

# Computing High-dimensional Confidence Sets for Arbitrary Distributions

**Chao Gao**

CHAOGAO@UCHICAGO.EDU

*Department of Statistics, University of Chicago, Chicago, USA*

**Liren Shan**

LIRENSHAN@TTIC.EDU

*Toyota Technological Institute at Chicago, Chicago, USA*

**Vaidehi Srinivas**

VAIDEHI@U.NORTHWESTERN.EDU

**Aravindan Vijayaraghavan**

ARAVINDV@NORTHWESTERN.EDU

*Department of Computer Science, Northwestern University, Evanston, USA*

**Editors:** Nika Haghtalab and Ankur Moitra

## Abstract

We study the problem of learning a high-density region of an arbitrary distribution over  $\mathbb{R}^d$ . Given a target coverage parameter  $\delta$ , and sample access to an arbitrary distribution  $\mathcal{D}$ , we want to output a confidence set  $S \subset \mathbb{R}^d$  such that  $S$  achieves  $\delta$  coverage of  $\mathcal{D}$ , i.e.,  $\mathbb{P}_{y \sim \mathcal{D}}[y \in S] \geq \delta$ , and the volume of  $S$  is as small as possible. This is a central problem in high-dimensional statistics with applications in high-dimensional analogues of finding confidence intervals, uncertainty quantification, and support estimation.

In the most general setting, this problem is statistically intractable, so we restrict our attention to competing with sets from a concept class  $\mathcal{C}$  with bounded VC-dimension. An algorithm for learning confidence sets is competitive with class  $\mathcal{C}$  if, given samples from an arbitrary distribution  $\mathcal{D}$ , it outputs in polynomial time a set that achieves  $\delta$  coverage of  $\mathcal{D}$ , and whose volume is competitive with the smallest set in  $\mathcal{C}$  with the required coverage  $\delta$ . This problem is computationally challenging even in the basic setting when  $\mathcal{C}$  is the set of all Euclidean balls. Existing algorithms based on coresets find in polynomial time a ball whose volume is  $\exp(\tilde{O}(d/\log d))$ -factor competitive with the volume of the best ball.

Our main result is an algorithm that finds a confidence set whose volume is  $\exp(\tilde{O}(d^{1/2}))$  factor competitive with the optimal ball having the desired coverage. It is surprisingly simple and also extends to finding confidence sets competitive against unions of  $k$  balls, and improved guarantees under additional assumptions. The algorithm is improper (it outputs an ellipsoid). Combined with our computational intractability result for proper learning balls within an  $\exp(\tilde{O}(d^{1-o(1)}))$  approximation factor in volume, our results provide an interesting separation between proper and (improper) learning of confidence sets.

**Keywords:** Confidence Sets, Conformal Prediction, High-dimensional Statistics, Robust Statistics, Separation between Proper and Improper Learning, Uncertainty Quantification

## 1. Introduction

We consider the problem of learning a high-density region of an arbitrary distribution. That is, given an arbitrary distribution  $\mathcal{D}$  over  $\mathbb{R}^d$ , and a coverage factor  $0 \leq \delta \leq 1$ , we want to find the set  $S \subset \mathbb{R}^d$  that minimizes volume, while achieving coverage at least  $\delta$ :

$$\min \text{vol}(S) \quad \text{s.t. } S \subset \mathbb{R}^d, \quad \mathbb{P}_{y \sim \mathcal{D}}[y \in S] \geq \delta, \quad (1)$$

where  $\text{vol}(S)$  denotes the volume or Lebesgue measure of the set  $S$ . The minimum of (1) is known as the generalized quantile function (Einmahl and Mason, 1992; Polonik, 1997), an important quantity in statistical inference that plays a central role in a number of problems including estimating density level sets (Garcia et al., 2003), distribution support estimation (Schölkopf et al., 2001) and conformal prediction (Gao et al., 2025). A set containing at least  $\delta$ , say 0.9, probability mass generalizes the idea of confidence intervals in one-dimension to a high-dimensional setting.

Solving (1) leads to a number of interesting applications in statistical inference. On the one hand, finding a set that contains a large probability mass of the distribution can be useful as a high-dimensional analogue of support estimation (Schölkopf et al., 2001). On the other hand, as  $\delta \rightarrow 0$ , the maximizer of (1) can be used as a mode estimator (Sager, 1978, 1979). Algorithms that learn confidence sets for arbitrary distributions can also be used as a tool for estimating the uncertainty in the predictions of black-box machine learning models, as in conformal prediction (Gammerman et al., 1998; Vovk et al., 2005). Finally, algorithms for this task have to be robust to contamination present in the data, as the remaining  $(1 - \delta)$  fraction of the data could be outliers. The formulation in (1) leads to a different notion of robustness that is non-parametric and may be of independent interest in robust estimation (Rousseeuw, 1984, 1985; Van Aelst and Rousseeuw, 2009). There is also a rich body of work on robust statistics, where the remaining  $1 - \delta$  fraction can be arbitrary outliers, but these works often make stronger model assumptions about the  $\delta$  fraction of points that are “inliers” (Huber, 1964; Diakonikolas and Kane, 2023). See Section 1.5 for applications in statistics like conformal prediction, goodness-of-fit, robust estimation and testing unimodality.

**Density estimation and curse of dimensionality.** A popular approach to learn minimum volume sets is based on the fact that the solution to (1) is given by level sets of the density function  $\{y \in \mathbb{R}^d : p(y) \geq t\}$ . The volume optimality of taking the level set of a density estimator can be established in some nonparametric settings (Rigollet and Vert, 2009). However, these methods typically assume additional properties of the density function of the distribution, and are usually limited to low dimensional settings since the sample complexity of nonparametric density estimation suffers from an exponential dependence on the dimension.

**Learning Confidence Sets with Bounded VC-dimension.** Consider the setting where there exists a good confidence set belonging to a family  $\mathcal{C}$  of finite VC-dimension. When independent samples from the distribution  $\mathcal{D}$  are available, one can solve the empirical version of (1) with polynomial sample complexity and establish approximate volume optimality as long as the concept class  $\mathcal{C}$  has finite VC-dimension (Polonik, 1999; Scott and Nowak, 2005). This is related to the (agnostic) PAC learning setting in classification where the confidence sets  $S$  can be identified with classifiers belonging to the concept class  $\mathcal{C}$ . This leads to the following broad class of learning problems that we study:

**Learning confidence sets competitive with  $\mathcal{C}$ .** An algorithm for learning confidence sets is  $\Gamma$ -competitive with  $\mathcal{C}$ , if given samples from an arbitrary distribution  $\mathcal{D}$  in  $\mathbb{R}^d$ , for any  $\gamma > 0$ , it finds a set  $S \subset \mathbb{R}^d$  in time polynomial (in  $d, 1/\gamma$ ) satisfying  $\mathbb{P}_{y \sim \mathcal{D}}[y \in S] \geq \delta$ , and

$$\text{vol}(S)^{1/d} \leq \Gamma \cdot \min \left\{ \text{vol}(C)^{1/d} : C \in \mathcal{C} \text{ s.t. } \mathbb{P}_{y \sim \mathcal{D}}[y \in C] \geq \delta + \gamma \right\}. \quad (2)$$

The small loss of the  $\gamma$  term in coverage accounts for any potential sampling error. Above, we have followed the standard convention of measuring  $\text{vol}^{1/d}$  instead of volume  $\text{vol}$  (e.g., for Euclidean balls this becomes proportional to the radius of the ball). The factor  $\Gamma$  corresponds to the approximation factor or competitive ratio of the algorithm.

If there exists a small confidence set  $C^* \in \mathcal{C}$  with (slightly more than) the required coverage  $\delta$ , then the algorithm will output a set  $S$  with coverage  $\delta$  and whose volume is competitive with  $C^*$ . The set  $S$  is not required to belong to the family  $\mathcal{C}$ . This is analogous to the standard setting in PAC learning, where the learning algorithm may output a classifier that does not belong to the concept class  $\mathcal{C}$  (see e.g., [Shalev-Shwartz and Ben-David, 2014](#)). If the algorithm outputs a confidence set  $S$  that belongs to  $\mathcal{C}$ , we will call it a *proper* learner of confidence sets (analogous to proper PAC learning). When we want to be more explicit, the term *improper* learner may be used for an algorithm that does not necessarily output a set belonging to the concept class  $\mathcal{C}$ . We emphasize the worst-case nature of the distribution that the samples are drawn i.i.d from. This is analogous to the distribution-free setting of PAC learning, where we make no assumptions on the distribution. While there is a simple algorithm that is sample efficient (but computationally inefficient) when the VC-dimension of  $\mathcal{C}$  is bounded (by a polynomial), the algorithmic problem becomes challenging in high dimensions.

**The family of Euclidean Balls.** Consider the basic setting where  $\mathcal{C}$  is the set of  $\ell_2$  balls (this has VC-dimension at most  $d$ ). This is a natural generalization of intervals in higher dimensions, and is motivated by applications (see e.g., [Wang et al., 2023](#)). It is NP-hard to find the smallest set in  $\mathcal{C}$  that covers a  $\delta$  fraction of points from an arbitrary set ([Shenmaier, 2015](#)).<sup>1</sup> The famous work of Badoiu, Har-Peled and Indyk on coresets for minimum volume enclosing balls ([Badoiu et al., 2002](#); [Agarwal et al., 2004](#); [Ding, 2020](#)) gives a proper learning algorithm that runs in polynomial time and is  $\Gamma = 1 + O(1/\log d)$ -competitive i.e., a  $\exp(O(d/\log d))$  factor approximation in volume. To the best of our knowledge, these works using core-sets ([Badoiu et al., 2002](#); [Ding, 2020](#)) remain the state-of-the-art guarantee for learning confidence sets competitive with Euclidean balls in  $d$  dimensions. We focus on Euclidean balls and unions of Euclidean balls in the rest of the paper.

### 1.1. Our Results

Our main algorithmic result is an algorithm that is  $\Gamma \approx 1 + d^{-1/2}$  competitive against Euclidean balls.

**Theorem 1 (Learning confidence sets competitive with Euclidean balls)** *There is a polynomial time algorithm that for any target coverage  $\delta \in (0, 1)$  and coverage slack  $\gamma \in (0, 1)$ , when given  $n = \Omega(d^2/\gamma^2)$  samples drawn i.i.d. from an arbitrary distribution  $\mathcal{D}$ , finds with high probability a set  $S \subset \mathbb{R}^d$  that is  $\Gamma = \exp(O_{\gamma,\delta}(d^{-1/2+o(1)}))$  competitive, i.e.,*

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

and

$$\text{vol}(S)^{1/d} \leq \text{vol}(B^*)^{1/d} \left(1 + O\left(d^{-1/2+o(1)}(\gamma\delta)^{-1}\right)\right).$$

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

---

1. On the other hand, finding the minimum volume ball that encloses all the given  $n$  points is polynomial time solvable.

Please see Appendix B for more details. To compare against existing methods, consider the setting when  $\delta, \gamma$  are constants. A straightforward solution that considers balls around the sample points that covers at least  $\delta$  fraction of the data points achieves a factor of  $\Gamma = 2$  (due to the triangle inequality).<sup>2</sup> Algorithms based on the powerful technique of coresets for minimum volume enclosing balls due to Badoiu et al. (2002) gives  $\Gamma = 1 + \tilde{O}(1/\log d)$ -competitive. In terms of the volume, the  $\Gamma \approx 1 + d^{-1/2}$  factor corresponds to volume approximation of

$$\Gamma^d = (1 + O_{\delta, \gamma}(d^{-1/2+o(1)}))^d = \exp\left(O_{\gamma, \delta}(d^{1/2+o(1)})\right),$$

which corresponds to a substantial improvement from the  $\exp(O_{\delta, \gamma}(d/\log d))$  volume approximation factor in existing work (Badoiu et al., 2002; Ding, 2020).

**Union of Balls  $\mathcal{C}_k$ .** We next consider the more general concept class  $\mathcal{C}_k$  which is the unions of  $k$  Euclidean balls. This family has VC-dimension of at most  $kd$ . We obtain a similar guarantee for unions of  $k = O(1)$  balls, by recursively applying the algorithm from Theorem 1.

**Theorem 2 (Union of Balls)** *Let  $\delta \in (0, 1), \gamma \in (0, 1), k \in \mathbb{N}$  be any constants. There is a polynomial time algorithm that for target coverage  $\delta \in (0, 1)$  and coverage slack  $\gamma \in (0, 1)$  when given  $n = \Omega(kd^2/\gamma^2)$  samples drawn i.i.d. from an arbitrary distribution  $\mathcal{D}$ , finds with high probability a set  $S \subset \mathbb{R}^d$  that is  $\Gamma = (1 + O_{\gamma, \delta}(d^{-1/2+o(1)}))$  competitive; more precisely, it satisfies  $\mathbf{P}_{y \sim \mathcal{D}}[y \in S] \geq \delta$ , and*

$$\text{vol}(S)^{1/d} \leq \text{vol}(C_k^*)^{1/d} \left(1 + O_{k, \delta}(d^{-1/2+o(1)})\right) \cdot \left(\frac{O(\log(k/\gamma))}{\gamma}\right)^{1/d}$$

where  $C_k^*$  is the minimum volume union of  $k$ -balls that achieves at least  $\delta + \gamma + O(\sqrt{kd^2/n})$  coverage over  $\mathcal{D}$ .

See Corollary 15 for details. Our algorithms are surprisingly simple and conceptually different from existing algorithms based on coresets. It is based on a new connection to robust estimation that follows from a structural statement about the mean and the center of a large fraction of points in the optimal ball with the desired coverage. We give a detailed description of the ideas in the technical overview.

## 1.2. Proper Learning: Separation Result and Algorithms

The algorithm in Theorem 1 outputs a set  $S \subset \mathbb{R}^d$  of small volume achieving the desired coverage, when there is a small ball achieving the required coverage. The set  $S$  may not be a ball, and may in general be an ellipsoid. Similarly the algorithm in Theorem 2 outputs a union of  $O(k/\gamma)$  ellipsoids. These correspond to “improper” learners for confidence sets. In contrast, the best *proper* learning algorithms achieve a worse competitive ratio of  $\Gamma = 1 + O(1/\log d)$ . The following theorem shows that such a loss is unavoidable assuming  $P \neq NP$ , unless we are willing to make additional assumptions on  $\mathcal{D}$ .

---

2. In fact this strategy achieves a  $\Gamma = 2$  factor approximation (with  $\text{vol}^{1/d}$ ) for any set  $\mathcal{C}$  that consists of scalings and translations of a fixed convex shape. However, it is NP-hard to obtain a  $(2 - o(1))$  factor approximation even for the  $\ell_\infty$  ball (Shenmaier, 2015).

**Theorem 3 (NP-hardness of Proper Learning for Balls)** *For any constant  $\varepsilon > 0$ , assuming  $P \neq NP$  there is no polynomial time learning algorithm that given samples from an arbitrary distribution over  $\mathbb{R}^d$  that outputs balls as confidence sets that are  $\Gamma \leq (1 + \frac{1}{d^\varepsilon})$  competitive with balls.*

See Theorem 17 for a more formal statement. Note that the above result is a worst-case hardness result, and reflects the arbitrary nature of the distribution. This is reminiscent of NP-hardness of proper learning in supervised learning (see e.g., Feldman et al., 2009). The above theorem follows from a simple reduction that amplifies the existing NP-hardness result due to (Shenmaier, 2015). In Theorem 19, we also show computational intractability (assuming the Small Set Expansion hypothesis of Raghavendra and Steurer (2010)) that shows a similar quantitative hardness, even when the coverage can be smaller by an arbitrary constant factor. Combined with our algorithmic result in Theorem 1, this shows a formal separation between proper and improper learning for a natural geometric learning problem.

**Proper learning algorithms.** When we are required to output a confidence set from  $\mathcal{C}$ , the algorithm in Theorem 1 can also be used to output a ball with a worse competitive ratio (which is expected due to Theorem 3). When we make no assumptions about the distribution  $\mathcal{D}$ , our algorithm can output a ball that is  $\Gamma = 1 + \tilde{O}(1/\log d)$  competitive. See Theorem 8 in Appendix A for a formal statement of the theorem. This guarantee recovers the guarantee from the coresset-based algorithm which achieves  $\Gamma = 1 + \tilde{O}(1/\log d)$  (Badoiu et al., 2002). In the worst-case instances of the algorithm (and the hard instances in Theorem 3), the points sampled from  $\mathcal{D}$  that lie inside the optimal ball  $B^*$  effectively lie on a lower-dimensional subspace.

On the other hand, our algorithm can give a significantly stronger guarantee when the points inside  $B^*$  are approximately isotropic. In the following theorem, we consider the setting where we have  $n$  points  $Y$  drawn from  $\mathcal{D}$ , and the guarantee depends on the covariance of the sampled points within the optimal ball  $B^*$ .

**Theorem 4 (Bounded variance implies better bounds)** *Let  $Y \subseteq \mathbb{R}^d$  be a set of  $n$  points, and  $\delta, \gamma \in (0, 1)$ , such that there is an unknown ball  $B^* = B(c^*, R^*)$  with  $Y^* = B^* \cap Y$  satisfying  $|Y^*| \geq \delta|Y|$ , and  $Y^*$  is  $\beta(R^*)^2/d$ -isotropic for some  $\beta \geq 0$ . That is, if  $\mu^*$  is the mean of the points  $Y^*$ , then*

$$\Sigma_{Y^*} = \frac{1}{|Y^*|} \sum_{y \in Y^*} (y - \mu^*)(y - \mu^*)^\top \preceq \beta \frac{(R^*)^2}{d} I.$$

*Then we can find a ball  $\hat{B} = B(\hat{c}, \hat{R})$  such that  $|Y \cap \hat{B}| \geq (1 - \gamma)|Y^*|$ , and*

$$\text{vol}(\hat{B})^{1/d} \leq \text{vol}(B^*)^{1/d} \sqrt{1 + O\left(\frac{\beta}{\gamma \delta d}\right)},$$

*in polynomial time.*

In the above Theorem 4,  $\Sigma_{Y^*}$  is a PSD matrix of dimension  $d \times d$  with  $\text{tr}(\Sigma_{Y^*}) \leq (R^*)^2$ , since  $Y^* \subseteq B^*$ . The isotropicity condition requires the maximum eigenvalue  $\lambda_{\max}(\Sigma_{Y^*}) \leq \beta(R^*)^2/d$  for some  $\beta > 0$ . Thus every  $Y^*$  satisfies the isotropic condition with  $\beta \leq d$ . When the points of  $Y^*$  are very spread out, this corresponds to a setting when  $\beta \leq 1$  (all the eigenvalues are equal). In this case, the theorem guarantees to find a ball with  $\Gamma = 1 + O(d^{-1})$ . In other words, it finds a ball whose

volume is within  $O_{\delta,\gamma}(1)$  factor of the optimum, which is significantly stronger. Finally, the ball  $\hat{B}$  output by the algorithm also has coverage  $(1-\gamma)\delta - O(\sqrt{d/n})$  by standard concentration arguments (see e.g., [Devroye and Lugosi, 2001](#)) since the VC-dimension of  $d$ -dimensional Euclidean balls is  $d+1$ .

All of our algorithmic results are distribution-free i.e., they do not make any assumptions on the distribution  $\mathcal{D}$ . While it may be impossible to learn the distribution or even estimate the density at a given point (these tasks often incur an exponential dependence on the dimension  $d$  even for smooth distributions), our results show that one can obtain polynomial time algorithms that find approximately optimal dense sets from  $\mathcal{C}$  covering at least  $\delta$  mass.

### 1.3. Application to Conformal Prediction

As a direct application of our result we get the following algorithm for conformal prediction in high dimensions. In conformal prediction, an algorithm is given training examples  $Y_1, \dots, Y_n$  in some set  $\mathcal{Y}$ , and a miscoverage rate  $\alpha$ , and must output a set  $\hat{C}$ , where

$$\mathbb{P}[Y_{n+1} \in \hat{C}] \geq 1 - \alpha,$$

for some unseen training example  $Y_{n+1}$ , assuming only that the training and test examples  $Y_1, \dots, Y_{n+1}$  are *exchangeable*.<sup>3</sup> There is a fairly simple way to “conformalize” our result to achieve this goal, resulting in the following algorithm that always outputs a valid conformal set when the data is exchangeable, and is additionally approximately volume optimal when the data is drawn i.i.d. from an (unknown) distribution  $\mathcal{D}$ .

**Theorem 5 (Conformal Prediction with Approximate Volume Optimality)** *We have an algorithm for conformal prediction over examples from  $\mathcal{Y} = \mathbb{R}^d$ , that achieves approximate volume optimality with respect to the set  $\mathcal{C}$  of Euclidean balls.<sup>4</sup> That is, we have an algorithm that, given training examples  $Y_1, \dots, Y_n \in \mathcal{Y}$ , miscoverage rate  $0 < \alpha < 1$ , and coverage slack factor  $0 \leq \gamma \leq 1$ , outputs a set  $\hat{C}$  (not necessarily in  $\mathcal{C}$ ), such that for an unknown test example  $Y_{n+1} \in \mathcal{Y}$ ,*

(a) *if  $Y_1, \dots, Y_{n+1}$  are exchangeable, then*

$$\mathbb{P}[Y_{n+1} \in \hat{C}] \geq 1 - \alpha.$$

(b) *if  $Y_1, \dots, Y_{n+1}$  are drawn i.i.d. from some (unknown) distribution  $\mathcal{D}$ , and  $n = \Omega(d^2/\gamma^2)$ , then*

$$\text{vol}(\hat{C})^{1/d} \leq \left(1 + O_{\gamma,\delta}(d^{-1/2+o(1)})\right) \text{vol}(C^*)^{1/d},$$

where

$$C^* = \underset{C \in \mathcal{C}}{\text{argmin}} \text{vol}(C) \quad \text{s.t.} \quad \mathbb{P}[Y_{n+1} \in C] \geq 1 - \alpha + \gamma.$$

We provide an in depth discussion of this application in Appendix E. See also Section 1.5 for more applications in high-dimensional statistics. We now review related literature and other methods that have been explored in literature, followed by an overview of the algorithm and techniques.

3. That is,  $\mathbb{P}[Y_1 = y_1, \dots, Y_{n+1} = y_{n+1}] = \mathbb{P}[Y_1 = y_{\pi(1)}, \dots, Y_{n+1} = y_{\pi(n+1)}]$ , for all  $y_1, \dots, y_{n+1} \in \mathcal{Y}$  and permutations  $\pi : [n+1] \rightarrow [n+1]$ .

4. This is an approximate form of *restricted volume optimality* defined by [Gao et al. \(2025\)](#).



#### 1.4. Prior and Related Algorithmic Approaches

The problem of learning minimum volume sets has a long history in statistical learning and computer science. It also goes under the names of learning density level sets, anomaly detection, or distribution support estimation in the literature. Learning algorithms that have been studied in the literature can roughly be categorized into the following three classes.

1. *Excess mass functional.* The empirical version of Equation (1) is

$$\min \text{vol}(C) \quad \text{s.t. } C \in \mathcal{C} \text{ and } \mathbb{P}_n(C) \geq \delta - \gamma, \quad (3)$$

where  $\mathbb{P}_n$  is the empirical distribution and  $\gamma$  is some slack parameter (Scott and Nowak, 2005). A more general objective function called the excess mass functional proposed by Hartigan (1987) is defined by

$$E_\lambda(C) = \mathbb{P}_n(C) - \lambda \text{vol}(C), \quad (4)$$

where the parameter  $\lambda > 0$  is determined by the confidence level. In fact, maximizing the excess mass functional over  $C \in \mathcal{C}$  is equivalent to Equation (3). This approach has been well studied in the literature. The original work (Hartigan, 1987) considers  $\mathcal{C}$  being the collection of convex sets. The collection of ellipsoids is considered by Nolan (1991). Convergence rates of the maximizer of Equation (4) are investigated by Tsybakov (1997) in a nonparametric setting for star-shaped sets and convex sets.

2. *One-class classification.* Another type of algorithms aim to learn a nonparametric function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , and the confidence set is taken as  $\{y \in \mathbb{R}^d : f(y) \geq 0\}$ . This approach was pioneered by Schölkopf et al. (2001) and they model the function  $f$  in a feature space by

$$f(y) = \text{sign}(w^T \Phi(y) - \rho), \quad (5)$$

where  $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}^k$  is some feature mapping. The parameters  $w$  and  $\rho$  are solutions to the following optimization problem,

$$\max_{w, \rho, \xi} \frac{1}{2} \|w\|^2 + \frac{1}{\lambda n} \sum_{i=1}^n \xi_i - \rho \quad \text{s.t. } w^T \Phi(Y_i) - \rho + \xi_i \geq 0, \quad (6)$$

where  $\xi \in \mathbb{R}^n$  is a margin vector and  $\lambda$  is determined by the confidence level. The optimization (6) is known as the one-class support vector machine (SVM) in the literature, which is equivalent to another optimization problem called support vector data description (SVDD) (Tax and Duin, 2004) using some particular feature mapping. Recasting learning minimum volume sets as solving classification is natural. The excess mass functional (4) already suggests that the problem is equivalent to testing between two measures, empirical distribution against volume (uniform). Formally, this connection was established by Steinwart et al. (2005). Unlike traditional classification, here all the data points are generated from the same distribution, which leads to the name “one-class classification” in the literature. Compared with directly learning confidence sets through (3) or (4), the one-class classification perspective is more flexible, and can be easily combined with deep learning models (Ruff et al., 2018).

3. *Density estimation plug-in.* The population version of the excess mass functional is  $P(C) - \lambda \text{vol}(C)$ , which is maximized by the density level set

$$\{y \in \mathbb{R}^d : p(y) \geq t\},$$

by Neyman–Pearson lemma, since the density function  $p$  can be regarded as the likelihood ratio between  $P$  and volume (Garcia et al., 2003). Thus, it is natural to replace  $p$  by some density estimator  $\hat{p}$ . The plug-in strategy has been considered by Hyndman (1996); Park et al. (2010); Lei et al. (2013) among others. Under certain nonparametric setting, the optimality of the plug-in strategy can be proved (Rigollet and Vert, 2009).

**Algorithms based on Coresets.** Coresets are a powerful algorithmic primitive for geometric problems involving an input set of data points (Agarwal et al., 2004). A coreset for a problem is a small summation of the data (e.g., a subset of the points) such that solving the problem on the summation gives an approximate solution to the whole instance. Badiou, Har-Peled and Indyk proved a surprising result about coresets for the problem of minimum volume enclosing ball for a set of points – they showed that there is a coreset of size  $1/\varepsilon$  (that is independent of the dimension) whose minimum volume enclosing ball approximates the minimum volume ball enclosing all the points up to a  $(1 + \varepsilon)$  factor in radius (Badoiu et al., 2002).<sup>5</sup> Badoiu et al. (2002) show how to find a ball enclosing  $\delta - \gamma$  fraction of the given points in polynomial time whose radius is within a factor  $\Gamma = \left(1 + O_\gamma\left(\frac{(\log \log d)^2}{\log d}\right)\right)$  (see Theorem 4.2 of Badoiu et al., 2002). The subsequent work of Ding (2020) also uses coresets and seems to obtain a guarantee of  $\Gamma = \left(1 + O_\gamma\left(\frac{1}{\log d}\right)\right)$  (see (24) following Corollary 11 in (Ding, 2020)), though this is not stated explicitly. This corresponds to an approximation factor of  $\Gamma^d = \exp(\tilde{O}(d/\log d))$  for the volume. We are not aware of any other work that get a better competitive ratio  $\Gamma$  for balls.

**Comparison to robust mean estimation and covariance estimation.** There is a rich literature for developing robust algorithms to perform basic statistical estimation tasks in high-dimensions when  $1 - \delta$  fraction of the points are contaminated (see e.g., Huber, 2004; Diakonikolas and Kane, 2023). This is usually studied in the context of parameter estimation problems like mean estimation or covariance estimation where a  $\delta$  fraction of points are *inliers* that are assumed to be drawn from a distribution with a location (mean) or covariance, while the rest of the points are arbitrary *outliers*. The inlier distribution is assumed to satisfy some nice property like bounded covariance and/or moments, or typically from a parametric family like Gaussians. Various contamination models have been considered. In the famous Huber contamination model and the stronger adversarial contamination model, a  $\delta = 1 - \eta > 1/2$  fraction of the points are inliers, while a  $\eta = 1 - \delta < 1/2$  fraction of the points are outliers (Huber, 1964; Diakonikolas and Kane, 2023). The setting when  $\delta < 1/2$  has also been studied as list-decodable robust estimation (Charikar et al., 2017). Most relevant to us are the works in algorithmic robust statistics including the important works of Diakonikolas et al. (2016); Lai et al. (2016). which gave polynomial time algorithms for robust mean and covariance estimation in high dimensions. Over the past decade there have been several polynomial time algorithms proposed based on outlier removal, sophisticated convex relaxations including sum-of-squares and many other techniques (please see Diakonikolas and Kane, 2023, for an excellent book on the topic).

5. The coreset size of  $\lceil 1/\varepsilon \rceil$  is also known to be tight (Agarwal et al., 2004).



However, the robust estimation literature usually assumes various distributional assumptions – for mean estimation, they assume at the very least that the inlier distribution has bounded variance in every direction; while for covariance estimation one needs various fourth moment conditions, or other distributional assumptions like Gaussianity. Such assumptions are necessary to define a certifiable property of the inlier distribution in parameter estimation tasks.

In our setting, our main task is not one of parameter estimation. More importantly, the distribution of points is completely arbitrary i.e., we cannot make any assumptions on the distribution of both the inliers and the outliers. But in some settings we use powerful algorithms for list-decodable mean estimation (Charikar et al., 2017) as a black-box subroutine once we reduce the task to mean estimation for bounded covariance distributions. Finally, algorithms for our problem can also be used to design robust estimators for various statistical tasks like mode estimation or robust support estimation, as detailed in Section 1.5.

### 1.5. Applications in Statistics

Successfully learning a small confidence set has implications in many statistical inference tasks, ranging from estimation to hypothesis testing. We will summarize a few important applications below.

1. *Conformal prediction.* The goal of conformal prediction (Gammerman et al., 1998; Vovk et al., 2005; Angelopoulos and Bates, 2023) is to find a set  $\hat{C}$  that is trained from  $y_1, \dots, y_n \sim \mathcal{D}$  such that  $\mathbb{P}(y_{n+1} \in \hat{C}) \geq 1 - \alpha$  for an independent future observation  $y_{n+1} \sim \mathcal{D}$ .

Minimum volume sets are natural candidates for conformal prediction to achieve finite sample coverage and volume optimality simultaneously. For example, conformalized density level sets have already been considered by Lei et al. (2013). More generally, any minimum volume set learning algorithm, including our algorithms in the paper (e.g., Theorem 1, Theorem 2), can be conformalized through the construction of nested systems (Gupta et al., 2022; Gao et al., 2025) to obtain similar approximate volume optimality guarantees. For more details see Appendix E.

2. *Robust estimation.* Suppose  $\hat{C}$  is a  $\delta$ -level confidence set with some  $\delta$  close to zero. It was suggested by Sager (1978, 1979) that one can use some point in  $\hat{C}$  as a mode estimator. More generally, robust location and scatter estimation through learning minimum volume sets is advocated by Rousseeuw (1984, 1985). Specifically, Rousseeuw first computes  $\hat{C}$  as the minimum volume ellipsoid (MVE), and then the center and the shape of the  $\hat{C}$  are used as a robust location estimator and a robust scatter matrix estimator. Other applications of MVE are discussed in Van Aelst and Rousseeuw (2009).
3. *Testing unimodality.* The original motivation of Hartigan (1987) in learning a minimum volume convex set is to test whether a distribution is unimodal. The paper suggested testing statistic  $\sup_{\lambda} \sup_{C \in \mathcal{C}} E_{\lambda}(C)$  that maximizes the excess mass functional (4). The idea was further developed and generalized by Müller and Sawitzki (1991); Polonik (1995); Cheng and Hall (1998).
4. *Anomaly detection.* The problem of anomaly detection, also known as outlier or novelty detection, aims to find a set that is not typical. Given a distribution  $\mathcal{D}$  with density  $p$ , its anomaly region is  $\{y \in \mathbb{R}^d : p(y) < t\}$ , the complement of a density level set. Thus, the

problem is mathematically equivalent to learning a confidence set. We refer the readers to [Chandola et al. \(2009\)](#); [Ruff et al. \(2021\)](#) for comprehensive review of this area.

5. *Goodness-of-fit.* Consider a null hypothesis  $H_0 : P = P_0$ . Whether data is generated from  $P_0$  can be determined by the value of  $|\mathbb{P}_n(C) - P_0(C)|$  for some set  $C$  that is informative. The idea of using minimum volume sets in the goodness-of-fit test was considered by [Polonik \(1999\)](#). In particular, let  $\hat{C}_\alpha$  (resp.  $C_\alpha$ ) be a  $(1 - \alpha)$ -level confidence set computed from  $\mathbb{P}_n$  (resp.  $P_0$ ). The statistic

$$\sup_{\alpha} \left( |\mathbb{P}_n(\hat{C}_\alpha) - P_0(\hat{C}_\alpha)| + |\mathbb{P}_n(C_\alpha) - P_0(C_\alpha)| \right)$$

was proposed and its asymptotic property was analyzed by [Polonik \(1999\)](#).

## 2. Technical Overview

Our goal is, given a set  $Y$  of samples from  $\mathcal{D}$ , to find a set  $S$  that captures  $(1 - \gamma)\delta|Y|$  points of  $Y$ , and has volume comparable to the minimum volume ball  $B^*$  that captures a subset of points  $Y^* \subseteq Y$ , with  $|Y^*| \geq \delta|Y|$ . If we could estimate the center of  $B^*$  well, then we could find  $B^*$  approximately by guessing an appropriate radius. We can focus on the worst-case problem over the samples  $Y$  because we will choose our output set  $S$  to be an ellipsoid. Since the class of ellipsoids in  $d$  dimensions has bounded VC-dimension, the empirical coverage of all ellipsoids enjoys uniform convergence. This means that the coverage of *all* ellipsoids will simultaneously generalize from the empirical samples to the population setting, so it suffices to find an ellipsoid that approximately minimizes volume subject to a coverage constraint on the empirical samples.

Intuitively, one could interpret our problem as trying to recover the *center* of the points  $Y^* \subseteq Y$ ,  $|Y^*| \geq \delta|Y|$  in a way that is robust to adversarial contamination, i.e., we only have access to  $Y$ . However, the center of the ball could depend on only a few points, i.e., a sphere in  $d$  dimensions can be defined by  $d + 1$  points.<sup>6</sup> Thus, it appears that the center is not a quantity that can be robust to outliers. This is in contrast to other quantities, such as the mean or median, that inherently depend on the whole dataset, and are not largely influenced by a few points.

In fact, the goal of recovering the center of  $B^*$  is too strong. Properly learning the minimum volume ball  $B^*$  within an approximation factor  $\Gamma = 1 + 1/d^\varepsilon$ , for any constant  $\varepsilon > 0$  is NP-hard. So we should not expect such a strategy to work in the worst case.

However, it turns out that for certain non-worst-case instances, it is indeed possible to recover  $B^*$  within a better approximation ratio than in the worst case. Our first main technical insight is that when the variance of the points is well-spread (i.e., the sampled points do not have too much of their total variance concentrated in any one direction), the *mean* of the points serves as a good proxy for the center. As a side result, this implies strong proper learning guarantees for non-worst-case instances, where the approximation ratio improves as the variance is more controlled (see [Theorem 8](#) and [Theorem 4](#) in [Appendix A](#)).

This assumption on the variance of the sampled points is quite strong, and we cannot expect it to hold for arbitrary distributions. Our second main technical insight is that it is possible to precondition the sampled points by applying a linear transformation  $T$  to the points, to move them

6. In fact the coresset is a subset of size only  $\lceil 1/\varepsilon \rceil$  that determines a ball that is within a  $(1 + \varepsilon)$  approximation in radius.

into a position that meets the strong variance criterion above. Then, we can solve the non-worst-case proper learning task in the transformed space, and apply the inverse transformation  $T^{-1}$  to the resulting ball to construct a confidence set for the original points.

The transformation  $T$  necessarily distorts the space, so when we apply the inverse transformation  $T^{-1}$  to a ball there are two things to keep in mind. First, the result will no longer be a ball, rather it will be an ellipsoid. This falls in the regime of improper learning, and allows us to sidestep the strong lower bounds against proper learning. Second, the transformation  $T$  “shrinks” the space, so the inverse transformation  $T^{-1}$  will expand the volume of the ball.

Our final result follows from carefully choosing the transformation  $T$  to balance two competing objectives:

- (a) We must control the variance of the transformed points to minimize the approximation ratio of the non-worst-case proper learning algorithm.
- (b) We must control the distortion of  $T$  to minimize the volume blow up from applying  $T^{-1}$  to the result of the proper learning algorithm.

Armed with these insights our final algorithm is quite simple: for each coarse set of candidate points (points contained in a ball centered at one of the sample points, there are  $O(n^2)$  such sets), we estimate a linear transformation  $T$ , we transform the points by  $T$ , we find the smallest ball centered at the mean of the transformed points that achieves the desired coverage, and we apply  $T^{-1}$  to the resulting ball to get a candidate ellipsoid. Finally we take the minimum volume ellipsoid that is found over all coarse sets of candidates.

## 2.1. Proper Learning for Non-Worst-Case Instances

To tackle (a) above, our main observation is that we can approximate the center  $c^*$  of the minimum volume ball  $B^*$  containing  $Y^*$  by using the mean  $\mu^*$  of  $Y^*$ . It is not in general true that  $\mu^*$  should be near  $c^*$ . However, *most* points in  $Y^*$  must be approximately as close to  $\mu^*$  as they are to  $c^*$ . This is illustrated in Figure 1.

**Lemma 6 (For bounded variance, most points are near the mean)** *Let  $Y$  be a set of points in  $\mathbb{R}^d$  that are contained in a ball  $B(c, R)$ . Let  $\mu_Y$  be the empirical mean of  $Y$ , and  $\Sigma_Y$  be the empirical covariance, where the largest eigenvalue of  $\Sigma_Y$  satisfies  $\lambda_{\max}(\Sigma_Y) \leq \sigma^2$ . Let  $\hat{\mu}$  be an arbitrary point such that  $\|\mu_Y - \hat{\mu}\|_2 \leq \tau$ . Then for any  $t \geq 1$ , there exists a subset  $\hat{Y}$  such that  $|\hat{Y}| \geq (1 - \frac{1}{t^2}) |Y|$ , and for every point  $y \in \hat{Y}$ ,*

$$\|y - \hat{\mu}\|_2^2 \leq \|y - c\|_2^2 + t^2(\sigma^2 + \tau^2).$$

**Proof** As a warm-up, we can bound the number of points that deviate significantly from  $\mu_Y$  using Chebyshev’s inequality. Let  $v_\mu = \frac{\mu_Y - c}{\|\mu_Y - c\|}$  be the unit vector in the direction of  $\mu_Y - c$ . We have that

$$\mathbf{P}_{y \sim \text{Unif}(Y)} [|\langle y - \mu_Y, v_\mu \rangle| \geq t\sigma] \leq \frac{1}{t^2}.$$

This means that there exists a subset  $\hat{Y} \subseteq Y$  such that  $|\hat{Y}| \geq (1 - \frac{1}{t^2}) |Y|$ , and for all  $y \in \hat{Y}$  we have  $|\langle y - \mu_Y, v_\mu \rangle| \leq t\sigma$ . For each  $y \in \hat{Y}$ , we can decompose  $y - \mu_Y$  into the portion

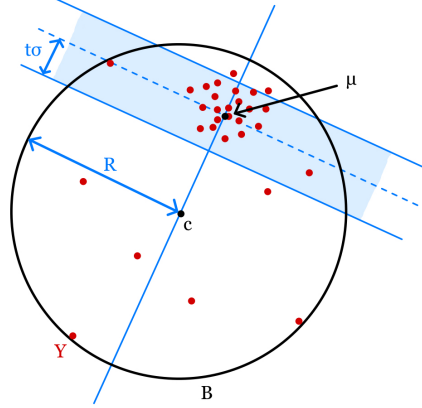


Figure 1: The figure shows the points  $Y$  (in red) with mean  $\mu$ , that are contained in ball  $B$  with center  $c$  and radius  $R$ . It is not necessarily the case that  $\mu$  is near  $c$ , as  $B$  can be defined by just a few points. However, Chebyshev's inequality tells us that *most* of the points  $Y$  (depicted here as the points in the shaded region) are within a few standard deviations ( $\sigma$ ) of  $\mu$  in the  $\mu - c$  direction. This allows us to bound the distance of these points to  $\mu$  as  $\leq \sqrt{R^2 + (t\sigma)^2}$ .

that is parallel to  $v_\mu$ :  $\overline{(y - \mu_Y)} = \langle y - \mu_Y, v_\mu \rangle v_\mu$ , and the portion that is perpendicular to  $v_\mu$ :  $(y - \mu_Y)^\perp = (y - \mu_Y) - \overline{(y - \mu_Y)}$ . Let  $\overline{(y - c)}$  be the portion of  $y - c$  that is parallel to  $v_\mu$ :  $\overline{(y - c)} = \langle y - c, v_\mu \rangle v_\mu$ . By the choice of  $v_\mu$ , we have that  $\overline{(y - c)} = \overline{(y - \mu_Y)}$ . Thus,

$$\|y - \mu_Y\|_2^2 = \|(y - \mu_Y)^\perp\|_2^2 + \|\overline{(y - \mu_Y)}\|_2^2 = \|(y - c)^\perp\|_2^2 + \|\overline{(y - \mu_Y)}\|_2^2 \leq \|y - c\|_2^2 + (t\sigma)^2.$$

Instead of bounding the number of points that deviate significantly from  $\mu_Y$ , we would like to bound the number of points that deviate significantly from  $\hat{\mu}$ . We observe that a similar argument goes through. We have that

$$\begin{aligned} \mathbf{E}_{y \sim \text{Unif}(Y)} \left[ (y - \hat{\mu})(y - \hat{\mu})^\top \right] &= \mathbf{E}_{y \sim \text{Unif}(Y)} \left[ (y - \mu_Y + \mu_Y - \hat{\mu})(y - \mu_Y + \mu_Y - \hat{\mu})^\top \right] \\ &= \mathbf{E}_{y \sim \text{Unif}(Y)} \left[ (y - \mu_Y)(y - \mu_Y)^\top \right] + \mathbf{E}_{y \sim \text{Unif}(Y)} \left[ (y - \mu_Y) \right] (\mu_Y - \hat{\mu})^\top \\ &\quad + (\mu_Y - \hat{\mu}) \mathbf{E}_{y \sim \text{Unif}(Y)} \left[ (y - \mu_Y)^\top \right] + (\mu_Y - \hat{\mu})(\mu_Y - \hat{\mu})^\top \\ &= \Sigma_Y + (\mu_Y - \hat{\mu})(\mu_Y - \hat{\mu})^\top \\ &\preceq \Sigma_Y + \tau^2 I. \end{aligned} \tag{7}$$

Let  $v_{\hat{\mu}} = \frac{\hat{\mu} - c}{\|\hat{\mu} - c\|}$  be a unit vector in the direction of  $\hat{\mu} - c$ . By (7) we have that

$$\mathbf{P}_{y \sim \text{Unif}(Y)} \left[ |\langle y - \hat{\mu}, v_{\hat{\mu}} \rangle| \geq t\sqrt{\sigma^2 + \tau^2} \right] \leq \frac{1}{t^2}.$$

Thus there exists a subset  $\hat{Y} \subseteq Y$  such that  $|\hat{Y}| \geq (1 - \frac{1}{t^2})|Y|$ , and for all  $y \in \hat{Y}$  we have  $|\langle y - \hat{\mu}, v_{\hat{\mu}} \rangle| \leq t\sqrt{\sigma^2 + \tau^2}$ . For each  $y \in \hat{Y}$ , we can decompose  $y - \hat{\mu}$  into the portion that is

parallel to  $v_{\hat{\mu}}$ :  $\overline{(y - \hat{\mu})} = \langle y - \hat{\mu}, v_{\hat{\mu}} \rangle v_{\hat{\mu}}$ , and the portion that is perpendicular to  $v_{\hat{\mu}}$ :  $(y - \hat{\mu})^\perp = (y - \hat{\mu}) - \overline{(y - \hat{\mu})}$ . Let  $\overline{(y - c)}$  be the portion of  $y - c$  that is parallel to  $v_{\hat{\mu}}$ :  $\overline{(y - c)} = \langle y - c, v_{\hat{\mu}} \rangle v_{\hat{\mu}}$ . By the choice of  $v_{\hat{\mu}}$ , we have that  $\overline{(y - c)} = \overline{(y - \hat{\mu})}$ . Thus we have

$$\|y - \hat{\mu}\|_2^2 = \|(y - \hat{\mu})^\perp\|_2^2 + \|\overline{(y - \hat{\mu})}\|_2^2 = \|(y - c)^\perp\|_2^2 + \|\overline{(y - \hat{\mu})}\|_2^2 \leq \|y - c\|_2^2 + t^2(\sigma^2 + \tau^2).$$

■

## 2.2. Reducing Worst-Case Improper Learning to Non-Worst-Case Proper Learning

Our goal is to use Lemma 6 to get a guarantee for learning a small ellipsoid, that does not depend on the variance of the points. As a stepping stone to our population statement, we first prove the following statement about the sample points.

**Theorem 7 (Finding small volume ellipsoid for  $n$  points)** *Suppose we are given a set of points  $Y \subseteq \mathbb{R}^d$ ,  $|Y| = n$ , and  $0 \leq \delta \leq 1$ ,  $0 \leq \gamma \leq 1$ , such that there exists a subset  $Y^* \subseteq Y$ ,  $|Y^*| \geq \delta|Y|$ , that is contained in an unknown ball  $B^* = B(c^*, R^*)$ . Then we can find an ellipsoid  $\hat{E}$  such that  $|\hat{E} \cap Y| \geq \delta(1 - \gamma)|Y|$ , and*

$$\text{vol}(\hat{E})^{1/d} \leq \text{vol}(B^*)^{1/d} \cdot \left(1 + O\left(d^{-1/2+o(1)}/\gamma\delta\right)\right).$$

Intuitively, the case where our non-worst-case proper learning algorithm fails is when the points have most of their variance concentrated in only a few directions, see for example Figure 2. This makes it hard to estimate the position of the center accurately in the large variance directions. However, this also means that the variance of the points in the other directions must be low, so it is *easier* to estimate the position of the center in these directions. Applying the correct linear transformation to the points essentially allows us to “reweigh” the accuracy that we are aiming for across the different directions, and trade off error in the high variance directions for accuracy in the low variance directions.

Our strategy is to find a good preconditioner  $\widehat{M}^{-1/2}$ , for the points in  $Y$ , such that the transformed points have small variance. (Writing the preconditioner as  $\widehat{M}^{-1/2}$  is convenient, as we can then argue about the distortion in terms of the eigenvalues of  $\widehat{M}$ , which we will choose to be positive definite.) Then, we can apply our procedure to find an approximately minimum volume ball on the transformed points, which will return a strong approximation since the variance of the points is controlled (via Lemma 6). Transforming this ball back into the original space will result in a confidence region that is shaped as an ellipsoid.

Our goal is to choose  $\widehat{M}$  so that the largest variance of the transformed points in any one direction is comparable to the *average* variance across all the directions (i.e., the variance is well-spread, and the distribution is “approximately isotropic”). We choose  $\widehat{M}$  to have eigenvectors aligned with those of the covariance  $\Sigma_{Y'}$  of the points  $Y'$ , where  $Y'$  is some subset of  $Y$  contained in a ball of radius  $R$ , so that the average variance across the  $d$  directions is at most  $R^2/d$ . We choose  $\widehat{M}$  to have the same eigenvectors as  $\Sigma_{Y'}$ , with eigenvalues  $a_i^2$  of  $\widehat{M}$  satisfying  $a_i^2 = d$  if  $\lambda_i \geq \hat{\tau}^2 R^2/d$  and  $a_i^2 = 1$  if  $\lambda_i < \hat{\tau}^2 R^2/d$ , where  $\hat{\tau}$  is a parameter that we choose later.

Essentially, this means that  $\widehat{M}^{-1/2}$  will shrink the large variance directions by a factor  $\sqrt{d}$ , ensuring that the resulting variance in those directions is constant (depending only on  $R$ ). It does nothing for the small variance directions. This choice of  $\widehat{M}$  has the following nice properties.

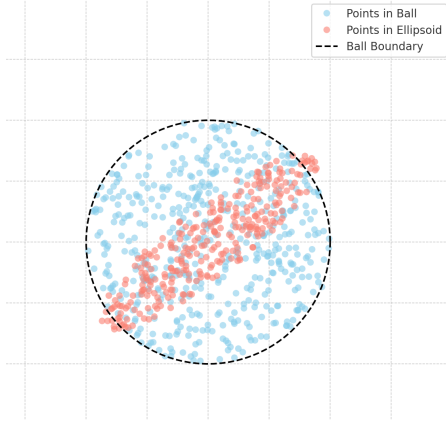


Figure 2: The figure shows a ball  $B^*$  of radius  $R^* = \sqrt{d}$  containing a set of points, whose the covariance matrix  $\Sigma_{Y^*}$  is not isotropic and has some directions of high variance. In this case, we can try to find a smaller ellipsoid containing many of the points.

- (i)  $\widehat{M} \succcurlyeq I$ , and therefore  $\widehat{M}^{-1/2}$  is “shrinking” (non-expanding) in all directions. This ensures that there exists a ball in the transformed space that contains  $Y^*$  and has volume at most that of  $B^*$  in the original space.
- (ii) We have that the variance of the points  $Y$  in the transformed space is bounded, i.e., the points in the transformed space are approximately isotropic:

$$\widehat{M}^{-1/2} \Sigma_Y \widehat{M}^{-1/2} \preccurlyeq O\left(\frac{\widehat{\tau}^2 R^2}{d}\right) I_d.$$

This allows us to bound the approximation ratio of applying Lemma 6 to the transformed points.

- (iii) At most  $d/\widehat{\tau}^2$  eigenvalues of  $\widehat{M}$  are set to  $d$ . This bounds the distortion of  $\widehat{M}$ . That is, for any shape  $S$ ,

$$\frac{\text{vol}(S)}{\text{vol}(\widehat{M}^{-1/2} S)} \leq d^{d/\widehat{\tau}^2} \leq \exp\left(\frac{d \ln d}{\widehat{\tau}^2}\right).$$

Properties (i) and (ii) allow us to reduce to the non-worst-case setting, and property (iii) ensures that this translation does not lose too much in volume. We will set  $\widehat{\tau} = d^{1/4}$  to trade-off the different losses from (ii) and (iii) and obtain the final guarantee. The full proof of this theorem can be found in Appendix B.

We remark that our main result (Theorem 1) gives a volume approximation guarantee of  $\exp(d^{1/2+o(1)})$  in the worst-case. At the same time, Corollary 4 gives a  $O(1)$  factor approximation guarantee in volume, when the distribution of points inside the optimal ball  $B^*$  is approximately isotropic. We leave as an open question whether or not there exists a polynomial time (improper) learning algorithm that obtains an  $O(1)$  volume approximation in the worst-case.



## Acknowledgments

This research project was supported by NSF-funded Institute for Data, Econometrics, Algorithms and Learning (IDEAL) through the grants NSF ECCS-2216970 and ECCS-2216912. The research started as part of the IDEAL special program on Reliable and Robust Data Science. Chao Gao was also supported in part by NSF Grants DMS-2310769, NSF CAREER Award DMS-1847590, and an Alfred Sloan fellowship. Vaidehi Srinivas was supported by the Northwestern Presidential Fellowship. We gratefully acknowledge the support of the NSF-Simons AI-Institute for the Sky (SkAI) via grants NSF AST-2421845 and Simons Foundation MPS-AI-00010513, and the support of the NSF-Simons National Institute for Theory and Mathematics in Biology (NITMB) via grants NSF DMS-2235451 and Simons Foundation MP-TMPS-00005320. We also thank an anonymous reviewer for helpful pointers to previous work involving coresets.

## References

- Pankaj Agarwal, Sariel Har Peled, and Kasturi R. Varadarajan. Geometric approximation via core-sets. *Combinatorial and Computational Geometry*, 52, 11 2004.
- Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.*, 16(4):494–591, March 2023. ISSN 1935-8237. doi: 10.1561/2200000101. URL <https://doi.org/10.1561/2200000101>.
- Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024.
- Mihai Badoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *Proceedings of the Thiry-Fourth Annual ACM Symposium on Theory of Computing*, STOC ’02, page 250–257, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 1581134959. doi: 10.1145/509907.509947. URL <https://doi.org/10.1145/509907.509947>.
- Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.
- Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, page 47–60, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450345286. doi: 10.1145/3055399.3055491. URL <https://doi.org/10.1145/3055399.3055491>.
- M-Y Cheng and Peter Hall. Calibrating the excess mass and dip tests of modality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):579–589, 1998.
- Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2001.
- Ilias Diakonikolas and Daniel M. Kane. *Algorithmic High-Dimensional Robust Statistics*. Cambridge University Press, 2023.

- Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 655–664. IEEE, 2016.
- Ilias Diakonikolas, Daniel Kane, Daniel Kongsgaard, Jerry Li, and Kevin Tian. List-decodable mean estimation in nearly-pca time. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10195–10208. Curran Associates, Inc., 2021. URL [https://proceedings.neurips.cc/paper\\_files/paper/2021/file/547b85f3fafdf30856386753dc21c4e1-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2021/file/547b85f3fafdf30856386753dc21c4e1-Paper.pdf).
- Hu Ding. A Sub-Linear Time Framework for Geometric Optimization with Outliers in High Dimensions. In Fabrizio Grandoni, Grzegorz Herman, and Peter Sanders, editors, *28th Annual European Symposium on Algorithms (ESA 2020)*, volume 173 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 38:1–38:21, Dagstuhl, Germany, 2020. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-162-7. doi: 10.4230/LIPIcs.ESA.2020.38. URL <https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ESA.2020.38>.
- John HJ Einmahl and David M Mason. Generalized quantile processes. *The Annals of Statistics*, pages 1062–1078, 1992.
- Uriel Feige. Vertex cover is hardest to approximate on regular graphs. 11 2003.
- Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. On agnostic learning of parities, monomials, and halfspaces. *SIAM Journal on Computing*, 39(2):606–645, 2009. doi: 10.1137/070684914. URL <https://doi.org/10.1137/070684914>.
- Herbert Fleischner, Gert Sabidussi, and Vladimir I Sarvanov. Maximum independent sets in 3-and 4-regular hamiltonian graphs. *Discrete mathematics*, 310(20):2742–2749, 2010.
- A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI’98, page 148–155, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 155860555X.
- Chao Gao, Liren Shan, Vaidehi Srinivas, and Aravindan Vijayaraghavan. Volume optimality in conformal prediction with structured prediction sets, 2025. URL <https://arxiv.org/abs/2502.16658>.
- Javier Nuñez Garcia, Zoltan Katalik, Kwang-Hyun Cho, and Olaf Wolkenhauer. Level sets and minimum volume sets of probability density functions. *International journal of approximate reasoning*, 34(1):25–47, 2003.
- Chirag Gupta, Arun K Kuchibhotla, and Aaditya Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, 2022.
- John A Hartigan. Estimation of a convex density contour in two dimensions. *Journal of the American Statistical Association*, 82(397):267–270, 1987.

- Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.
- P.J. Huber. *Robust Statistics*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley, 2004. ISBN 9780471650720. URL <https://books.google.com/books?id=e62RhdqIdMkC>.
- Rob J Hyndman. Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126, 1996.
- Subhash Khot. On the power of unique 2-prover 1-round games. In *Proceedings of the Thiry-Fourth Annual ACM Symposium on Theory of Computing*, STOC ’02, page 767–775, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 1581134959. doi: 10.1145/509907.510017. URL <https://doi.org/10.1145/509907.510017>.
- Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE, 2016.
- Jing Lei, James M. Robins, and Larry A. Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108:278 – 287, 2013. URL <https://api.semanticscholar.org/CorpusID:17499892>.
- Dietrich W Müller and Günther Sawitzki. Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, 86(415):738–746, 1991.
- Deborah Nolan. The excess-mass ellipsoid. *Journal of multivariate analysis*, 39(2):348–371, 1991.
- Chiwoo Park, Jianhua Z Huang, and Yu Ding. A computable plug-in estimator of minimum volume sets for novelty detection. *Operations Research*, 58(5):1469–1480, 2010.
- Wolfgang Polonik. Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *The annals of Statistics*, pages 855–881, 1995.
- Wolfgang Polonik. Minimum volume sets and generalized quantile processes. *Stochastic processes and their applications*, 69(1):1–24, 1997.
- Wolfgang Polonik. Concentration and goodness-of-fit in higher dimensions:(asymptotically) distribution-free methods. *The Annals of Statistics*, 27(4):1210–1229, 1999.
- Prasad Raghavendra and David Steurer. Graph expansion and the unique games conjecture. In *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, STOC ’10, page 755–764, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450300506. doi: 10.1145/1806689.1806792. URL <https://doi.org/10.1145/1806689.1806792>.
- Prasad Raghavendra, David Steurer, and Madhur Tulsiani. Reductions between expansion problems. In *2012 IEEE 27th Conference on Computational Complexity*, pages 64–73, 2012. doi: 10.1109/CCC.2012.43.

- Philippe Rigollet and Régis Vert. Optimal rates for plug-in estimators of density level sets. *Bernoulli*, pages 1154–1178, 2009.
- P Rousseeuw. Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications B*, 1985.
- Peter J Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984.
- Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.
- Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.
- Thomas W Sager. Estimation of a multivariate mode. *The Annals of Statistics*, 6(4):802–812, 1978.
- Thomas W Sager. An iterative method for estimating a multivariate mode and isopleth. *Journal of the American Statistical Association*, 74(366a):329–339, 1979.
- Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.
- Clayton Scott and Robert Nowak. Learning minimum volume sets. *Advances in neural information processing systems*, 18, 2005.
- Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014. ISBN 1107057132.
- Vladimir Shenmaier. Complexity and approximation of the smallest k-enclosing ball problem. *European Journal of Combinatorics*, 48:81–87, 2015. ISSN 0195-6698. doi: <https://doi.org/10.1016/j.ejc.2015.02.011>. URL <https://www.sciencedirect.com/science/article/pii/S0195669815000335>. Selected Papers of EuroComb’13.
- Jacob Steinhardt. Lecture notes for stat260 (robust statistics). 2019.
- Ingo Steinwart, Don Hush, and Clint Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6(2), 2005.
- David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54:45–66, 2004.
- Alexandre B Tsybakov. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3):948–969, 1997.
- Stefan Van Aelst and Peter Rousseeuw. Minimum volume ellipsoid. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):71–82, 2009.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387001522.

Zhendong Wang, Ruijiang Gao, Mingzhang Yin, Mingyuan Zhou, and David Blei. Probabilistic conformal prediction using conditional random samples. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 8814–8836. PMLR, 25–27 Apr 2023. URL <https://proceedings.mlr.press/v206/wang23n.html>.

Minxing Zheng and Shixiang Zhu. Generative conformal prediction with vectorized non-conformity scores, 2025. URL <https://arxiv.org/abs/2410.13735>.

**Contents**

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Our Results . . . . .	3
1.2	Proper Learning: Separation Result and Algorithms . . . . .	4
1.3	Application to Conformal Prediction . . . . .	6
1.4	Prior and Related Algorithmic Approaches . . . . .	7
1.5	Applications in Statistics . . . . .	9
<b>2</b>	<b>Technical Overview</b>	<b>10</b>
2.1	Proper Learning for Non-Worst-Case Instances . . . . .	11
2.2	Reducing Worst-Case Improper Learning to Non-Worst-Case Proper Learning . . .	13
<b>A</b>	<b>Proper Learning of Euclidean Balls</b>	<b>21</b>
A.1	Completing the Proof of Theorem 8 . . . . .	25
A.2	Better volume approximation guarantee for almost isotropic distributions . . . . .	28
<b>B</b>	<b>Finding Confidence Sets <math>(\Gamma = 1 + d^{-1/2+o(1)})</math>-Competitive with Euclidean Balls</b>	<b>30</b>
<b>C</b>	<b>Greedy Algorithm and Unions of Sets</b>	<b>33</b>
<b>D</b>	<b>Hardness of Proper Learning</b>	<b>38</b>
D.1	NP-hardness of Proper Learning . . . . .	38
D.2	Computational Intractability even with Slack in Coverage . . . . .	39
<b>E</b>	<b>Application to Conformal Prediction</b>	<b>41</b>



## Appendix A. Proper Learning of Euclidean Balls

The main technical insight that allows us to get a proper learning algorithm for Euclidean balls in a non-worst-case setting is Lemma 6, which tells us that for a set of points with bounded variance, the mean of the points is a good proxy for the center (up to a slack factor in coverage).

We use this result as a subroutine in our main result for improper learning. In this section we provide some additional results that extend our result for proper learning to other settings. In particular, Theorem 8 gives a proper learning algorithm for *fully worst-case* instances, and achieves an approximation ratio of  $\left(1 + \tilde{O}_{\gamma,\delta}\left(\frac{1}{\log d}\right)\right)$  in radius, which is comparable to the factor of  $\left(1 + \tilde{O}_{\gamma,\delta}\left(\frac{1}{\log d}\right)\right)$  due to (Badoiu et al., 2002; Ding, 2020). While we do achieve a slightly weaker approximation guarantee in this setting, our techniques are significantly different than the coresot based techniques of the prior work, and extend well to non-worst-case settings, as evidenced by our improper learning result. We also provide Theorem 4, which shows a way to achieve a better approximation ratio, assuming non-worst-case properties only on the *inlier* points.

**Theorem 8 (Finding approximate minimum volume  $\delta$ -coverage ball)** *Given a set of points  $Y \subseteq \mathbb{R}^d$ ,  $|Y| = n$ , and  $0 \leq \delta \leq 1$ ,  $0 \leq \gamma \leq 1$ , such that there exists a subset  $Y^* \subseteq Y$ ,  $|Y^*| \geq \delta|Y|$ , that is contained in an unknown ball  $B^* = B(c^*, R^*)$ , we can find a ball  $\hat{B} = B(\hat{c}, \hat{R})$  such that  $|Y \cap \hat{B}| \geq (1 - \gamma)|Y^*|$ , and*

$$\text{vol}(\hat{B})^{1/d} \leq \text{vol}(B^*)^{1/d} \left(1 + O\left(\frac{\log \log d}{\gamma \delta \log d}\right)\right)$$

*in polynomial time in  $n, d$ .*

The algorithm for finding this small volume coverage ball is given in Figure 3. By using the fact that the VC-dimension of Euclidean balls in  $d$ -dimensions is at most  $d + 1$ , we immediately get the following corollary.

**Corollary 9 (Population version of Theorem 8)** *Suppose  $Y \subset \mathbb{R}^d$  is a set of  $n$  i.i.d. samples generated from an arbitrary distribution  $\mathcal{D}$ . For a target coverage  $\delta \in (0, 1)$  and a coverage approximation factor  $\gamma \in (0, 1)$  such that  $n = \Omega(d)$ , in  $\text{poly}(n, d)$  time, with high probability, we can find a ball  $\hat{B}$  such that with high probability,*

$$\mathbf{P}_{y \sim \mathcal{D}} [y \in \hat{B} \mid Y] \geq \delta,$$

*and*

$$\text{vol}(\hat{B})^{1/d} \leq \text{vol}(B^*)^{1/d} \left(1 + O\left(\frac{\log \log d}{\gamma \delta \log d}\right)\right)$$

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

Our approach starts with the following structural observation about the true high-density ball  $B^*$ . It is not in general true that the mean  $\mu^*$  of the points  $Y^* = B^* \cap Y$  should be near the center  $c^*$  of  $B^*$ . However, it turns out that if the mean is far from the center, taking a ball centered at the mean of  $B^*$  must still capture *most* of the points in  $Y^*$ . This is a simple consequence of Chebyshev's inequality— if we are able to control the variance of the points in  $Y^*$  (Lemma 6, Figure 1). Thus

our problem reduces to the following. Given only the full set of points  $Y$ , we must estimate the mean of  $Y^*$ , while also controlling the variance of the points that we are taking the mean of.

The following simple lemma from robust statistics shows that the empirical mean of a set of points with bounded covariance that contains both inliers and outliers is close to the true mean.

**Lemma 10 (High probability event has close-by mean)** (see e.g., Lemma 2.1 of [Steinhardt, 2019](#)) *Let  $X \in \mathbb{R}^d$  be a random variable with mean  $\mu \in \mathbb{R}^d$  and covariance matrix  $\Sigma$ , where the largest eigenvalue satisfies  $\lambda_{\max}(\Sigma) \leq \sigma^2$ . Then for any event  $E$  with probability at least  $\delta > 0$ , the conditional expectation satisfies:*

$$\|\mathbb{E}[X|E] - \mu\|_2 \leq \sigma \cdot \sqrt{2(1-\delta)/\delta}.$$

The following simple claim shows that the variance does not increase much when we condition on an event with non-negligible probability. [High probability event has similar variance]

Let  $X \in \mathbb{R}^d$  be a random variable whose maximum variance in any direction at most  $\sigma^2$ , and let  $E$  be an event that occurs with probability at least  $\delta$ . Then the maximum variance in any direction of  $X$  conditioned on  $E$  is at most  $\frac{\sigma^2}{\delta}$ .

**Proof** This follows from a simple averaging argument. Let  $\mu$  be the mean of  $X$ . We have that the variance of  $X$  is at most  $\sigma^2$ . That is, for all unit vectors  $v$ ,

$$\mathbf{E} [\langle X - \mu, v \rangle^2] \leq \sigma^2.$$

Now, we can decompose this into the variance conditioned on  $E$ ,

$$\begin{aligned} \sigma^2 &\geq \mathbf{E} [\langle X - \mu, v \rangle^2] \\ &\geq \mathbf{E} [\langle X - \mu, v \rangle^2 | E] \mathbf{P}[E] + \mathbf{E} [\langle X - \mu, v \rangle^2 | \neg E] \mathbf{P}[\neg E] \geq \mathbf{E} [\langle X - \mu, v \rangle^2 | E] \delta \\ \frac{\sigma^2}{\delta} &\geq \mathbf{E} [\langle X - \mu_E, v \rangle^2 | E], \end{aligned}$$

where in the last line  $\mu_E$  is the mean of  $X$  conditioned on  $E$ , and the last line follows because the mean minimizes the squared deviation. Thus we can conclude that the variance of  $X$  given  $E$  is bounded by  $\frac{\sigma^2}{\delta}$  in every direction.  $\blacksquare$

To use the above observation to find an approximate center, we need to do two things. First, we need to control the standard deviation of the points that we are considering. In particular, to achieve a subconstant approximation of radius, we need that the standard deviation of our points is  $\frac{1}{\omega(1)} R^*$ . Second, we need to be able to find  $\mu^*$  approximately, which is akin to robust mean estimation.

In the first step, we use a simple search procedure that gives us a coarse estimate of  $B^*$ . That is, we find a ball  $B'$  such that  $Y^* \subseteq B'$  and the radius  $R'$  of  $B'$  is bounded by  $R^* \leq R' \leq 2R^*$ . In the worst case, we cannot expect that the variance is less than  $(R')^2$ , for example if all of the variance of  $Y'$  is concentrated in one direction. However, there can be at most  $\log d$  directions with variance higher than  $(R')^2 / \log d$ . Thus we can find a list of candidate centers for  $Y^*$  by estimating the location of the center separately in the high variance space and the low variance space. In the high variance space, we can grid search to find a list of potential candidates for the center. This is tractable since the high variance space has dimension  $\log d$ . In the low variance space we can estimate the location of  $\mu^*$ .

**Lemma 11 (Grid search)** Suppose we are given a ball  $B' = B(0, R')$  containing a set of points  $Y' \in \mathbb{R}^q$ , such that there exists a subset of points  $Y^* \in Y'$  that are contained in a smaller ball  $B^* = B(c^*, R^*)$  with  $c^* \in B'$ , where  $|Y^*| \geq \delta|Y'|$ .

Then for a tolerance  $0 < \tau < R'$ , there exists a list  $L$  of size  $\left(\frac{2+\tau/R'}{\tau/R'}\right)^q$ , such that  $L$  contains a point  $\hat{c} \in \mathbb{R}^q$  such that for every  $y \in Y^*$ ,

$$\|y - \hat{c}\| \leq \|y - c^*\| + \tau.$$

**Proof** We first construct a  $\frac{\tau}{R'}$ -net for the unit ball in  $\mathbb{R}^q$ , denoted as  $L_0$ , with size  $\left(\frac{2+\tau/R'}{\tau/R'}\right)^q$ . This ensures that every point in the unit ball is at most distance  $\frac{\tau}{R'}$  from some point in the net  $L_0$ . Next, we scale the net  $L_0$  to  $L = R'L_0$ . This ensures that every point in  $B' = B(0, R')$  is at most distance  $\tau$  from some point in  $L$ . Thus, for the center  $c^* \in B'$ , there is a point  $\hat{c} \in L$  such that  $\|\hat{c} - c^*\| \leq \tau$ . By the triangle inequality, for every  $y \in Y^*$ , we have  $\|y - \hat{c}\| \leq \|y - c^*\| + \|\hat{c} - c^*\| \leq \|y - c^*\| + \tau$ . ■

**Lemma 12 (Coarse-to-fine refinement)** Suppose we are given a ball  $B' = B(0, R')$  containing a set of points  $Y' \subseteq \mathbb{R}^d$ , where the  $(q+1)$ -th largest eigenvalue of the covariance matrix  $\Sigma_{Y'}$  of  $Y'$ , denoted as  $\lambda_{q+1}(\Sigma_{Y'})$ , is bounded by

$$\lambda_{(q+1)}(\Sigma_{Y'}) \leq \sigma^2.$$

Suppose there exists a subset of points  $Y^* \subseteq Y'$  are contained in a smaller ball  $B^* = B(c^*, R^*)$ , where  $|Y^*| \geq \delta|Y'|$  and  $R' \leq 2R^*$ . Let  $\gamma > 0$  be a coverage slack factor.

Then there exists a list  $C$  of size at most  $\left(\frac{3R'}{\tau}\right)^q$  and a set  $\hat{Y} \subseteq Y^*$  with  $|\hat{Y}| \geq (1 - \gamma)|Y^*|$ , such that there exists a point  $\hat{c} \in C$  which for every  $y \in \hat{Y}$  satisfies

$$\begin{aligned} \|y - \hat{c}\|_2^2 &\leq \|\Pi_{\text{low}}(y - c^*)\|_2^2 + \frac{\sigma^2}{\gamma} \left(1 + \frac{2(1 - \delta)}{\delta}\right) + (\|\Pi_{\text{high}}(y - c^*)\|_2 + \tau)^2 \\ &\leq \left(\|y - c^*\|_2 + O\left(\frac{\sigma}{\sqrt{\gamma\delta}}\right) + \tau\right)^2, \end{aligned}$$

where  $\Pi_{\text{high}}$  is the projection onto the top- $q$  eigenspace of  $\Sigma_{Y'}$  and  $\Pi_{\text{low}}$  is the projection onto the space orthogonal to the top- $q$  eigenspace of  $\Sigma_{Y'}$ . For  $q = 0$ , we get a single estimate  $\hat{c}$ , and  $\hat{Y} \subseteq Y^*$  with  $|\hat{Y}| \geq (1 - \gamma)|Y^*|$ , that for every  $y \in \hat{Y}$  satisfies

$$\|y - \hat{c}\|_2^2 \leq \|y - c^*\|_2^2 + \frac{\sigma^2}{\gamma} \left(1 + \frac{2(1 - \delta)}{\delta}\right).$$

**Proof** We start by considering  $\Sigma_{Y'}$ , the covariance matrix of  $Y'$  and partition its eigenvalues into two groups. The directions associated with the largest  $q$  eigenvalues of  $\Sigma_{Y'}$  are referred to as the “high variance directions”; while the directions corresponding to the remaining eigenvalues are the “low variance directions”. Let  $\Pi_{\text{high}} : \mathbb{R}^d \rightarrow \mathbb{R}^q$  denote the projection onto the “high variance”

eigenspace and  $\Pi_{\text{low}} : \mathbb{R}^d \rightarrow \mathbb{R}^{d-q}$  denote the projection onto the “low variance” eigenspace. We estimate the mean  $\mu^*$  of  $Y^*$  separately in the high variance and low variance directions.

First, if  $q > 0$ , we estimate the location of  $\mu^*$  in the high variance directions. Let  $Y'_{\text{high}} = \Pi_{\text{high}} Y'$  denote the projection of points in  $Y'$  onto the high variance directions. Note that the set  $Y'_{\text{high}}$  lies within the  $q$ -dimensional ball  $B'_{\text{high}} = \Pi_{\text{high}} B'$ , which is centered at the origin and has a radius of at most  $R'$ . Next, consider the projection of  $Y^*$  onto the high variance directions, denoted by  $Y^*_{\text{high}} = \Pi_{\text{high}} Y^*$ . Then, this set  $Y^*_{\text{high}}$  is contained in the ball  $B^*_{\text{high}} = \Pi_{\text{high}} B^*$  centered at  $\Pi_{\text{high}} c^*$  with radius at most  $R^*$ . Since  $B^*_{\text{high}} \subseteq B'_{\text{high}}$ , and  $\mu^* \in B'_{\text{high}}$  as well, by Lemma 11, there exists a list  $L \subset \mathbb{R}^q$  with size  $|L| \leq \left(\frac{2+\tau/R'}{\tau/R'}\right)^q \leq \left(\frac{3R'}{\tau}\right)^q$  such that there is a point  $\hat{\ell} \in L$  such that for every point  $y \in Y^*$ ,

$$\|\Pi_{\text{high}} y - \hat{\ell}\| \leq \|\Pi_{\text{high}} y - \Pi_{\text{high}} c^*\| + \tau. \quad (8)$$

Next, we estimate the location of  $\mu^*$  in the low variance directions. Let  $Y'_{\text{low}} = \Pi_{\text{low}} Y'$  denote the projection of  $Y'$  onto the low variance directions. By construction, the set  $Y'_{\text{low}}$  has variance at most  $\lambda_{(q+1)}(\Sigma_{Y'}) \leq \sigma^2$  in every direction. Since the set  $Y^*$  contains at least a  $\delta$ -fraction of the points in  $Y'$ , by Appendix A, we have  $Y^*_{\text{low}} = \Pi_{\text{low}} Y^*$  has variance at most  $\sigma^2/\delta$  in every direction. By Lemma 10, the mean  $\mu^*_{\text{low}}$  of  $Y^*_{\text{low}}$  is close to the mean  $\mu'_{\text{low}}$  of  $Y'_{\text{low}}$ , i.e.,

$$\|\mu^*_{\text{low}} - \mu'_{\text{low}}\|_2 \leq \sigma \cdot \sqrt{2(1-\delta)/\delta}. \quad (9)$$

Since the variance of  $Y^*_{\text{low}}$  is bounded by  $\sigma^2/\delta$ , and  $\mu'_{\text{low}}$  is close to  $\mu^*_{\text{low}}$ , Lemma 6 tells us that, for a fixed choice of  $\gamma$ , there exists a subset  $\hat{Y} \subseteq Y^*$  such that  $|\hat{Y}| \geq (1-\gamma)|Y^*|$ , and for every point  $y \in \hat{Y}$ ,

$$\|\Pi_{\text{low}} y - \mu'_{\text{low}}\|_2^2 \leq \|\Pi_{\text{low}}(y - c^*)\|_2^2 + \frac{\sigma^2}{\gamma} \left(1 + \frac{2(1-\delta)}{\delta}\right).$$

For  $q = 0$ , this gives the final bound.

For  $q > 0$ , we take  $\hat{c}$  such that  $\Pi_{\text{high}} \hat{c} = \hat{\ell}$  and  $\Pi_{\text{low}} \hat{c} = \mu'_{\text{low}}$ . Putting the bounds together, we have that there exists a set  $\hat{Y} \subseteq Y^*$ , with  $|\hat{Y}| \geq (1-\gamma)|Y^*|$ , such that for every  $y \in \hat{Y}$ ,

$$\begin{aligned} \|y - \hat{c}\|_2^2 &= \|\Pi_{\text{low}}(y - \hat{c})\|_2^2 + \|\Pi_{\text{high}}(y - \hat{c})\|_2^2 \\ &\leq \|\Pi_{\text{low}}(y - c^*)\|_2^2 + \frac{\sigma^2}{\gamma} \left(1 + \frac{2(1-\delta)}{\delta}\right) + (\|\Pi_{\text{high}}(y - c^*)\|_2 + \tau)^2 \\ &\leq \left(\|y - c^*\|_2 + O\left(\frac{\sigma}{\sqrt{\gamma\delta}}\right) + \tau\right)^2, \end{aligned}$$

where both the tighter bound (second line) and looser bound (third line) will be useful in applications of the lemma. ■

**Lemma 13 (Most directions are low variance)** *Let  $Y$  be a set of points contained in  $B(0, R)$  with mean  $\mu_Y$ . The covariance of  $Y$  can have at most  $q$  eigenvalues greater than  $\frac{R^2}{q}$ .*

**Proof** Denote the covariance of  $Y$  by  $\Sigma_Y = \frac{1}{|Y|} \sum_{y \in Y} (y - \mu_Y)(y - \mu_Y)^\top$ , where  $\mu_Y$  is the mean of the points  $Y$ . We have that

$$\Sigma_Y \preceq \frac{1}{|Y|} \sum_{y \in Y} yy^\top,$$

since the mean minimizes the deviation. Since trace follows the PSD domination, we have that

$$\text{tr}(\Sigma_Y) \leq \text{tr} \left( \frac{1}{|Y|} \sum_{y \in Y} yy^\top \right).$$

Since trace is a linear operator and  $Y$  is contained in  $B(0, R)$ , we have that

$$\text{tr} \left( \frac{1}{|Y|} \sum_{y \in Y} yy^\top \right) = \frac{1}{|Y|} \sum_{y \in Y} \text{tr}(yy^\top) = \frac{1}{|Y|} \sum_{y \in Y} \|y\|_2^2 \leq R^2.$$

So we have that  $\text{tr}(\Sigma_Y) \leq R^2$ . Since the trace is equal to the sum of the eigenvalues, and  $\Sigma_Y$  is positive semidefinite and thus has all non-negative eigenvalues, we have by an averaging argument that  $\Sigma_Y$  can have at most  $q$  eigenvalues  $\geq \frac{R^2}{q}$ .  $\blacksquare$

### A.1. Completing the Proof of Theorem 8

**Proof** [Proof of Theorem 8] We begin by finding a list of coarse guesses for  $B^*$ . We create a list of balls

$$\mathfrak{B} = \{B(y_1, \|y_1 - y_2\|) \mid y_1, y_2 \in Y\}.$$

Note that  $|\mathfrak{B}| \leq n^2$  since  $|Y| = n$ . Consider the subset  $Y^* \subseteq Y$  contained in ball  $B^* = B(c^*, R^*)$ . Choosing

$$y_1^*, y_2^* = \underset{y_1, y_2 \in Y^*}{\text{argmax}} \|y_1 - y_2\|,$$

gives us that the ball

$$B' = B(y_1^*, \|y_1^* - y_2^*\|) \tag{10}$$

contains all points in  $Y^*$ , has radius  $R' \leq 2R^*$ , and  $B' \in \mathfrak{B}$ .

We now do a filtering step to get a coarse estimate of  $R^*$ , and remove any balls from  $\mathfrak{B}$  that are too large. In particular, let  $R_{\min}$  be the minimum radius among balls in  $\mathfrak{B}$  that contain at least  $\delta n$  points. Since  $B^*$  is the volume optimal ball that contains at least  $\delta$ -fraction of  $Y$ , we have  $R_{\min} \geq R^*$ . Thus, we know that

$$R^* \leq R_{\min} \leq 2R^*.$$

We filter  $\mathfrak{B}$  to only contain balls that contain at least  $\delta n$  points from  $Y$ , and have radius at most  $2R_{\min} \leq 4R^*$ . Finding  $R_{\min}$  and filtering the list take  $\tilde{O}(n^3)$  time.

Now we apply the coarse-to-fine refinement of Lemma 12 to each ball in (the filtered)  $\mathfrak{B}$ , to generate a list of candidate centers for  $\hat{B}$ . In particular, we apply the procedure for  $q = \frac{\log d}{\log \log d}$  and  $\tau = \frac{R_{\min}}{\log d}$ . For each ball  $B$  in the filtered  $\mathfrak{B}$ , we have that the radius  $R$  of  $B$  is at most  $2R_{\min}$ . Thus, by Lemma 12, the size of the list of potential centers  $C_B$  that we generate for  $B$  is at most

$$|C_B| \leq \left( \frac{3R'}{\tau} \right)^q \leq (6 \log d)^{\frac{\log d}{\log \log d}} \leq d^3.$$

**Algorithm DENSE\_BALL**

**Input:** a set of point  $Y \in \mathbb{R}^d$ , target fraction  $\delta \in (0, 1)$ , and slack  $\gamma \in (0, 1)$

**Output:** an ball  $\hat{B} \subset \mathbb{R}^d$

1. Create a list of coarse balls  $\mathfrak{B} = \{B(y_1, \|y_1 - y_2\|) \mid y_1, y_2 \in Y\}$  and remove from  $\mathfrak{B}$  all balls that contain less than  $\delta n$  points in  $Y$ .
2. Set  $R_{\min}$  to be the minimum radius among all balls in  $\mathfrak{B}$ . Then remove from  $\mathfrak{B}$  all balls with a radius greater than  $2R_{\min}$ .
3. Create a list of refined balls  $\mathfrak{B}_2$  as follows:
  - (a) Create a list of candidate centers  $C_{B'}$  for each ball  $B'$  in  $\mathfrak{B}$  as follows.
  - (b) Compute the eigenvalues  $\lambda_1, \dots, \lambda_d$  and eigenvectors  $v_1, \dots, v_d$  of the covariance  $\Sigma_{Y'}$  of points  $Y' = Y \cap B'$ .
  - (c) Set  $\Pi_{\text{high}}$  be the projection onto the eigenspace of  $\Sigma_{Y'}$  with eigenvalues greater than  $(R')^2 \log \log d / \log d$  and  $\Pi_{\text{low}} = I - \Pi_{\text{high}}$ . Set  $B'_{\text{high}} = \Pi_{\text{high}} B'$  and  $Y'_{\text{low}} = \Pi_{\text{low}} Y'$ . Note that there are at most  $q = \log d / \log \log d$  eigenvalues of  $\Sigma_{Y'}$  is greater than  $(R')^2 \log \log d / \log d$ . When  $q > 0$ , we run a grid search over these high-variance directions; however, when  $q = 0$ , no grid search is needed.
  - (d) Set the list  $C_{B'} = \{c : \Pi_{\text{high}} c = \ell, \Pi_{\text{low}} c = \mu'_{\text{low}}, \ell \in L\}$ , where  $\mu'_{\text{low}}$  is the mean of  $Y'_{\text{low}}$  and  $L$  is the  $\tau$ -net of the ball  $B'_{\text{high}}$  for  $\tau = R_{\min} / \log d$ .
  - (e) Combine all candidate centers  $\mathfrak{C} = \cup_{B' \in \mathfrak{B}} C_{B'}$ .
  - (f) Set the list  $\mathfrak{B}_2 = \{B(c, \|y - c\|) \mid c \in \mathfrak{C}, y \in Y\}$ .
4. Return the ball  $\hat{B} \in \mathfrak{B}_2$  with the smallest volume that contains at least  $\delta(1 - \gamma)$  points in  $Y$ .

Figure 3: Algorithm DENSE\_BALL for finding a small volume ball that contains at least  $\delta' = \delta(1 - \gamma)$  fraction of points

Aggregating over the up to  $n^2$  balls in  $\mathfrak{B}$ , we collect a list  $\mathfrak{C} = \bigcup_{B \in \mathfrak{B}} C_B$  of most  $n^2 d^{\frac{7}{2}}$  potential centers.

Now, we do a more refined search for  $B^*$  using the candidate centers in  $\mathfrak{C}$ . That is, we generate the list

$$\mathfrak{B}_2 = \{B(c, \|y - c\|) \mid c \in \mathfrak{C}, y \in Y\}.$$

Note that  $|\mathfrak{B}_2| \leq |\mathfrak{C}| \cdot n \leq n^3 d^{\frac{7}{2}}$ . Then, we output the minimum volume ball  $B \in \mathfrak{B}_2$ , such that  $|B \cap Y| \geq \delta(1 - \gamma)n$ . Finding this ball can be done in time  $\tilde{O}(|\mathfrak{B}_2| \cdot n) = \tilde{O}(n^4 d^{\frac{7}{2}})$ . Thus the total runtime of our algorithm is bounded by  $\tilde{O}(n^4 d^{\frac{7}{2}})$ .

Now we argue that  $\mathfrak{B}_2$  contains a ball that is a good approximation to  $B^*$  in volume. Consider the ball  $B'$  from (10). We have that  $B'$  contains all points in  $Y^*$ , has radius  $R' \leq 2R^*$ , and is in



the list  $\mathfrak{B}$ . Let  $Y' = Y \cap B'$  be the subset of points of  $Y$  that are captured by  $B'$ . An averaging argument (Lemma 13) tells us that  $\lambda_{(q+1)}(\Sigma_{Y'}) \leq \frac{R'^2}{q}$ , where  $\Sigma_{Y'}$  is the covariance of the points  $Y'$ , and  $\lambda_{(q+1)}(\Sigma_{Y'})$  is the  $(q+1)$ -th largest eigenvector of  $\Sigma_{Y'}$ . Thus, when we apply the procedure of Lemma 12 to  $B'$  with  $q = \frac{\log d}{\log \log d}$  and  $\tau = \frac{R_{\min}}{\log d}$ , we also have that the eigenvalue  $\lambda_{(q+1)}(\Sigma_{Y'})$  is upper bounded by  $\sigma^2 \leq \frac{R'^2}{q} \leq \frac{(2R^*)^2 \log \log d}{\log d}$ . This guarantees that, for the list of potential centers  $C_{B'}$  that we get for this ball  $B'$ , there is a center  $\hat{c} \in C_{B'}$  and a subset  $\hat{Y} \subseteq Y^*$ ,  $|\hat{Y}| \geq (1 - \gamma)|Y^*|$ , such that for every  $y \in \hat{Y}$ ,

$$\|y - c^*\|_2^2 \leq \|\Pi_{\text{low}}(y - c^*)\|_2^2 + O\left(\frac{\sigma^2}{\gamma\delta}\right) + (\|\Pi_{\text{high}}(y - c^*)\|_2 + \tau)^2,$$

where  $\Pi_{\text{high}}$  is the projection onto the top- $q$  eigenspace of  $\Sigma_{Y'}$  and  $\Pi_{\text{low}}$  is the projection onto the space orthogonal to the top- $q$  eigenspace of  $\Sigma_{Y'}$ .

To bound the second term, we use the fact that

$$(a + b)^2 \leq (1 + \varepsilon)a^2 + \left(1 + \frac{1}{\varepsilon}\right)b^2, \quad \forall \varepsilon > 0.$$

Using  $\tau = \frac{R_{\min}}{\log d}$ , and  $\varepsilon = \frac{1}{\log d}$ , we get

$$\begin{aligned} \|y - c^*\|_2^2 &\leq \|\Pi_{\text{low}}(y - c^*)\|_2^2 + O\left(\frac{R'^2 \log \log d}{\gamma\delta \log d}\right) \\ &\quad + \left(1 + \frac{1}{\log d}\right) \|\Pi_{\text{high}}(y - c^*)\|_2^2 + (1 + \log d) \cdot \frac{R_{\min}^2}{\log^2 d} \\ &\leq \left(1 + \frac{1}{\log d}\right) \|y - c^*\|_2^2 + O\left(\frac{(R^*)^2 \log \log d}{\gamma\delta \log d}\right) \\ &\leq (R^*)^2 \left(1 + O\left(\frac{\log \log d}{\gamma\delta \log d}\right)\right). \end{aligned}$$

Thus, it is guaranteed that the list of potential centers  $C_{B'}$  that we get for this ball  $B'$  contains a center  $\hat{c} \in C_{B'}$  such that  $\hat{B} = B(\hat{c}, \hat{R})$  contains at least  $(1 - \gamma)|Y^*|$  points for

$$\hat{R} \leq R^* \left(1 + O\left(\frac{\log \log d}{\gamma\delta \log d}\right)\right)^{1/2},$$

which corresponds to a volume bound of

$$\text{vol}(\hat{B}) \leq \text{vol}(B^*) \exp\left(O\left(\frac{d \log \log d}{\gamma\delta \log d}\right)\right).$$

Since  $\hat{c} \in \mathfrak{C}$ , we are guaranteed that the smallest ball  $B_2 = B(\hat{c}, \|y - c\|)$  for  $y \in Y$  that contains at least  $(1 - \gamma)|Y^*|$  points can have radius at most that of  $\hat{B}$ . Since  $B_2 \in \mathfrak{B}_2$ , we know that the minimum volume ball in  $\mathfrak{B}_2$  that contains at least  $\delta(1 - \gamma)n$  points from  $Y$  can only have radius at most that of  $B_2$ , which is at most  $\hat{R}$ . ■

## A.2. Better volume approximation guarantee for almost isotropic distributions

Our strategy starts by finding a ball in  $B'$  that contains all of the points in  $B^*$  and has radius at most twice that of  $B^*$ . (More precisely, we find a list of candidate balls, which can be done simply.) Trivially, this means that in every direction the standard deviation of the points in  $B'$  is at most  $R^*$ . Lemma 12 tells us that choosing the mean of the points in  $B'$  as the center of  $\hat{B}$  essentially allows us to capture most of the points in  $B^*$  with radius that is  $R^*$  plus the maximum standard deviation in any direction. This gives a 2-approximation to the radius. We note that even a slightly stronger bound on the standard deviation can improve the approximation factor to  $1 + o(1)$ . In Theorem 8, we do this by grid-searching the  $\log d$  many highest variance directions of the points in  $B'$ , and taking the mean in the low-variance directions. This allows us to argue about the standard deviation of the  $(\log d + 1)$ th highest variance direction, which can be at most  $\frac{R^*}{\sqrt{\log d}}$  (Lemma 13).

This bound on the variance that we use in Theorem 8 is the best that our approach achieves, since it is possible that the variance of the points is concentrated in a few directions. However, if the points in  $B^*$  (the “inliers”) are not worst-case, and are instead approximately isotropic, we expect the standard deviation in any direction to be bounded by  $\frac{R^*}{\sqrt{d}}$ , which we can use to give a better bound on the radius of the ball that we output. In our earlier argument, we bounded the variance not only of the inliers— the points in  $B^*$ — but also all of  $Y'$ , which includes some of the outliers. To utilize this weaker assumption on the variance, we appeal to algorithms for list-decodable robust mean estimation, which work exactly in this setting where the inliers are well-behaved, but the outliers may be arbitrary.

This observation is also useful as a subroutine for our algorithm that finds a small volume ellipsoid in Appendix B. At a high level, that algorithm will first estimate a linear transformation that limits the variance of the points in  $B^*$ . Then, after applying this transformation, we essentially use Algorithm DENSE-BALL-ISOTROPIC to search for a ball in the transformed space with a better volume approximation. In that application, we will apply the transformation to and control the variance of all of our points, inliers and outliers, and thus avoid the need for list-decodable robust mean estimation.

**Theorem 4 (Bounded variance implies better bounds)** *Let  $Y \subseteq \mathbb{R}^d$  be a set of  $n$  points, and  $\delta, \gamma \in (0, 1)$ , such that there is an unknown ball  $B^* = B(c^*, R^*)$  with  $Y^* = B^* \cap Y$  satisfying  $|Y^*| \geq \delta|Y|$ , and  $Y^*$  is  $\beta(R^*)^2/d$ -isotropic for some  $\beta \geq 0$ . That is, if  $\mu^*$  is the mean of the points  $Y^*$ , then*

$$\Sigma_{Y^*} = \frac{1}{|Y^*|} \sum_{y \in Y^*} (y - \mu^*)(y - \mu^*)^\top \preceq \beta \frac{(R^*)^2}{d} I.$$

*Then we can find a ball  $\hat{B} = B(\hat{c}, \hat{R})$  such that  $|Y \cap \hat{B}| \geq (1 - \gamma)|Y^*|$ , and*

$$\text{vol}(\hat{B})^{1/d} \leq \text{vol}(B^*)^{1/d} \sqrt{1 + O\left(\frac{\beta}{\gamma \delta d}\right)},$$

*in polynomial time.*

**Proof** We begin by finding a coarse estimate of  $R^*$ . We do this as in Theorem 8: we can create a list of coarse guesses for  $B^*$ :

$$\mathfrak{B} = \{B(y_1, \|y_1 - y_2\|) \mid y_1, y_2 \in Y\}.$$

**Algorithm DENSE-BALL-ISOTROPIC**

**Input:** a set of point  $Y \in \mathbb{R}^d$ , target fraction  $\delta \in (0, 1)$ , slack  $\gamma \in (0, 1)$ , isotropic parameter  $\beta \in (0, 1)$

**Output:** an ball  $\hat{B} \subset \mathbb{R}^d$

1. Create a list of coarse balls  $\mathfrak{B} = \{B(y_1, \|y_1 - y_2\|) \mid y_1, y_2 \in Y\}$  and remove from  $\mathfrak{B}$  all balls that contain less than  $\delta n$  points in  $Y$ .
2. Set  $R_{\min}$  to be the minimum radius among all balls in  $\mathfrak{B}$ .
3. Create a list of candidate centers  $L$  by running the list-decodable mean estimation algorithm (Diakonikolas et al., 2021; Diakonikolas and Kane, 2023) on points  $Y$ , target fraction  $\delta$ , and variance  $\beta(R_{\min})^2$ .
4. Create a list of refined balls  $\mathfrak{B}_2 = \{B(c, \|y - c\|) \mid c \in L, y \in Y\}$ .
5. Return the ball  $\hat{B} \in \mathfrak{B}_2$  with the smallest volume that contains at least  $\delta(1 - \gamma)$  points in  $Y$ .

Figure 4: Algorithm DENSE-BALL-ISOTROPIC for finding a small volume ball that contains at least  $\delta' = \delta(1 - \gamma)$  fraction of points for isotropic distributions

By taking  $y_1, y_2$  to be maximally distant points in  $B^*$ , we have that there is a  $\hat{B} \in \mathfrak{B}$  such that the radius  $\hat{R}$  of  $\hat{B}$  satisfies

$$R^* \leq \hat{R} \leq 2R^*.$$

Taking the minimum radius  $R_{\min}$  over all balls in  $\mathfrak{B}$  that covers at least  $\delta n$  points gives us an estimate of  $R^*$  such that  $R^* \leq R_{\min} \leq \hat{R} \leq 2R^*$ .

Let  $\mu^*$  be the mean of the points in  $Y^*$ . We know that  $Y^*$  has covariance p.s.d. dominated by  $\frac{\beta(R^*)^2}{d}I$ , which is in turn p.s.d. dominated by  $\frac{\beta(R_{\min})^2}{d}I$ . The polynomial-time algorithm for list-decodable mean estimation (Diakonikolas et al., 2021; Diakonikolas and Kane, 2023) outputs a list  $L$  of length  $O(\log(1/p)/\delta)$  such that with probability  $1 - p$ , there exists a  $\hat{\mu} \in L$  such that

$$\|\hat{\mu} - \mu^*\| = O\left(\left(\frac{\beta}{\delta d}\right)^{1/2} R_{\min}\right) = O\left(\left(\frac{\beta}{\delta d}\right)^{1/2} R^*\right).$$

By Lemma 6 we have that there exists a subset  $\hat{Y} \subseteq Y \cap B^*$ , such that  $|\hat{Y}| \geq (1 - \gamma)|Y \cap B^*|$ , and for points  $y \in \hat{Y}$ ,

$$\begin{aligned} \|y - \hat{\mu}\|_2^2 &\leq \|y - c^*\|_2^2 + \gamma^{-1} \left( \frac{\beta}{d} (R_{\min})^2 + O(\delta^{-1}(\beta/d)(R^*)^2) \right) \\ &\leq (R^*)^2 + O(\gamma^{-1}\delta^{-1}(\beta/d)(R^*)^2) \\ &\leq (R^*)^2 \left( 1 + O\left(\frac{\beta}{\gamma\delta d}\right) \right) \end{aligned}$$

Now, we can do one more search step and create a refined list of candidate balls

$$\mathfrak{B}_2 = \{B(\hat{\mu}, \|\hat{\mu} - y\|) \mid \hat{\mu} \in L, y \in Y\}.$$

Since  $|\mathfrak{B}_2| \leq |L| \cdot |Y|$ , we have that this list is polynomially sized. We output the minimum volume ball  $\hat{B} \in \mathfrak{B}_2$  such that  $|\hat{B} \cap Y| \geq (1 - \gamma)\delta|Y|$ . This guarantees that  $\hat{B}$  has radius  $\hat{R}$  with

$$\hat{R} \leq R^* \sqrt{1 + O\left(\frac{\beta}{\gamma\delta d}\right)}.$$

■

## Appendix B. Finding Confidence Sets $(\Gamma = 1 + d^{-1/2+o(1)})$ -Competitive with Euclidean Balls

In this section, we provide an algorithm to find a confidence set whose volume is within a  $\exp(O(d^{1/2+o(1)}))$ -factor of the volume of the smallest ball that contains a  $\delta$  fraction of  $Y$ .

**Theorem 1 (Learning confidence sets competitive with Euclidean balls)** *There is a polynomial time algorithm that for any target coverage  $\delta \in (0, 1)$  and coverage slack  $\gamma \in (0, 1)$ , when given  $n = \Omega(d^2/\gamma^2)$  samples drawn i.i.d. from an arbitrary distribution  $\mathcal{D}$ , finds with high probability a set  $S \subset \mathbb{R}^d$  that is  $\Gamma = \exp(O_{\gamma,\delta}(d^{-1/2+o(1)}))$  competitive, i.e.,*

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

and

$$\text{vol}(S)^{1/d} \leq \text{vol}(B^*)^{1/d} \left(1 + O\left(d^{-1/2+o(1)}(\gamma\delta)^{-1}\right)\right).$$

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

The above is an immediate consequence of the following theorem which is for the empirical version of the problem, using standard concentration tools (Devroye and Lugosi, 2001) by incurring an additive  $O(\sqrt{d^2/n})$  term in the coverage probability (since the VC-dimension of  $d$ -dimensional ellipsoids is at most  $d^2 + d$ ).

**Theorem 7 (Finding small volume ellipsoid for  $n$  points)** *Suppose we are given a set of points  $Y \subseteq \mathbb{R}^d$ ,  $|Y| = n$ , and  $0 \leq \delta \leq 1, 0 \leq \gamma \leq 1$ , such that there exists a subset  $Y^* \subseteq Y$ ,  $|Y^*| \geq \delta|Y|$ , that is contained in an unknown ball  $B^* = B(c^*, R^*)$ . Then we can find an ellipsoid  $\hat{E}$  such that  $|\hat{E} \cap Y| \geq \delta(1 - \gamma)|Y|$ , and*

$$\text{vol}(\hat{E})^{1/d} \leq \text{vol}(B^*)^{1/d} \cdot \left(1 + O\left(d^{-1/2+o(1)}/\gamma\delta\right)\right).$$

**Proof** [Proof of Theorem 7] We begin by finding a list of coarse guesses for  $B^*$ , as in Theorem 8. We create a list of balls

$$\mathfrak{B} = \{B(y_1, \|y_1 - y_2\|) \mid y_1, y_2 \in Y\},$$

**Algorithm DENSE\_ELLIPSOID**

**Input:** a set of point  $Y \in \mathbb{R}^d$ , target fraction  $\delta \in (0, 1)$ , and slack  $\gamma \in (0, 1)$

**Output:** an ellipsoid  $\hat{E} \subset \mathbb{R}^d$

1. Create a list of coarse balls  $\mathfrak{B} = \{B(y_1, \|y_1 - y_2\|) \mid y_1, y_2 \in Y\}$  and remove from  $\mathfrak{B}$  all balls that contains less than  $\delta n$  points in  $Y$ .
2. Create an ellipsoid  $\hat{E}_i$  for each coarse ball  $B'_i$  in the list  $\mathfrak{B}$  as follows:
  - (a) Set  $\hat{\tau} = d^{1/4}$  and  $R'_i$  be the radius of  $B'_i$ .
  - (b) Compute the eigenvalues  $\lambda_1, \dots, \lambda_d$  and eigenvectors  $v_1, \dots, v_d$  of the covariance  $\Sigma_{Y'_i}$  of points  $Y'_i = Y \cap B'_i$ .
  - (c) Create an ellipsoidal shape  $\hat{M}_i = \sum_{j=1}^d a_j^2 v_j v_j^\top$ , where  $a_j^2 = d$  if  $\lambda_j \geq \hat{\tau}^2 (R'_i)^2 / d$  and  $a_j^2 = 1$  if  $\lambda_j < \hat{\tau}^2 (R'_i)^2 / d$ .
  - (d) Compute the transformed points  $Z_i = \hat{M}_i^{-1/2} Y$ . Find a ball  $\hat{B}_i$  using subroutine from Lemma 12 for  $q = 0$ ,  $\lambda_{(q+1)}(\Sigma_{Y'}) \leq \sigma^2$ ,  $\sigma^2 = \hat{\tau}^2 (R')^2 / d$ , slack factor  $\gamma$ .
  - (e) Set  $\hat{E}_i = \hat{M}_i^{1/2} \hat{B}_i$ .
3. Return the ellipsoid  $\hat{E}$  with the smallest volume in the list  $\{\hat{E}_i\}$ .

Figure 5: Algorithm DENSE\_ELLIPSOID for finding a small volume ellipsoid that contains at least  $\delta' = \delta(1 - \gamma)$  fraction of points

(step 1 of Algorithm DENSE\_ELLIPSOID). Note that  $|\mathfrak{B}| \leq n^2$  since  $|Y| = n$ . Consider the subset  $Y^* \subseteq Y$  contained in ball  $B^* = B(c^*, R^*)$ . Choosing

$$y_1^*, y_2^* = \operatorname{argmax}_{y_1, y_2 \in Y^*} \|y_1 - y_2\|,$$

gives us that the ball

$$B' = B(y_1^*, \|y_1^* - y_2^*\|)$$

contains all points in  $Y^*$ . It has radius  $R' \leq 2R^*$  and  $B' \in \mathfrak{B}$ . Let  $Y'$  be all points contained in this ball  $B'$ .

For a parameter  $\hat{\tau}$  that we will set later, Algorithm DENSE\_ELLIPSOID chooses an ellipsoidal shape  $\hat{M}$  to have the same eigenvectors as  $\Sigma_{Y'}$ , with eigenvalues  $a_i^2$  of  $\hat{M}$  satisfying  $a_i^2 = d$  if  $\lambda_i \geq \hat{\tau}^2 (R')^2 / d$  and  $a_i^2 = 1$  if  $\lambda_i < \hat{\tau}^2 (R')^2 / d$ .  $\hat{M}$  has the following nice properties.

- (i) The eigenvalues of  $\hat{M}$  are in  $\{1, d\}$ . Thus  $\hat{M} \succeq I$ , and in particular, for any set of points  $P$ ,

$$\operatorname{vol}(\operatorname{encball}(\hat{M}^{-1/2} P)) \leq \operatorname{vol}(\operatorname{encball}(P)), \quad (11)$$

where  $\operatorname{encball}(S)$  is the minimum volume ball enclosing  $S$ .

- (ii) Let  $\Sigma_{Y'}$  be the covariance matrix of  $Y'$ , and  $\lambda_1, \dots, \lambda_d$  be the eigenvalues of  $\Sigma_{Y'}$ . We show that

$$a_i^2 \geq \frac{d}{\hat{\tau}^2(R')^2} \lambda_i \quad \text{for all } i. \quad (12)$$

When  $\lambda_i < \hat{\tau}^2(R')^2/d$ , we set  $a_i^2 = 1$ , which satisfies (12). When  $\lambda_i \geq \hat{\tau}^2(R')^2/d$ , we set  $a_i^2 = d$ . We know that since  $Y' \subseteq B'$ , every eigenvalue of  $\Sigma_{Y'}$  is upper bounded as  $\lambda_i \leq (R')^2$ . Thus  $(d/\hat{\tau}^2(R')^2)\lambda_i \leq d/\hat{\tau}^2 \leq d$ , satisfying (12). By definition, (12) implies that

$$\widehat{M} \succeq \frac{d}{\hat{\tau}^2(R')^2} \Sigma_{Y'}.$$

In particular, this implies that the points in  $Y'$  have bounded variance in the transformed space  $\widehat{M}^{-1/2}(\mathbb{R}^d) = \{\widehat{M}^{-1/2}y : y \in \mathbb{R}^d\}$ ,

$$\widehat{M}^{-1/2} \Sigma_{Y'} \widehat{M}^{-1/2} \preceq O\left(\frac{\hat{\tau}^2(R')^2}{d}\right) I_d. \quad (13)$$

- (iii) Since  $Y' \subseteq B'$ , the total variance of the points  $Y'$  is bounded as

$$\sum_i \lambda_i \leq (R')^2.$$

Since the  $\lambda_i$  are non-negative, this tell us that there can be at most  $d/\hat{\tau}^2$  values of  $i$  for which  $\lambda_i \geq \hat{\tau}^2(R')^2/d$ . Thus at most  $d/\hat{\tau}^2$  eigenvalues of  $\widehat{M}$  are set to  $d$ .

In particular, this implies that for any shape  $S$ ,

$$\frac{\text{vol}(S)}{\text{vol}(\widehat{M}^{-1/2}S)} \leq d^{d/\hat{\tau}^2} \leq \exp\left(\frac{d \ln d}{\hat{\tau}^2}\right). \quad (14)$$

The algorithm transforms the points  $Y'$  by  $\widehat{M}^{-1/2}$ . We denote the points in the transformed space as  $T'$ ,

$$T' = \{\widehat{M}^{-1/2}y \mid y \in Y'\}.$$

Recall that by the choice of  $Y'$ ,  $Y^\star \subseteq Y'$ , and we denote

$$T^\star = \{\widehat{M}^{-1/2}y \mid y \in Y^\star\} \subseteq T'.$$

We apply the procedure from Lemma 12 with  $q = 0$  to the minimum volume ball  $B_T' = B(c_T', R_T')$  containing  $T'$ . We have that  $\lambda_{\max}(\Sigma_{Y'}) \leq \sigma^2$ , where  $\sigma^2 = \hat{\tau}^2(R')^2/d$  (due to (13)), and slack factor  $\gamma$ . Let  $B_T^\star = B(c_T^\star, R_T^\star)$  be the minimum volume ball containing  $T^\star$ . Lemma 12 allows us to find a ball  $\widehat{B}_T = B(\widehat{c}_T, \widehat{R}_T)$  that contains a subset  $\widehat{T} \subseteq T'$ , where  $|\widehat{T}| \geq (1 - \gamma)|T^\star|$ , and

$$\widehat{R}_T \leq \sqrt{(R_T^\star)^2 + \frac{\sigma^2}{\gamma} \left(1 + \frac{2(1 - \delta)}{\delta}\right)}. \quad (15)$$

(Lemma 12 has an additional parameter also named  $\tau$ , that is different from  $\hat{\tau}$ , but this parameter does not appear here since we set  $q = 0$  and avoid the grid search element of the procedure entirely.)



Finally, we transform  $\widehat{B}_T$  back to the transformed space to get the candidate ellipsoid  $\widehat{E}$ ,

$$\widehat{E} = \widehat{M}^{1/2} \widehat{B}_T.$$

Note that  $|Y' \cap \widehat{E}| \geq (1 - \gamma)|Y^*|$  (since  $|\widehat{T}| \geq (1 - \gamma)|T^*|$ ). We can bound the volume of  $\widehat{E}$  as

$$\begin{aligned}
 \text{vol}(\widehat{E}) &\leq \exp\left(\frac{d \ln d}{\widehat{\tau}^2}\right) \cdot \text{vol}(\widehat{B}_T) && \text{by (14)} \\
 &\leq \exp\left(\frac{d \ln d}{\widehat{\tau}^2}\right) \cdot c_d \widehat{R}_T^d && c_d : \text{vol. of } d \text{ dim. unit ball} \\
 &\leq \exp\left(\frac{d \ln d}{\widehat{\tau}^2}\right) \cdot c_d \left( (R_T^*)^2 + \sigma^2 \gamma \left(1 + \frac{2(1 - \delta)}{\delta}\right) \right)^{d/2} && \text{by (15)} \\
 &\leq \exp\left(\frac{d \ln d}{\widehat{\tau}^2}\right) \cdot c_d \left( (R^*)^2 + \frac{\sigma^2}{\gamma} \left(1 + \frac{2(1 - \delta)}{\delta}\right) \right)^{d/2} && R^* \geq R_T^* \text{ by (11)} \\
 &\leq \exp\left(\frac{d \ln d}{\widehat{\tau}^2}\right) \cdot c_d (R^*)^d \left(1 + \frac{\sigma^2}{\gamma (R^*)^2} \left(1 + \frac{2(1 - \delta)}{\delta}\right)\right)^{d/2} \\
 &\leq \exp\left(\frac{d \ln d}{\widehat{\tau}^2}\right) \cdot \exp\left(\frac{d}{2} \cdot \frac{\sigma^2}{\gamma (R^*)^2} \left(1 + \frac{2(1 - \delta)}{\delta}\right)\right) \cdot \text{vol}(B^*) \\
 &\leq \exp\left(\frac{d \ln d}{\widehat{\tau}^2}\right) \cdot \exp\left(\frac{\widehat{\tau}^2 (R')^2}{2\gamma (R^*)^2} \left(1 + \frac{2(1 - \delta)}{\delta}\right)\right) \cdot \text{vol}(B^*) && \text{since } \sigma^2 = \frac{\widehat{\tau}^2 (R')^2}{d} \\
 &\leq \exp\left(\frac{d \ln d}{\widehat{\tau}^2} + \widehat{\tau}^2 \cdot \frac{2}{\gamma} \left(1 + \frac{2(1 - \delta)}{\delta}\right)\right) \cdot \text{vol}(B^*) && \text{since } R' \leq 2R^* \\
 &\leq \exp\left(d^{1/2} \left(\ln d + \frac{2}{\gamma} \left(1 + \frac{2(1 - \delta)}{\delta}\right)\right)\right) \cdot \text{vol}(B^*) && \text{setting } \widehat{\tau}^2 = d^{1/2} \\
 &\leq \exp\left(d^{1/2+o(1)} (\gamma\delta)^{-1}\right) \cdot \text{vol}(B^*).
 \end{aligned}$$

Rephrasing the above bound in terms of  $\text{vol}(\widehat{E})^{1/d}$  gives the stated bound. ■

## Appendix C. Greedy Algorithm and Unions of Sets

In this section, we provide a greedy algorithm for constructing a union of sets from a given set system that achieves the desired coverage with a small volume compared to the minimum volume union of  $k$  sets.

**Theorem 14 (Greedy algorithm)** *Let  $Y \subseteq \mathbb{R}^d$  be a set of  $n$  points, and  $\mathcal{C}$  be a set system over  $\mathbb{R}^d$  with bounded VC-dimension  $D$ ,  $k \in \mathbb{N}$ ,  $\delta \in (0, 1)$  be a coverage level, and  $\gamma \in (0, 1 - \delta)$  be a slack parameter in coverage. Let  $\mathcal{A}_\alpha(Y', \delta', \gamma')$  for approximation factor  $\alpha \geq 1$  be an algorithm, that given a set of points  $Y'$ , and a coverage level  $\delta'$  and a slack  $\gamma' \in (0, 1)$ , outputs a  $C \in \mathcal{C}$  such that  $|C \cap Y'| \geq (1 - \gamma')\delta|Y'|$ , and  $\text{vol}(C) \leq \alpha \cdot \text{vol}(C')$ , where  $C'$  is the minimum volume set in  $\mathcal{C}$  such that  $|C' \cap Y'| \geq \delta|Y'|$ .*

Then, in  $\text{poly}(n, d, k)$  time and calls to  $\mathcal{A}_\alpha$ , Algorithm GREEDY\_DENSITY outputs a set  $\widehat{C} \subseteq \mathbb{R}^d$  such that  $\widehat{C}$  is a union of  $O(\delta k / \gamma)$  sets from  $\mathcal{C}$ , and

$$\text{vol}(\widehat{C}) \leq O\left(\frac{\alpha \log(k/\gamma)}{\gamma}\right) \cdot \text{vol}(C^*),$$

where  $C^*$  is the minimum volume union of  $k$  sets from  $\mathcal{C}$  such that  $|C^* \cap Y| \geq (\delta + \gamma)n$ .

**Algorithm GREEDY\_DENSITY**

**Input:** points  $Y \subseteq \mathbb{R}^d$  with  $|Y| = n$ , coverage level  $\delta \in [0, 1]$ , slack factor  $\gamma$ , number of sets  $k \in \mathbb{N}$ , black-box algorithm  $\mathcal{A}_\alpha$  (see description in Theorem 14)

**Output:** confidence set  $\widehat{C} \subseteq \mathbb{R}^d$

1. Set  $t = 0$  and  $\widehat{C} = \emptyset$ ,  $Y_t = Y$ .
2. While  $|\widehat{C} \cap Y| \leq \delta n$ :
  - (a) For  $i \in \{0, \dots, \lceil \log_2 n \rceil\}$ , let  $S_i^{(t)} = \mathcal{A}_\alpha(Y_t, \frac{2^i}{n}, \gamma')$
  - (b) Let
 
$$S^{(t)} = \max_{S_i^{(t)}} \frac{|S_i^{(t)} \cap Y_t|}{\text{vol}(S_i^{(t)})}$$

such that  $|S_i^{(t)} \cap Y_t| \geq (1 - \gamma') \frac{\gamma}{4k} \cdot n$
  - (c) Set  $\widehat{C} = \widehat{C} \cup S^{(t)}$ ,  $Y_{t+1} = Y_t \setminus S^{(t)}$ ,  $t = t + 1$
3. Output  $\widehat{C}$

Figure 6: Algorithm for finding the small confidence set

**Proof** Let  $C_1^*, \dots, C_k^* \in \mathcal{F}$  be the optimal  $k$  sets in  $\mathcal{F}$  that covers at least a  $\delta + \gamma$  fraction of points in  $Y$ . Let  $C^* = C_1^* \cup \dots \cup C_k^*$ .

First we bound the number of iterations of the algorithm. For each iteration of the algorithm, we have that  $|\widehat{C} \cap Y| \leq \delta n$ . This means that  $|C^* \cap Y_t| \geq \gamma n$ , so there exists a  $C_\ell^* \in C^*$  such that  $|C_\ell^* \cap Y_t| \geq \frac{\gamma n}{k}$ . Thus, step (b) of our algorithm will successfully find a set in  $S^{(t)} \in \mathcal{C}$  such that  $|S^{(t)} \cap Y_t| \geq (1 - \gamma') \frac{\gamma}{4k} \cdot n$ . Setting  $\gamma' = \frac{1}{2}$  we have that the algorithm must terminate in

$$O\left(\frac{\delta}{\gamma} \cdot k\right)$$

iterations, and output a union of at most that many balls.

Now we bound the volume of the output set. We group the iterations into phases. In particular, phase  $j$  consists of iterations in which  $|\widehat{C} \cap Y| \in [(1 - \frac{1}{2^j})\delta n, (1 - \frac{1}{2^{j+1}})\delta n]$ . Appendix C tells us that the volume of sets added in each phase is

$$O\left(\frac{\alpha}{\gamma}\right) \cdot \text{vol}(C^*).$$

Since each iteration, the new picked set  $S^{(t)}$  covers at least  $(1 - \gamma') \frac{\gamma}{4k} \cdot n$  points, there can be at most  $\log_2(\delta k / \gamma) \leq \log(k / \gamma)$  many phases. This means that the total volume of the sets output over all phases is bounded by

$$O\left(\frac{\alpha}{\gamma} \cdot \log(k / \gamma)\right) \cdot \text{vol}(C^*).$$

■

[Marginal volume added in each phase is bounded] In every phase  $j$ , the total marginal volume added is

$$\leq O(\alpha) \cdot (\text{vol}(C_1^*) + \dots + \text{vol}(C_k^*)).$$

**Proof** Let  $t_j$  be the first iteration in phase  $j$ , and  $t_{j+1} - 1$  be defined as the last iteration in phase  $j$ .

We first show that in every iteration  $t$ , one of the sets  $C_1^*, \dots, C_k^*$  is a feasible solution with good density. Note that at the beginning of each iteration, the confidence set covers at most  $\delta$  fraction of points in  $Y$ . Let  $C^* = C_1^* \cup \dots \cup C_k^*$ . Since  $|C^* \cap Y| \geq (\delta + \gamma)|Y| = (\delta + \gamma)n$ , we have that  $|C^* \cap Y_t| \geq \gamma n$  for any iteration  $t$ .

Appendix C tells us that there is a set  $C_\ell^*$  that achieves individually high coverage and high density compared to  $C^*$  over  $Y_t$ . That is, there is an  $\ell \in [k]$  such that

$$|C_\ell^* \cap Y_t| \geq \frac{\gamma n}{2k} \quad \text{and} \quad \frac{|C_\ell^* \cap Y_t|}{\text{vol}(C_\ell^*)} \geq \frac{|C^* \cap Y_t|}{2 \cdot \text{vol}(C^*)}.$$

Since we are in phase  $j$ , we have that the number of points that we have covered thus far is  $|Y \setminus Y_t| \leq (1 - \frac{1}{2^{j+1}})\delta n$ , so  $|C^* \cap Y_t| \geq \frac{\delta n}{2^{j+1}}$ , so

$$\frac{|C_\ell^* \cap Y_t|}{\text{vol}(C_\ell^*)} \geq \frac{|C^* \cap Y_t|}{2 \cdot \text{vol}(C^*)} \geq \frac{\delta n}{2^{j+2} \cdot \text{vol}(C^*)}.$$

This means that, by Appendix C, will choose a set  $C^{(t)} \in \mathcal{C}$  such that

$$|C^{(t)} \cap Y_t| \geq (1 - \gamma') \frac{\gamma n}{4k} \quad \text{and} \quad \frac{|C^{(t)} \cap Y_t|}{\text{vol}(C^{(t)})} \geq \frac{|C_\ell^* \cap Y_t|}{\alpha \cdot \text{vol}(C_\ell^*)} \geq \frac{\delta n}{\alpha \cdot 2^{j+2} \cdot \text{vol}(C^*)}.$$

This implies that

$$\begin{aligned} \frac{|C^{(t)} \cap Y_t|}{\text{vol}(C^{(t)})} &\geq \frac{\delta n}{\alpha \cdot 2^{j+2} \cdot \text{vol}(C^*)} \\ \frac{|C^{(t)} \cap Y_t| \cdot \alpha \cdot 2^{j+2} \cdot \text{vol}(C^*)}{\delta n} &\geq \text{vol}(C^{(t)}). \end{aligned}$$

Now we can aggregate over the iterations in phase  $j$ . We first consider all but the last iteration, that is  $t$  such that  $t_j \leq t \leq t_{j+1} - 2$ . (It can be the case that  $t_{j+1} - 2 < t_j$ , in which case there are no  $t$  in this category and the bound holds trivially.)

$$\left( \sum_{t=t_j}^{t_{j+1}-2} |C^{(t)} \cap Y_t| \right) \cdot \frac{\alpha \cdot 2^{j+2} \cdot \text{vol}(C^*)}{\delta n} \geq \sum_{t=t_j}^{t_{j+1}-2} \text{vol}(C^{(t)})$$

Now, since for all iterations  $t$  such that  $t_j \leq t \leq t_{j+1} - 2$ , we have that  $t + 1$  is still in phase  $j$ , we know that the chosen sets could not have covered more than  $\frac{1}{2^j}n$  points.

$$\begin{aligned} \frac{\delta n}{2^j} \cdot \frac{\alpha \cdot 2^{j+2} \cdot \text{vol}(C^*)}{\delta n} &\geq \sum_{t=t_j}^{t_{j+1}-2} \text{vol}(C^{(t)}) \\ O(\alpha) \cdot \text{vol}(C^*) &\geq \sum_{t=t_j}^{t_{j+1}-2} \text{vol}(C^{(t)}). \end{aligned}$$

Finally, for the iteration  $t = t_{j+1} - 1$ , we have that  $|C^{(t)} \cap Y_t| \leq n$ . So we get that

$$\begin{aligned} \frac{|C^{(t)} \cap Y_t|}{\text{vol}(C^{(t)})} &\geq \frac{1}{2\alpha} \cdot \frac{|C_\ell^* \cap Y_t|}{\text{vol}(C_\ell^*)} && \text{by Appendix C} \\ \frac{|C^{(t)} \cap Y_t|}{\text{vol}(C^{(t)})} &\geq \frac{1}{4\alpha} \cdot \frac{|C^* \cap Y_t|}{\text{vol}(C^*)} && \text{by Appendix C} \\ \frac{n}{\text{vol}(C^{(t)})} &\geq \frac{1}{4\alpha} \cdot \frac{\gamma n}{\text{vol}(C^*)} \\ \frac{4\alpha}{\gamma} \text{vol}(C^*) &\geq \text{vol}(C^{(t)}). \end{aligned}$$

Thus in total, over all days in phase  $j$ , we have that the total volume of sets chosen by our algorithm is at most

$$O\left(\frac{\alpha}{\gamma}\right) \cdot \text{vol}(C^*).$$

■

Given  $a_1, a_2, \dots, a_k, b_1, b_2, \dots, b_k \geq 0$  such that  $\sum_{i=1}^k a_i \geq \beta$ , and  $\sum_{i=1}^k a_i / \sum_{i=1}^k b_i \geq \gamma$ . Then there exists  $i \in [k]$ , satisfying  $a_i \geq \beta/(2k)$  and  $a_i/b_i \geq \gamma/2$ .

**Proof** Let  $\beta' = \sum_{i=1}^k a_i$ . Suppose there is no  $a_i, b_i$  such that  $a_i \geq \beta'/(2k)$  and  $a_i/b_i \geq \gamma/2$ . Let  $B = \{i \in [k] : a_i \geq \beta'/(2k)\}$ . Then, we have

$$\sum_{i \in B} a_i = \sum_{i \in [k]} a_i - \sum_{i \in [k] \setminus B} a_i > \sum_{i \in [k]} a_i - k \cdot \frac{\beta'}{2k} = \frac{\beta'}{2}.$$

Moreover for all  $i \in B$  we have  $a_i/b_i < \gamma/2$  i.e.

$$\sum_{i \in B} b_i > \frac{2}{\gamma} \sum_{i \in B} a_i > \frac{2}{\gamma} \cdot \frac{\beta'}{2} = \frac{\beta'}{\gamma}.$$

But this contradicts  $\sum_{i \in [k]} b_i \leq \sum_{i=1}^k a_i / \gamma = \beta' / \gamma$ , which completes the proof. ■

[Each iteration finds approximate density maximizer] In each iteration, our algorithm chooses a set  $S^{(t)}$  such that

$$\frac{|S^{(t)} \cap Y_t|}{\text{vol}(S^{(t)})} \geq \frac{1}{2\alpha} \cdot \frac{|C^{(t)} \cap Y_t|}{\text{vol}(C^{(t)})},$$

and

$$(1 - \gamma') \frac{\gamma}{4k} \cdot n \leq |S^{(t)} \cap Y_t|,$$

where  $C^{(t)}$  is the maximum density set such that  $\frac{\gamma}{2k}n \leq |C^{(t)} \cap Y_t|$ .

**Proof** Fix an iteration  $t$ , and let  $C^{(t)}$  be the set in  $\mathcal{C}$  such that  $\frac{\gamma}{2k}n \leq |C^{(t)} \cap Y_t|$ , that maximizes

$$\frac{|C^{(t)} \cap Y_t|}{\text{vol}(C^{(t)})}.$$

Let  $i^* \in \{0, \dots, \lceil \log_2 n \rceil\}$  be the value that satisfies

$$2^{i^*} \leq |C^{(t)} \cap Y_t| \leq 2 \cdot 2^{i^*}.$$

Since  $C^{(t)}$  achieves coverage at least  $\frac{2^{i^*}}{n}$ , we have that the set  $S_{i^*}^{(t)}$  chosen in step (a) of the algorithm must have

$$\text{vol}(S_{i^*}^{(t)}) \leq \alpha \cdot \text{vol}(C^{(t)}) \quad \text{and} \quad |S_{i^*}^{(t)} \cap Y_t| \geq (1 - \gamma') \frac{2^{i^*}}{n}.$$

Thus, we have that

$$\frac{|S_{i^*}^{(t)} \cap Y_t|}{\text{vol}(S_{i^*}^{(t)})} \geq \frac{1}{2\alpha} \cdot \frac{|C^{(t)} \cap Y_t|}{\text{vol}(C^{(t)})},$$

and that

$$(1 - \gamma') \frac{\gamma}{4k} \cdot n \leq |S_{i^*}^{(t)} \cap Y_t|.$$

■

**Corollary 15 (Union of Balls)** *We give an algorithm that, given a set of points  $Y \subseteq \mathbb{R}^d$ ,  $|Y| = n$ , a coverage fraction  $\delta \in (0, 1)$ , a slack parameter in coverage  $\gamma \in (0, 1 - \delta)$  and  $k \in \mathbb{N}$ , can find a set  $\hat{C}$  such that*

$$\text{vol}(\hat{C}) \leq \exp\left(O_\delta\left(d^{1/2+o(1)}\right)\right) \cdot O\left(\frac{1}{\gamma} \cdot \log(k/\gamma)\right) \cdot \text{vol}(C^*),$$

and  $\hat{C}$  is a union of  $O(\frac{\delta k}{\gamma})$  ellipsoids, where  $C^*$  is the minimum volume union of  $k$  balls that covers at least  $(\delta + \gamma)$  fraction of the points in  $Y$ .

**Corollary 16 (Proper Learning Union of balls)** *Let  $\delta \in (0, 1)$ ,  $\gamma \in (0, 1)$ ,  $k \in \mathbb{N}$  be any constants. There is a polynomial time algorithm that for target coverage  $\delta \in (0, 1)$  and coverage slack  $\gamma \in (0, 1)$  when given  $n = \Omega(kd^2/\gamma^2)$  samples drawn i.i.d. from an arbitrary distribution  $\mathcal{D}$ , finds with high probability a set  $S \subset \mathbb{R}^d$  that is a union of balls, and is  $\Gamma = (1 + O_{\gamma,\delta}(\log \log d / \log d))$  competitive; more precisely, it satisfies  $\mathbf{P}_{y \sim \mathcal{D}}[y \in S] \geq \delta$ , and*

$$\text{vol}(S)^{1/d} \leq \text{vol}(C_k^*)^{1/d} \left(1 + O_{k,\delta}(\log \log d / \log d)\right) \cdot \left(\frac{O(\log(k/\gamma))}{\gamma}\right)^{1/d}$$

where  $C_k^*$  is the minimum volume union of  $k$ -balls that achieves at least  $\delta + \gamma + O(\sqrt{kd^2/n})$  coverage over  $\mathcal{D}$ .

## Appendix D. Hardness of Proper Learning

### D.1. NP-hardness of Proper Learning

In this section, we show the approximation hardness of finding the smallest volume ball that achieves the required coverage.

**Theorem 17 (NP-hardness of Proper Learning)** *For any small constant  $\varepsilon > 0$  and there is some constant  $\delta \geq 1/4$ , unless  $P = NP$ , there is no algorithm that given a set of points  $Y \subseteq \mathbb{R}^d$  runs in polynomial time in  $|Y|, d$  that finds a ball  $\hat{B} = B(\hat{c}, \hat{R})$  that contains at least a  $\delta$  fraction of points in  $Y$  with radius  $\hat{R} \leq (1 + d^{-\varepsilon})R^*$ , where  $R^*$  is the radius of the smallest ball  $B^*$  that contains at least a  $\delta$  fraction points in  $Y$ .*

We remark that the above hardness result rules out polynomial time proper learning algorithms that work for arbitrary distributions—this is called “the distribution-free setting” in PAC learning (as opposed to the distribution-specific setting). More formally, one can consider the following reduction from the worst-case problem of finding the minimum volume ball containing a  $\delta$  fraction of an adversarially chosen point set  $Y \subset \mathbb{R}^d$ , to our problem. We construct a distribution  $D_Y$  that is the uniform distribution over  $Y$ . By the construction of  $Y$ , finding a set that contains probability mass  $\delta$  over  $D_Y$  corresponds exactly to finding a set that contains at least a  $\delta$  fraction of  $Y$ . We can simulate running our algorithm on  $D_Y$  by sampling from  $D_Y$  and providing these samples to the algorithm. Thus our problem is only harder than the worst-case problem, because it only gets sample access to  $D_Y$ , rather than access to  $Y$  itself.

The above theorem involves a reduction from the maximum clique problem to the smallest  $k$ -enclosing ball due to [Shenmaier \(2015\)](#). We now proceed to the proof of Theorem 17.

**Proof** [Shenmaier \(2015\)](#) proves the strong NP-hardness of the smallest  $k$ -enclosing ball in Euclidean space through a reduction from the  $k$  clique problem on regular graphs. We utilize this construction in our reduction as follows. Let  $G$  be a  $\Delta$ -regular graph with  $n$  vertices and  $m$  edges. We construct a set of  $n$  points  $Y \subset \mathbb{R}^d$  with dimension  $d = m^{3/\varepsilon}$  for a small constant  $\varepsilon \in (0, 1)$ . Each point  $y \in Y$  corresponds to a vertex  $v$  in graph  $G$ . The first  $m$  coordinates of  $y$  are the  $m$ -dimensional row for vertex  $v$  in the incident matrix of  $G$ . Then, for  $i = 1, 2, \dots, m$ , we have  $y_i = 1$  if the edge  $i$  connects vertex  $v$ ; otherwise  $y_i = 0$ . The rest  $d - m$  coordinates of  $y$  have value 0. Note that this reduction takes time  $O(nd)$ , which is polynomial in  $m$  when  $\varepsilon$  is a small constant in  $(0, 1)$ .

We now show that this construction of  $Y$  provides the hardness of proper learning. Since the graph  $G$  is  $\Delta$  regular, we have the distance between two points in  $Y$  is  $\sqrt{2\Delta - 2}$  if the corresponding vertices are connected in  $G$ ; otherwise, the distance is  $\sqrt{2\Delta}$ . By Lemma 3 in [Shenmaier \(2015\)](#), we have if  $G$  has a  $k$  clique, then the smallest ball that contains at least  $k$  points in  $Y$  has a radius  $R^* \leq \sqrt{A_k}$  where  $A_k = (\Delta - 1)(1 - 1/k)$ ; otherwise, the smallest ball that contains  $k$  points in  $Y$  has a radius at least  $\sqrt{A_k + 2/k^2}$ . Note that  $\sqrt{A_k + 2/k^2} \geq \sqrt{A_k}(1 + 1/n^3) \geq \sqrt{A_k}(1 + 1/m^3)$ . The  $k$ -clique problem is NP-hard even for regular graphs [Fleischner et al. \(2010\)](#); [Feige \(2003\)](#). [Fleischner et al. \(2010\)](#) show that it is NP-hard to find the maximum independent set for planar three regular graphs. Note that for three regular graphs, the maximum independent set has a size of at least  $n/4$ . The complement of a regular graph is still a regular graph and the maximum independent set corresponds to the maximum clique in the complement graph. Thus, for some constant  $\delta \geq 1/4$ , given a regular graph  $G$  on  $n$  vertices that contains  $k = \delta n$  clique, it is NP-hard to find a  $k$  clique in  $G$ .

Since we pick  $d = m^{3/\varepsilon}$ , we have it is NP-hard to approximate the radius of the smallest ball that contains a  $\delta$  fraction of points in  $Y$  within a factor of

$$1 + \frac{1}{m^3} = 1 + \frac{1}{d^\varepsilon}.$$

Thus, it is NP-hard to find a ball that contains a  $\delta$  fraction of points in  $Y$  with a radius at most  $1 + \frac{1}{d^\varepsilon}$  times the radius  $R^*$  of the minimum ball that contains a  $\delta$  fraction of points in  $Y$ .  $\blacksquare$

## D.2. Computational Intractability even with Slack in Coverage

We give different reduction to provide evidence of strong computational intractability even when we are allowed to violate the coverage by a constant factor. Our hardness result is assuming the Small Set Expansion (SSE) hypothesis of [Raghavendra and Steurer \(2010\)](#), which is closely related to the Unique Games Conjecture [Khot \(2002\)](#).

**Conjecture 18 (SSE hypothesis of [Raghavendra and Steurer \(2010\)](#))** *(See e.g., Theorem IV.5 of [Raghavendra et al. \(2012\)](#).) For any constant  $\eta \in (0, 1/2)$ , there is a constant  $\tau \in (0, 1)$  such that there is no polynomial time algorithm to distinguish between the following two cases given a graph  $G = (V, E)$  on  $n$  vertices with degree  $D$ :*

- **YES:** Some subset  $S \subseteq V$  with  $|S| = \tau n$  satisfies that the induced subgraph on  $S$  is dense i.e., the number of edges going out of  $S$  is  $|E(S, V \setminus S)| \leq \eta D|S|$  edges.
- **NO:** Any set  $S \subseteq V$  with  $\eta\tau n \leq |S| \leq 2\tau n$  has most of the edges incident on it going outside i.e.,  $|E(S, V \setminus S)| \geq (1 - \eta)|S|D$ .

We prove the following theorem.

**Theorem 19 (Computational Intractability with Slack in Coverage)** *For any constant  $\gamma > 0$ , there exists a constant  $\delta \in (0, 1)$ , such that assuming the SSE hypothesis for any constant  $\varepsilon > 0$  there is no algorithm that given a set of points  $Y \subseteq \mathbb{R}^d$  runs in polynomial time and finds a ball  $\hat{B} = B(\hat{c}, \hat{R})$  that contains at least a  $\gamma\delta$  fraction of points in  $Y$  with radius  $\hat{R} \leq (1 + d^{-\varepsilon})R^*$ , where  $R^*$  is the radius of the smallest ball  $B^*$  that contains at least a  $\delta$  fraction points in  $Y$ .*

**Proof** Set  $\eta := \min\{\frac{\gamma^2}{9}, \frac{1}{16}\}$ . Given a  $D$ -regular graph  $G = (V, E)$  of the SSE problem in Conjecture 18 on  $n$  vertices with parameter  $\eta$ , we first construct an instance for ball coverage in  $\mathbb{R}^m$  where  $m = |E|$  as follows. Every edge corresponds to a coordinate, and every vertex corresponds to a point in  $\mathbb{R}^m$  that corresponds to its edge incidences i.e.,  $u_i(e) = 1$  if vertex  $i$  is an endpoint of edge  $e$ , and 0 otherwise. Note that  $\|u_i - u_j\|^2 = 2D - 2$  if  $(i, j) \in E$  and  $\|u_i - u_j\|^2 = 2D$  when  $(i, j) \notin E$ . Let  $k = \tau n$  where  $\tau \in (0, 1)$  is the parameter in Conjecture 18. Let  $R^* = \sqrt{D - \frac{D}{k}}$ , and  $R' = \sqrt{D - \frac{D}{2k}}$ . We first prove that it is SSE-hard to distinguish between the following two cases:

- (YES) case when there exists a ball of radius  $R^*$  that contains at least  $\delta n$  points,
- (NO) case when every ball of radius at most  $R'$  has at most  $\gamma\delta n$  points inside it.



*Completeness argument (YES case):* In the YES case of SSE hypothesis, there exists a subset  $S \subset V$  of size  $k$  with  $|E(S, S)| \geq \frac{kD(1-\eta)}{2}$ . Consider a candidate center

$$c^* = \begin{cases} \frac{2}{k} & \text{if } (i, j) \in E(S, S) \\ 0 & \text{otherwise} \end{cases}.$$

Let  $S^* = \{i \in S : \deg_S(i) \geq (1 - \sqrt{\eta})D\}$ . Since  $|E(S, S^c)| \leq \eta kD$ , by Markov's inequality,  $|S^*| \geq (1 - \sqrt{\eta})|S| \geq 3/4 \cdot |S| = \delta n$  where  $\delta = 3\tau/4$ . We have  $S^* \subset \text{Ball}(c^*, R^*)$  since

$$\begin{aligned} \forall i \in S^*, \quad \|u_i - c^*\|_2^2 &= \sum_{e \in E} \left(u_i(e) - \frac{2}{k}\right)^2 = (D - \deg_S(i)) + \deg_S(i) \left(1 - \frac{2}{k}\right)^2 \\ &\quad + (|E(S, S)| - \deg_S(i)) \left(\frac{2}{k}\right)^2 \\ &= D - \deg_S(i) \left(\frac{4}{k} - \frac{4}{k^2}\right) + (|E(S, S)| - \deg_S(i)) \cdot \left(\frac{4}{k^2}\right) \\ &\leq D + \frac{kd}{2} \cdot \frac{4}{k^2} - \frac{4(1 - \sqrt{\eta})D}{k} \leq D - (1 - 2\sqrt{\eta}) \frac{2D}{k} \\ &\leq (R^*)^2, \text{ since } \eta \leq 1/16. \end{aligned}$$

*Soundness argument (NO case):* Suppose there exists a subset of points corresponding to vertices  $S' \subset V$  with  $|S'| = k'$  for  $3\gamma k/4 \leq k' \leq k$  that are contained in a ball of radius  $R'$  around a point  $c' \in \mathbb{R}^n$ . Note that  $|S'| \geq \eta\tau n$  and  $|S'| = k' \geq 3\gamma k/4 = \gamma\delta n$ . We have

$$\begin{aligned} (R')^2 &= \max_{i \in S'} \|u_i - c'\|^2 \geq \text{Avg}_{i \in S'} \|u_i - c'\|^2 \geq \text{Avg}_{i \in S'} \left\| u_i - \text{Avg}_{i \in S'} u_i \right\|^2 \\ &= \frac{1}{2} \text{Avg}_{i, j \in S'} \|u_i - u_j\|^2 = D - \frac{2|E(S', S')|}{|S'|(|S'| - 1)}. \end{aligned}$$

Since  $3\gamma k/4 \leq |S'| \leq k$ , we have

$$\begin{aligned} D - \frac{2|E(S', S')|}{k'(k' - 1)} &\leq (R')^2 \leq D - \frac{D}{2k} \\ |E(S', S')| &\geq \frac{D}{4k} \cdot k'(k' - 1) \geq \frac{3\gamma(3\gamma k/4 - 1)D}{16} \geq \frac{\gamma^2}{8} kd > \eta D |S'|, \end{aligned}$$

since  $\gamma^2 > 8\eta$ . This contradicts the NO case of the SSE conjecture. This finishes the soundness argument. Thus, we have established the hardness of approximating the minimum radius ball containing  $\delta n$  points, even when we are allowed an arbitrary constant slackness factor  $\gamma > 0$  in the coverage. It is hard to approximate the radius within a factor of

$$\frac{R'}{R^*} = \sqrt{\frac{D - D/(2k)}{D - D/k}} \geq \sqrt{1 + \frac{1}{2k}} \geq 1 + \frac{1}{4m}.$$

Now as in Theorem 17 we can pad the instance with dummy coordinates to make it a  $d$  dimensional instance with  $d = (4m)^{1/\varepsilon}$  to get the desired  $1 + d^{-\varepsilon}$  inapproximability factor in radius. ■

## Appendix E. Application to Conformal Prediction

As an immediate application of our result, we obtain an algorithm and guarantee for conformal prediction with approximate volume optimality in the high-dimensional setting.

Conformal prediction is the statistical problem of finding prediction intervals. That is, given training examples  $Y_1, \dots, Y_n$  lying in some space  $\mathcal{Y}$ , and a target miscoverage rate  $\alpha > 0$ , our goal is to output a set  $C$ , such that for an unknown test example  $Y_{n+1} \in \mathcal{Y}$ ,

$$\mathbb{P}[Y_{n+1} \in C] \geq 1 - \alpha, \quad (16)$$

assuming *exchangeability* of the training examples and the test example, i.e.,

$$\mathbb{P}[Y_1 = y_1, \dots, Y_{n+1} = y_{n+1}] = \mathbb{P}[Y_1 = y_{\pi(1)}, \dots, Y_{n+1} = y_{\pi(n+1)}],$$

for all  $y_1, \dots, y_{n+1} \in \mathcal{Y}$  and permutations  $\pi$  over  $\{1, \dots, n+1\}$ .<sup>7</sup> We refer the reader to the book of [Angelopoulos et al. \(2024\)](#) for a thorough introduction and treatment of conformal prediction.

There are many ways to construct conformal predictors that achieve (16). For example, a trivial predictor can simply output the whole space  $\mathcal{Y}$ , and achieve coverage 1. The common measure of efficiency that is used to compare conformal predictors is volume. That is, if  $\mathcal{Y} = \mathbb{R}^d$ , then we would like a conformal predictor that minimizes  $\mathbb{E}[\text{vol}(C)]$ , where  $\text{vol}(C)$  is the Lebesgue measure of the set  $C \subseteq \mathbb{R}^d$ . Typically in the literature, a conformal predictor must *provably* achieve coverage (16) under only the weak condition of exchangeability, and has efficiency (volume) that is empirically validated on datasets.

Much of the work in conformal prediction has focused on the one-dimensional setting where  $\mathcal{Y} = \mathbb{R}$ . In these settings, it is clear that any natural set  $S \subseteq \mathbb{R}$  is a union of intervals. However, a recent line of work has explored the setting where  $\mathcal{Y} = \mathbb{R}^d$  is higher dimensional. In this setting, a priori it is not even clear what form the output set  $S \subseteq \mathbb{R}^d$  should take. [Wang et al. \(2023\)](#) and [Zheng and Zhu \(2025\)](#) provide methods to tackle the problem by outputting sets  $S$  that are unions of balls. They provide theoretical guarantees that their methods achieve coverage, and validate the efficiency of their methods empirically on real and synthetic data.

One may hope to design a conformal predictor that is *provably* volume optimal subject to achieving coverage (16). However, even when the points  $Y_i$  are drawn i.i.d. from some arbitrary distribution  $\mathcal{D}$  over  $\mathbb{R}^d$ , the problem of finding the minimum volume set  $C \subseteq \mathbb{R}^d$  such that

$$\mathbb{P}[Y_{n+1} \in C] = \mathbb{P}_{y \sim \mathcal{D}}[y \in C] \geq 1 - \alpha$$

is statistically intractable even for  $d = 1$ . Thus any provable guarantee for volume optimality must restrict the problem in some way. [Gao et al. \(2025\)](#) observes that the problem becomes statistically tractable (for i.i.d. samples) when we restrict our set  $C$  to come from some class of bounded VC-dimension  $\mathcal{C}$ . That is, we compete with

$$\min_{C \in \mathcal{C}} \text{vol}(C) \quad \text{s.t.} \quad \mathbb{P}_{y \sim \mathcal{D}}[y \in C] \geq 1 - \alpha,$$

7. We refer to this setting as the *unsupervised* setting. In the more general *supervised* setting, our training examples are feature-label pairs  $(X_i, Y_i) \in \mathbb{X} \times \mathcal{Y}$ , and our task is given  $X_{n+1}$  to output a set  $C(X_{n+1})$  that contains the (unknown)  $Y_{n+1}$  with probability  $\geq 1 - \alpha$ , assuming exchangeability of the training and test examples. Strategies for the unsupervised setting often translate to the supervised setting where a regression model is trained to predict  $Y_i$  from  $X_i$ , by applying the unsupervised conformal inference procedure on the *residuals* (error between the true  $Y_i$  and predicted  $\hat{Y}_i$ ), or by considering the conditional distribution  $Y_i \mid X_i$ . See e.g., [Angelopoulos et al. \(2024\)](#); [Gao et al. \(2025\)](#) for details.

which they term  $\mathcal{C}$ -restricted volume optimality. In fact, in this setting, the problem essentially reduces to finding the minimum volume set in  $\mathcal{C}$  that achieves coverage  $1 - \alpha$  assuming the samples are i.i.d., which is precisely the problem that we solve (approximately) in this work. A conformal predictor must additionally satisfy (16) when the samples are exchangeable, but this can be done with a standard “conformalizing” step.

This gives us the following sample theorem when the class  $\mathcal{C}$  is the set of Euclidean balls. (Equivalent statements go through for the other settings we consider in this work: properly learning Euclidean balls, and properly/improperly competing with unions of balls. See Remark 20.)

**Theorem 5 (Conformal Prediction with Approximate Volume Optimality)** *We have an algorithm for conformal prediction over examples from  $\mathcal{Y} = \mathbb{R}^d$ , that achieves approximate volume optimality with respect to the set  $\mathcal{C}$  of Euclidean balls.<sup>8</sup> That is, we have an algorithm that, given training examples  $Y_1, \dots, Y_n \in \mathcal{Y}$ , miscoverage rate  $0 < \alpha < 1$ , and coverage slack factor  $0 \leq \gamma \leq 1$ , outputs a set  $\hat{C}$  (not necessarily in  $\mathcal{C}$ ), such that for an unknown test example  $Y_{n+1} \in \mathcal{Y}$ ,*

(a) *if  $Y_1, \dots, Y_{n+1}$  are exchangeable, then*

$$\mathbb{P}[Y_{n+1} \in \hat{C}] \geq 1 - \alpha.$$

(b) *if  $Y_1, \dots, Y_{n+1}$  are drawn i.i.d. from some (unknown) distribution  $\mathcal{D}$ , and  $n = \Omega(d^2/\gamma^2)$ , then*

$$\text{vol}(\hat{C})^{1/d} \leq \left(1 + O_{\gamma,\delta}(d^{-1/2+o(1)})\right) \text{vol}(C^*)^{1/d},$$

where

$$C^* = \underset{C \in \mathcal{C}}{\text{argmin}} \text{vol}(C) \quad \text{s.t.} \quad \mathbb{P}[Y_{n+1} \in C] \geq 1 - \alpha + \gamma.$$

**Proof** Without loss of generality, we assume that  $n$  is even. Theorem 1 gives an algorithm that satisfies case (b). That is, given samples  $Y_1, \dots, Y_{n/2}$  drawn i.i.d. from  $\mathcal{D}$ , for  $n = \Omega(d^2/\gamma^2)$ , it outputs a set

$$\text{vol}(\hat{C}) \leq \left(1 + O_{\gamma,\delta}(d^{-1/3+o(1)})\right) \text{vol}(C^*),$$

where

$$C^* = \underset{C \in \mathcal{C}}{\text{argmin}} \text{vol}(C) \quad \text{s.t.} \quad \mathbb{P}_{y \sim \mathcal{D}}[y \in C] \geq 1 - \alpha + \gamma.$$

Since  $\mathbb{P}_{y \sim \mathcal{D}}[y \in C] = \mathbb{P}[Y_{n+1} \in C]$ , this satisfies case (b).

Now, to construct a conformal predictor, it suffices to give a *conformity score*.<sup>9</sup> We can do this by constructing a *nested set system*, following the strategy in Gupta et al. (2022).<sup>10</sup> Let  $\mu \in \mathbb{R}^d$  be the center of  $\hat{C}$ . We define the natural scaling of  $\hat{C}$  by a scalar  $\lambda \geq 0$  as

$$\lambda \hat{C} = \{\lambda(x - \mu) + \mu \mid x \in \hat{C}\}. \quad (17)$$

For  $0 < \tau < 1$ , define

$$\lambda_\tau = \underset{\lambda \geq 0}{\text{argmin}} \text{ s.t. } \left| \{Y_{n/2+1}, \dots, Y_{n+1}\} \cap \lambda \hat{C} \right| \geq \tau n/2. \quad (18)$$

8. This is an approximate form of *restricted volume optimality* defined by Gao et al. (2025).

9. See Definition 3.1 in Angelopoulos et al. (2024).

10. See Assumption 2.4 in Gao et al. (2025) for a formal statement.

That is,  $\lambda_\tau$  is the smallest scaling of  $\widehat{C}$  that achieves coverage  $\tau$  over the training samples. Our nested set system will consist of  $\lambda_\tau \widehat{C}$  for values  $\tau$  from a grid over the interval  $[0, 1]$ . It will also contain  $\widehat{C}$  itself. Since all sets in this system are scalings of the same convex set centered at the same point, they are indeed nested.

We can use this set system to construct a conformity score, see e.g. equation (10) and Assumption 2.4 in [Gao et al. \(2025\)](#), which implies a conformal predictor via split conformal prediction, see e.g. Algorithm 3.6 in [Angelopoulos et al. \(2024\)](#). These have the property that they will only output sets  $C$  from the input nested set system. Thus, we have that

(a) when  $Y_1, \dots, Y_{n+1}$  are exchangeable, then we achieve coverage

$$\mathbb{P}[Y_{n+1} \in \widehat{C} \geq 1 - \alpha].$$

(b) if  $Y_1, \dots, Y_{n+1}$  are drawn i.i.d. from some (unknown) distribution  $\mathcal{D}$ , and  $n = \Omega(d^2/\gamma^2)$ , then since  $\widehat{C}$  is in our nested set system, and Theorem 1 guarantees that  $\widehat{C}$  achieves coverage  $\geq 1 - \alpha$ , the conformal predictor will not output any set in the system larger than  $\widehat{C}$ .

■

**Remark 20** *The main step that is necessary to “conformalize” our approximation result is to create a nested set system. That is, given the output  $\widehat{C}$  of our approximation algorithm, which is a set that provably achieves coverage  $\geq 1 - \alpha$  over  $Y_{n+1}$  if the samples  $Y_1, \dots, Y_n$  were drawn i.i.d. from some (unknown) distribution  $\mathcal{D}$ , we must create a family of nested sets  $\widehat{\mathcal{C}}$ , such that:*

(a)  $\widehat{C} \in \widehat{\mathcal{C}}$ ,

(b) *For a wide range of coverage levels  $\tau$  in some set  $T$ ,<sup>11</sup> we have that there exists a set  $C_\tau \in \widehat{\mathcal{C}}$  such that*

$$|\{Y_{n/2+1}, \dots, Y_{n+1}\} \cap \lambda C_\tau| \geq \tau n,$$

*and ideally this quantity is close to  $\tau n/2$ .*

*All of the forms of  $\widehat{C}$  that we work with in this work have natural notions of scaling (equivalent of Equation (17)) that result in nested sets. That is, for balls we can use the same scaling as Equation (17), and for unions of balls/ellipsoids, we can scale each set in the union individually. Thus there is a natural way to create this set system, and the rest of the argument from Equation (18) onward goes through unchanged to get the equivalent guarantees.*

11. Technically speaking, to achieve coverage (16), we only need that  $T$  contains some coverage level close to 1. However, for the conformal predictor to be efficient, it is good to think of  $T$  as being a grid over  $[0, 1]$ .