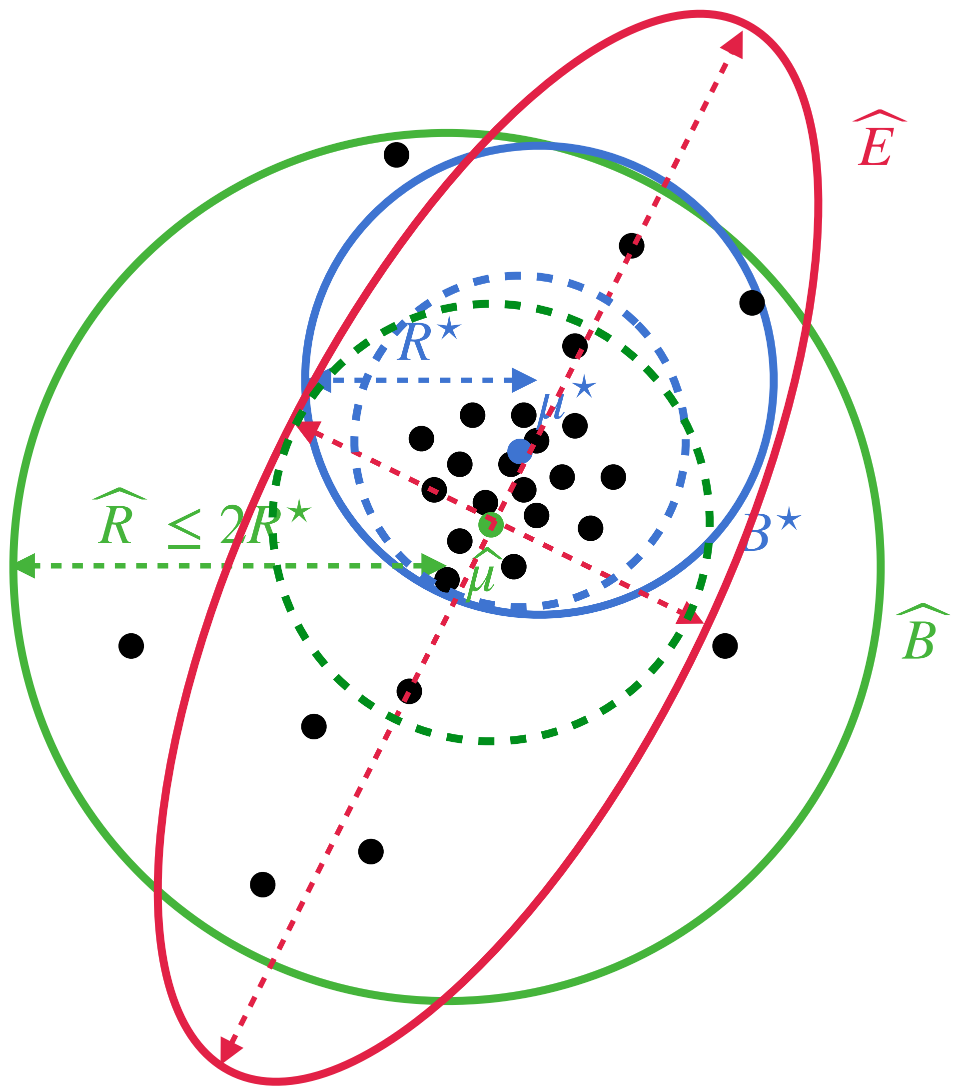
Can hope to robustly estimate $\mu^\star$




Requires bound on the variance of the points in $B^\star$

- to bound radius of ball around $\mu^\star$
- to accurately recover $\mu^\star$, the mean of inliers

# Bootstrapping Variance Bound




**Recall:** Simple algorithm finds $\widehat{B}$ that contains all points in optimal $B^\star$, with at most $2\times$ radius of $B^\star$

Because points are bounded in $\widehat{B}$, **on average** over directions their variance is low!

Centering a ball around $\mu^\star$, mean of $B^\star$, captures most of the mass in most directions

$\mu^\star$ is near $\widehat{\mu}$ since $B^\star$ contains at least $\delta$ mass, so can center ball at $\widehat{\mu}$, mean of $\widehat{B}$

There can only be a few directions in which the variance of points in $\widehat{B}$ is much higher than average, in which we expand our set

$\rightarrow$ log-concavity of volume means it is ok to expand a lot in a few directions

Algorithmic via PCA

# Result

**Theorem:** We give a polynomial-time algorithm, that for a target coverage $\delta \in (0,1)$, and coverage slack $\gamma \in (0,1)$, given $n = \Omega(d^2/\gamma^2)$ samples drawn i.i.d. from an arbitrary $\mathscr{D}$, finds with high probability a set $S$ such that

$$\mathbb{P}_{y \sim \mathscr{D}}[y \in S] \geq \delta,$$

and

$$\mathrm{vol}(S) \leq \mathrm{vol}(B^\star) \cdot \exp\left(O_{\delta,\gamma}(d^{1/2+o(1)})\right),$$

where $B^\star$ is the minimum volume ball that achieves $\delta + \gamma + O(\sqrt{d^2/n})$ coverage over $\mathscr{D}$ .

- Combined with hardness for approximating balls with balls, gives a **separation between proper and improper learning** for a natural task
- Can use ideas to compete against sets that are unions of balls
- **Beyond-worst-case extension:** Can use **list-decodable mean estimation** as a black-box to get an $O(1)$ volume approximation factor for nicely-behaved distributions (inliers are approximately isotropic)

# Conclusion

**Problem:** Estimating the high-density region of an **arbitrary** distribution

**Application:** Conformal prediction, and more!

**Techniques:** High-dimensional robust estimation toolkit

**Results:** Polynomial-time approximation algorithm, and separation between proper and improper learning

**Future directions:**

- Can we improve the approximation factor for balls, or prove hardness?
- Can we approximate other natural set families? Ex: $\ell_p$ balls for $p$ other than 2 [Braun Aolaritei Jordan Bach '25]
- Give a statistical characterization of tractability (i.e., bounded VC-dimension is sufficient, but is it necessary?)
- Online and/or streaming algorithms? [Angelopoulos Candes Tibshirani '23][**S.** '25]

# Thanks!

**vaidehi@u.northwestern.edu**