High-Dimensional Convexity Testers

Eric Liu

December 31, 2023

1 Introduction

This project is about testing for convexity in high dimensions. This report will give a survey of various different notions of convexity and property testers for them. The main goal of this project was to explore further avenues for extending the results by Chen et al. in [1]. In the setting of their paper, there is a set $S \subseteq \mathbb{R}^n$, and we would like to test if S is convex or ϵ-far from convex. The set S is ϵ-far from convex if for every convex set $D \subseteq \mathbb{R}^n$, we have that $\text{vol}(D \Delta S) > \epsilon$, where $D \Delta S$ denotes the symmetric difference and vol is the Gaussian volume. In their paper, they gave a sample-based testing algorithm: the algorithm has access to independent draws $(x, S(x))$, where $x \sim \mathcal{N}(0, I^n)$, and, abusing notation, we let S be the indicator function for the set S. The natural open question is to give an algorithm that can also utilize membership oracle access (i.e. queries, or an oracle which takes as input any point x in $\mathbb{R}^n$ and returns whether $x \in S$ or not) to give an even more efficient algorithm. We will spend much of this report discussing related work, most importantly including a recent line of work that highlights an analogy between monotone boolean functions and high-dimensional convex sets in $\mathbb{R}^n$.

We will also discuss two other settings: one, due to Rademacher and Vempala [3], where the algorithm additionally has a random oracle that returns a point uniformly at random from the set, and the other by Black et al. [4], where they test discrete convexity over the ternary hypercube.

2 CFSS Setting

We will briefly recall the results and techniques of the CFSS [1] paper. For one-sided testing, they gave a $2^{\Omega(n)}$ sample lower bound and $2^{O(n \log(n/\epsilon))}$ sample upper bound. For two-sided testing, they gave a $2^{\Omega(\sqrt{n})}$ sample lower bound and a $2^{O(\sqrt{n} \log(n)/\epsilon^2)}$ sample upper bound.

For the one-sided upper bound, they use a gridding approach, classifying each cell as external, internal, or boundary. These, loosely speaking, refer to where the cell is relative to the set. The key structural lemma used is a bound on the volume of $\partial S + \text{Ball}(\alpha)$, the so-called "$\alpha$-thickened boundary" of a convex set, which is then used in the algorithm to say that if we have too many boundary cubes, we can reject. This result was generalized by Harms and Yoshida [5] to any arbitrary product distribution. They show that this idea of "downsampling", a generalization of gridding, can be applied to various problems by analyzing problem-dependent isoperimetric quantities.

Returning to the original paper, for the lower bound, they used properties of convex sets in high dimension to show that after sampling 2^{cn} points, with high probability none of the points would be in the convex hull of the rest of the points. This means the class of convex sets shatters the sampled points. This approach is similar to (and precedes) the approach by Blais et al. which uses (lower) VC Dimension to give lower bounds on distribution-free sample-based testing [6]. This included a polynomial lower bound on intersections of k halfspaces. However, their method cannot be directly applied, since the VC Dimension of convex sets is infinite.

Relatedly, the Blais paper discusses that two-sided (sample-based) testing is not much more efficient than learning. This is exactly the approach used in the CFSS paper, which gives a two-sided upper bound by testing by implicit learning.

Thus far, the techniques we have seen heavily rely on the fact that in this model we are only given sample access. We saw one way this manifested in that the proofs all had later generalizations. The proof of the two-sided lower bound uses Yao's lemma and is more of an ad-hoc construction. I spent some time trying to see if it could provide lower bounds for testing algorithms with access to queries. The reason why I thought about this is that the construction over the 'no' instances is essentially a one-dimensional object.

In their proof, there were three distributions over $\mathbb{R}^n \times \{0,1\}$ analyzed: $\mathcal{E}_{yes}$, $\mathcal{E}^*_{no}$, and $\mathcal{E}_{no}$. In the first, a random set $\mathbf{S}$ is sampled as the intersection of $2^{\sqrt{N}}$ random halfspaces, defined by normal vectors chosen uniformly at random from a sphere of radius $\Theta(n^{1/4})$. Then the samples $(x, \mathbf{S}(x))$ are returned. In the distribution $\mathcal{E}^*_{no}$, each sample is marginally the same as in the previous distribution, but they are independent; there is no underlying set $\mathbf{S}$. We have that the total variation distance between these two distributions is $o(1)$, and the proof relied on some elementary probabilistic and convex geometric arguments which took a few pages. To me, even though this was the bulk of the technical arguments of the paper, it was not very illuminating. In fact, the construction for the random set $\mathbf{S}$ was originally used by Nazarov [7] to give a set with large Gaussian surface area, but in his paper he also mentioned that the arguments were tedious and not very insightful. Thus, it would be interesting to see if there were a way to simplify either of these arguments.

The next step is to define $\mathcal{E}_{no}$ and show that the total variation distance between it and $\mathcal{E}^*_{no}$ is $o(1)$. The definition is given by partitioning the ball of radius $2\sqrt{n}$ into at least $2^{\sqrt{n}}$ shells of equal volume, and then choosing each shell according to the marginal distribution in the previous two distributions. The union of the chosen sets is our random set. Then the idea is that with high probability, no two samples lie in the same shell, and since the shells were defined via the marginals of $\mathcal{E}^*_{no}$, then the TV-distance is not too far. However, because we are including each shell uniformly at random, with high probability the constructed set is far from convex. The analysis of this is given by taking the intersection of the random set with any line segment. In this sense, the construction is one dimensional: the only variable that determines membership is the distance to the origin. A simplistic way of viewing this is that perhaps an object which is higher dimensional could provide lower bounds to algorithms with query access, but the barrier is that with queries, we no longer can assume independence of the samples, and thus the distance to $\mathcal{E}^*_{no}$ no longer holds. On the other hand, it seems reasonable that the Nazarov construction is still what should be used for $\mathcal{E}_{yes}$.

3 Boolean Analogy

In a recent line of work, De, Nadimpalli, and Servedio have outlined various examples in which monotone boolean functions and symmetric convex functions are analogous [2]. This line of work was my reason for choosing this topic. The sample-based model of testing convex sets that we have discussed so far is motivated by the sample-based learning perspective, originally studied for learning boolean functions. We will contrast this later to the model by Vempala.

Their simplest evidence for why there is an anology between monotone boolean functions and symmetric convex sets is that "moving an input up towards 1^n" will never decrease the value of a monotone boolean function, similar to how for a centered symmetric convex set, moving an input towards the origin will never decrease the value. On the one hand, what they describe is only evidence for star-convexity being the analogy (which is why I considered the problem of testing star-convexity for a bit). On the other hand, they have ample actual analogous theorems between the structures in their stated analogy.

For instance, in [2], De and Servedio show an analogy to the Kruskal-Katona theorem, which is a quantitative statement about the density increment for monotone boolean functions. That density

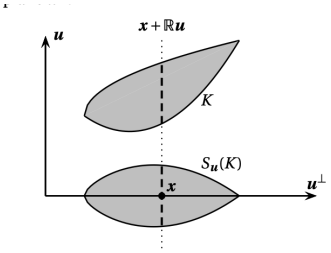

Figure 1: Steiner Symmetrization (taken from [11])

increment is about how the fraction of weight-j inputs which output 1 increases as j increases. They show a similar behavior holds for convex sets, and use this to give a *weak learner* for convex sets. This highlights one of the differences between property testing and weak learning: in the case of weak learning, for a target function f, you are only required to output a hypothesis h such that $Pr[h(x) = f(x)] \geq 1/2 + 1/\text{poly}(n)$. In this case, they only need to use three functions are hypotheses: the two constant functions, and the majority function, which they note is analogous to the case of weak learning monotone boolean functions.

On the other hand, they show important structural results related to the (Fourier/Hermite) spectrum, which was used in the strong learning of monotone boolean functions/convex sets [9] (and hence in the two-sided upper bound for sample testing). More specifically, it is known that the Fourier spectrum of an n-variable monotone boolean function is concentrated in the first $O(\sqrt{n})$ levels, and they show that the same concentration holds for the *Hermite* spectrum of indicator functions for convex sets. The details of the Hermite basis are not required here; but it is, as they argue, the appropriate basis to use for the dual space when dealing with convex sets.

Testing monotonicity and learning convex sets have both been the subjects of intense study. Thus, given this analogy, the topic of property testing high dimensional convex sets is well motivated.

The simplest tester for monotonicity is the edge tester, and the proof of correctness uses a shifting argument to relate the number of violating edges to the distance to monotonicity. It is also known that queries are necessary to achieve polynomial size bounds. Thus, one could hope to somehow use these ideas in the case of convexity testing. The problem is that, despite the analogy, there are only few concrete examples for objects that correspond under the analogy, and there is no real canonical way to transfer between the two settings. There are many results in Boolean Analysis that use approximation by Gaussian/the invariance principle [10], but they will not help us here.

One simple idea to try is a "midpoint-tester", which samples three collinear points, and checks for a violation to convexity. However, the problem is that given two points, their midpoint is likely to be in the inner ball, which we always assume to be part of the set without loss of generality. More specifically, since most of the Gaussian mass is concentrated in a shell of width $c \log n$ around the radius $\sqrt{n}$, we can assume the set contains $\text{Ball}(\sqrt{n} - c \log n)$ and excludes any point which has norm greater than $\sqrt{n} + c \log n$. Choosing magnitudes is one of the difficulties of testing in the continuous setting.

We will next discuss a possible analogue to the shifting operator: *Steiner Symmetrization*. The idea is that for our body S, given a unit vector u, we take each point $x \in u^{\perp}$ and shift the interval $(x + \mathbb{R}u) \cap S$ until it is centered around x. This technique was used to show that among all bodies of a given volume, the ball minimizes surface area. Accordingly, one needs to define a sequence of vectors with which to symmetrize in order to approach a ball with respect to the *Hausdorff* metric [11]. In recent work (and also towards the goal of property testing convexity), De, Nadimpalli, and Servedio investigated approximation of a convex set by polytope with respect to Gaussian distance [12] instead of Hausdorff. The other basic unanswered question to this approach is that even if we had some finite number of symmetrizations, it is still not clear what the tester should query.

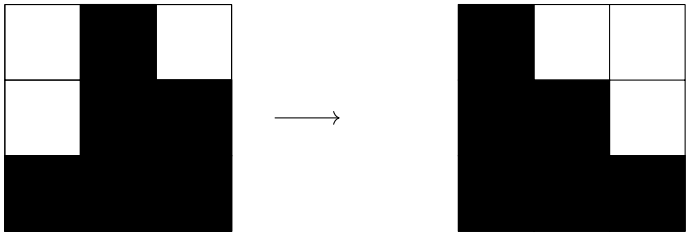

Figure 2: Monotone Rearrangement

Along the lines of isoperimetry, we will revisit how it was used in the CFSS paper. There, they explicitly used the α-thickened surface area, which built on Gaussian surface area (GSA), and implicitly used the Nazarov construction, which had large GSA. Indeed, the paper by Klivans et al. [9] argues that GSA is a natural measure of complexity for learning convex concepts. We will next discuss the relationship to influence and noise sensitivity, two fundamental concepts in Boolean Analysis.

The Ornstein-Uhlenbeck operator, defined for $0 \leq \rho \leq 1$, is given by $(T_\rho f)(x) = E_{\mathbf{y} \sim \mathcal{N}^n}[f(\rho x + \sqrt{1-\rho^2}\mathbf{y})]$ is analogous to the Bonami-Beckner noise operator for boolean functions, and acts similarly on the Hermite basis: $T_\rho f = \sum_{S \in \mathbb{N}^n} \rho^{|S|} \hat{f}(S) H_S$. The noise sensitivity for our convex body K is defined as $NS_\delta(K) = \frac{1}{2} - \frac{1}{2}\langle K, T_{1-\delta} K \rangle$. The connection to GSA is given by a theorem which states $NS_\delta(K) \leq \sqrt{\pi}\sqrt{\delta}\text{GSA}(K)$. The importance of these noise operators is that they give a smooth interpolation when sometimes the objects are nasty but could be analyzed more easily with noise.

The other case is with influence: directed influence has been applied to get relatively tight bounds for monotonicity testing for the boolean hypercube. On the other hand, convex influences were defined by De et al. [2], but have not found a direct application to testing convexity.

We will now discuss a monotonicity tester in the continuous setting, which does not exactly fit in to the Boolean-Convex analogy, but shows some difficulty of testing on continuous domains. Given a function $f : \mathbb{R}^n \to \mathbb{R}$, which is guaranteed to be Lipschitz, test if it is monotone or ϵ-far from monotone under L_1 distance [13]. Their work is analogous to the edge tester for monotonicity. Their techniques include using a monotone rearrangement that is similar to shifting, but in the continuous setting. Indeed, they have to check all the same things hold under monotone rearrangement that hold using the shifting operator, e.g. decreasing distance to monotonicity, different order of operations being applied still preserves monotonicity. Additionally, they must check that the Lipschitz property is preserved.

Given their structural theorems, their tester also assumes partial derivative query access, and the test is done by sampling a point uniformly at random and sampling a partial derivative in a random direction, rejecting if the partial derivative is negative. Their result is that, assuming the Lipschitz constant is L and the domain is $[0,1]^n$, the query bound is $\Theta(\frac{nL}{\epsilon})$.

Assuming the partial derivative oracle and also the Lipschitz condition are two relaxations that made testing monotonicity tractable. In the case of convex sets, their indicator functions are 0/1 and thus are not continuous and in some sense have infinite Lipschitz constant. On the other hand, the n-dimensional Gaussian measure of the boundary can be assumed to be 0. This is in contrast to the proof of the continuous monotonicity tester, which argues correctness using the following: given a function f which is ϵ-far from monotone, letting the sets S_i for $i \in [n]$ be given by $S_i = \{x \in [0,1]^n : \partial_i f(x) < 0\}$, we have that $\sum_{i=1}^n \mu(S_i) > \frac{\epsilon}{2L}$. In an extremely informal sense, integrating the 'infinite' Lipschitz constant over the boundary is the Gaussian surface area, which could be evidence for GSA as the correct notion of complexity for testing.

In an attempt to make this more rigorous, I considered the testing $T_\rho K$ instead of K itself. It is known that $T_\rho K$ is a smooth function, but I did not attempt to determine the Lipschitz constant, if it even is tractable. A simple toy calculation is for a step function $f : \mathbb{R} \to \mathbb{R}$ which is 0 on the negatives

and 1 on the nonnegatives. Then for $x \geq 0$ we have that $T_\rho f(x) = Pr_{\mathbf{y} \sim \mathcal{N}(0,1)}(\rho x + \sqrt{1-\rho^2}\mathbf{y} \geq 0)$. I did not further pursue this direction for two reasons. The first was that doing the relevant calculations seemed difficult. In fact, one reason why I considered this angle was a paper I read over the summer by Lee and Eldan [14] which showed that "under the Ornstein-Uhlenbeck semigroup, any non-negative measurable $f : \mathbb{R}^n \to \mathbb{R}$ exhibits a uniform tail bound better than implied by the Markov inequality". Their paper required heavy usage of stochastic calculus. (Actually, after re-reading the abstract just now, I wonder if the fact that $T_\rho K$ is semi-log-convex, i.e. the eigenvalues of $\nabla^2 \log T_\rho K$ are at least $-\beta$ for some β, could be leveraged somehow). The other reason I did not pursue this direction further was the polynomial relationship mentioned earlier between noise sensitivity and GSA, which means that in some sense this idea was already implicitly used.

The question still remains what relaxations are possible for convexity testing. For example, the indicator function for convex sets are *log-concave*, where a function $f : \mathbb{R}^n \to \mathbb{R}$ is log-concave if for any $\theta \in [0, 1]$, we have $f(\theta x + (1-\theta)y) \geq f(x)^\theta f(y)^{1-\theta}$. Perhaps we could consider testing Lipschitz log-concave functions.

4 Rademacher and Vempala Setting

We now briefly discuss a different setting. We once again have a set $S \subseteq \mathbb{R}^n$ and want to determine if it is convex or far from convex. However, in this new setting, they ask whether $\text{vol}(S \Delta D) \geq \epsilon \text{vol}(S)$, where vol is the Lebesgue volume. Also, their algorithm uses a membership oracle and a random oracle, where a random oracle returns a point in S uniformly at random. In *this* setting, the midpoint tester mentioned before (they use a similar line tester) makes sense, due to the random oracle access. Additionally, due to their definition of distance, the volume of S can be exponentially small. This contrasts with the original setting, where we may assume that the volume of S is constant. Thus, their ideas do not transfer over. For example, in a follow-up work [16], one of their lower bounds in is given by the union of two truncated cones, which is not far from convex in our original setting. Whereas the original setting is motivated by sample-based learning, their random oracle access can be implemented using Monte Carlo Markov Chains. This itself uses a membership oracle, but if S is not convex, then there are no guarantees that the Markov Chain will output points uniformly at random. This line of work was originated by Dyer et al [15].

5 Discrete Convexity

There are various notions of discrete convexity. The one focused on by Black et al. [4] is given by the following: a discrete set $S \subseteq \{-1, 0, 1\}^n$ is convex if there exists a convex set $C \subseteq \mathbb{R}^n$ such that $C \cap \{-1, 0, 1\}^n = S$. The question they are testing is whether a discrete subset of the ternary hypercube S is convex or ϵ-far from every convex set, i.e. for every set D we have that $\text{vol}(D \Delta S) \geq \epsilon$, where vol is given by the uniform distribution. They study the ternary hypercube because in the boolean hypercube, all sets are convex. On the other hand, for an m-ary hypercube, if $m = \text{poly}(n)$, then they claim the gridding results mentioned earlier can be used. On the other hand, for a constant $m > 3$, it is still open. For motivation of this notion of discrete convexity, we see that this is exactly the sort of relation that occurs when relaxing an integer linear program to a linear program.

For their results, they have structural results on the maximum edge boundary/influence of a convex set, and use this to show that one-sided sample-based testing has complexity $3^{\Theta(n)}$. Additionally, in contrast to the lack of query testing results in our original setting, they show nearly matching upper and lower bounds of $3^{\tilde{\Theta}(\sqrt{n})}$ for a non-adaptive query tester. They also have learning/two-sided upper and lower bounds of $3^{\tilde{O}(n^{3/4})}$ and $3^{\Omega(\sqrt{n})}$.

Many of their results build on concepts that were mentioned earlier in this report, but their results do not apply to the continuous setting. Although the discrete convex case is interesting because the

sets can act counterintuitively, e.g. not being connected, being in the discrete world allowed them to leverage the large body of work already done there. For example, for a high influence set, they took the construction from Kane [17], which was essentially a discretization of the Nazarov construction. For the sample-based tests, they were able to prove Fourier concentration for convex sets and use low-degree tests. For the lower bound, they adapted Talagrand's random DNFs, which have been well studied in property testing literature. Their nonadaptive algorithm is only needs to query points 'above' randomly sampled points; in the continuous world, one would have to worry about step sizes and potentially skipping over gaps.

Their most interesting idea is their relation of influence to the sign-changes of a random walk. Since a discrete convex set S can be written as the intersection of finitely many half spaces, one can take a certain outwardly moving random walk $(\mathbf{X}^{(s)})_s$. To be clear, we are not choosing neighbors uniformly at random. Next, for each halfspace, indexed by i define the walk $\mathbf{W}_i(s) = \langle \mathbf{X}^{(s)}, v^i \rangle - \tau_i$ for each halfspace. Then because S is the intersection of halfspaces, we are interested in $\max \mathbf{W}_i$.

6 Future Work

Given the discussion on the Monotone Boolean-Convex analogy, and the previous discussion on relating random walks to influence, it seems like further using stochastic processes such as the Ornstein-Uhlenbeck operator could lead to some further insights. Additionally, there was recent work by De, Nadimpalli, and Servedio [12] about approximating convex sets by polytopes, which seems promising, and towards the direction of convexity testing.

References

[1] Xi Chen, Adam Freilich, Rocco A. Servedio, and Timothy Sun. *Sample-based high-dimensional convexity testing*. 2017. `https://arxiv.org/abs/1706.09362`. *arXiv:1706.09362 [cs.CC]*.

[2] Anindya De, Shivam Nadimpalli, and Rocco A. Servedio. *Convex Influences*. 2021. `https://arxiv.org/abs/2109.03107`. *arXiv:2109.03107 [cs.CC]*.

[3] László Rademacher and Santosh Vempala. *Testing Geometric Convexity*. In: Kamal Lodaya and Meena Mahajan (eds), *FSTTCS 2004: Foundations of Software Technology and Theoretical Computer Science*, Lecture Notes in Computer Science, vol. 3328, Springer, Berlin, Heidelberg, 2004. `https://doi.org/10.1007/978-3-540-30538-5_39`.

[4] Hadley Black, Eric Blais, and Nathaniel Harms. *Testing and Learning Convex Sets in the Ternary Hypercube*. 2023. `https://arxiv.org/abs/2305.03194`. *arXiv:2305.03194 [cs.DS]*.

[5] Nathaniel Harms and Yuichi Yoshida, *Downsampling for Testing and Learning in Product Distributions*, 2021, `arXiv:2007.07449 [cs.DS]`.

[6] Eric Blais, Renato Ferreira Pinto Jr., and Nathaniel Harms, *VC Dimension and Distribution-Free Sample-Based Testing*, 2020, `arXiv:2012.03923 [cs.LG]`.

[7] F. Nazarov, *On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a Gaussian measure*, In *Geometric aspects of functional analysis (2001-2002)*, pages 169–187, Lecture Notes in Math., Vol. 1807, Springer, 2003.

[8] Anindya De and Rocco A. Servedio. *Weak Learning Convex Sets under Normal Distributions*. In *Conference on Learning Theory, COLT 2021*, volume 134 of *Proceedings of Machine Learning Research*, pages 1399–1428. PMLR, 2021.

[9] Adam R. Klivans, *Learning geometric concepts via Gaussian surface area*, In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, Pages 541-550, Publisher: IEEE, Publication date: October 25, 2008.

[10] Ryan O'Donnell, *Analysis of Boolean Functions*, Cambridge University Press, 2014.

[11] Thomas Rothvoss. *Asymptotic Convex Geometry.* Lecture Notes, 2021. `https://sites.math.washington.edu/~rothvoss/lecturenotes/AsymptoticConvexGeometry-30-AUG-2021.pdf`.

[12] Anindya De, Shivam Nadimpalli, and Rocco A. Servedio, *Gaussian Approximation of Convex Sets by Intersections of Halfspaces*, 2023, `arXiv:2311.08575 [cs.CC]`.

[13] Renato Ferreira Pinto Jr., *Directed Poincaré Inequalities and L^1 Monotonicity Testing of Lipschitz Functions*, 2023, `arXiv:2307.02193 [cs.DS]`.

[14] Ronen Eldan and James R. Lee, *Regularization under diffusion and anticoncentration of the information content*, *Duke Mathematical Journal*, vol. 167, no. 5, April 2018, ISSN: 0012-7094, DOI: 10.1215/00127094-2017-0048.

[15] Martin Dyer, Alan Frieze, and Ravi Kannan, *A random polynomial-time algorithm for approximating the volume of convex bodies*, *Journal of the ACM (JACM)*, vol. 38, no. 1, pp. 1–17, January 3, 1991, Publisher: ACM.

[16] Eric Blais and Abhinav Bommireddi, *On Testing and Robust Characterizations of Convexity*, In *APPROX/RANDOM 2020*, LIPIcs 176, pp. 18:1–18:15, DOI: 10.4230/LIPIcs.APPROX/RANDOM.2020.18.

[17] Daniel M. Kane, *The average sensitivity of an intersection of half spaces*, In David B. Shmoys, editor, *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 437–440, ACM, 2014.