SIAM J. MATH. DATA SCI.
Vol. 0, No. 0, pp. 000–000

Stochastic Tverberg Theorems With Applications in Multiclass Logistic Regression, Separability, and Centerpoints of Data*

Jesus A. De Loera† and Thomas Hogan†

Abstract. We present new stochastic geometry theorems that give bounds on the probability that m random data classes all contain a point in common in their convex hulls. These theorems relate to the existence of maximum likelihood estimators in multinomial logistic regression, to the separability of data, and to the computation of centerpoints of data clouds.

Key words. logistic regression, generalized linear models, maximum likelihood estimation, high-dimensional logistic regression, Tverberg's theorem, separability of data, geometric probability, combinatorial convexity, computational geometry in statistics, depth of data point, centerpoints, Tukey median

AMS subject classifications. 52C45, 60D05, 62H30, 62J02, 68T10

DOI. 10.1137/19M1277102

1. Introduction.

Convex geometry has been shown to be a valuable tool for the mathematical foundations of data science (see, e.g., [1, 2, 3, 5, 13] and the many geometric references there). Our paper further develops connections between combinatorial convex geometry and supervised learning. Before we discuss our results, we discuss our key motivation.

Logistic regression is a well-known nonlinear model in multivariate statistics and supervised learning [14]. Statistical inference and classification for this model relies on the theory of maximum likelihood estimation. In the multiclass logistic regression we have n observations $(\boldsymbol{x}_i, y_i), i = 1, \dots, n$. Here $\boldsymbol{x}_i$ is a vector of d real or discrete variables and y_i is a variable that takes discrete values $c_1, c_2, \dots, c_m$, the indicators of class membership. The logistic approach connects the response in y_i to the covariates $\boldsymbol{x}_i$ via the model

$$\mathbb{P}(y_i = c_s | \boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i' \boldsymbol{\beta}_s)}{\sum_{l=1}^m \exp(\boldsymbol{x}_i' \boldsymbol{\beta}_l)},$$

where the entries of the vectors $\boldsymbol{\beta}_s \in \mathbb{R}^d$ are the regression coefficients. We let E_s be the set of data point indices i, where $\boldsymbol{x}_i$ was labeled as $y_i = c_s$. In this model, the *log-likelihood* is given by the function

$$l(\boldsymbol{\beta}) = \sum_{l=1}^m \sum_{i \in E_j} \log \left(1 / \sum_{t=1}^m \exp(\boldsymbol{x}_i'(\boldsymbol{\beta}_t - \boldsymbol{\beta}_j)) \right).$$

*Received by the editors July 24, 2019; accepted for publication (in revised form) September 8, 2020; published electronically DATE.
https://doi.org/10.1137/19M1277102
Funding: This work was partially supported by NSF grants DMS-1522158 and DMS-1818169 and an NSF TRIPODS gran (NSF award CCF-1934568).

†Department of Mathematics, University of California at Davis, Davis, CA 95616 (deloera@math.ucdavis.edu, tahogan@math.ucdavis.edu).

A *maximum likelihood estimate* (MLE) is defined to be any maximizer of this function. Finding conditions for the existence of the MLE amounts to specifying the conditions under which MLE-based inference for these important models is possible. Unfortunately, one difficulty arising in machine learning is that the MLE for the logistic model does not exist in all situations.

There are two common approaches for multi-class classification: *"one-versus-rest"* and *"one-versus-one."* In "one-versus-rest," we train C separate binary classification models. Each classifier f_c for $c \in \{1, \dots, C\}$ is trained to determine whether or not an example is part of class c or not. To predict the class for a new sample $\boldsymbol{x}$, we run all C classifiers on $\boldsymbol{x}$ and choose the class with the highest score: $\hat{y} = \arg\max_{c \in \{1,\dots,C\}} f_c(\boldsymbol{x})$. In "one-versus-one" regression, we train $\binom{C}{2}$ separate binary classification models, one for each possible pair of classes. To predict the class for a new sample $\boldsymbol{x}$, we run all $\binom{C}{2}$ classifiers on $\boldsymbol{x}$ and choose the class with the most votes.

Albert and Anderson proved that the existence of the MLE is determined by how the data classes overlap [1, 21]. In "one-versus-one" multiclass logistic regression, the desired MLEs exist if, for any pair of classes, say a and b, the convex hull of the samples from class a intersect with the convex hull of the samples from class b. Similarly, for "one-versus-rest," the desired MLEs exist if the convex hull of the samples from any single class intersects with the convex hull of the remaining points. Although an appealing criterion for existence, this geometric characterization leads to another practical question: *How much training data do we need, as a function of the dimension of the covariates of the data, before we expect the MLEs for multiclass logistic models to exist with high probability?*

This question can be partially addressed using the geometric work of Cover [8] (adapting a technique originally due to Schläfli [20]), an investigation of class overlaps in random bicolorings for any large point set in general position, a sort of stochastic extension of Radon's theorem. Suppose that the n points $\boldsymbol{x}_i \in \mathbb{R}^d$'s are drawn i.i.d. from a continuous probability distribution and labeled into two classes with equal probability, i.e., $m = 2, c_1 = 1, c_2 = -1$ with $\mathbb{P}(y_i = 1 | \boldsymbol{x}_i) = 1/2$. If we assume that class labels are independent from $\boldsymbol{x}_i$, Cover's result can be used to show that under an asymptotic regime where d and n grow but $d/n \to k$ with probability tending to one, the convex hulls of the two types of colored points overlap if $k < 1/2$, or they are separated if $k > 1/2$. When the class labels are not independent from the $\boldsymbol{x}_i$, the problem is more difficult. In this case, Candès and Sur [5] proved that a similar phase transition occurs, and is parametrized by two scalars measuring the overall magnitude of the unknown sequence of regression coefficients. Their analysis of the general two-class logistic regression problem relies on the analysis of the geometry of convex cones from [2].

Our paper presents generalizations and variations of Cover's results to more than two colors using *Tverberg partitions*. A partition of a data set into $m \geq 2$ classes is Tverberg if the intersection of all the convex hulls of all the classes is nonempty. Tverberg's theorem guarantees Tverberg partitions exists with sufficiently large data sets and it is a key result in combinatorial geometry (see [4, 9] and the dozens of references there). Section 2 presents background on Tverberg-type theorems and some relevant prior work. Our new versions of Tverberg's theorem are presented in section 3. They yield three main applications we discuss next. All our proofs are based on basic combinatorial geometry and probability. They are presented in section 4.

Existence of MLE on multiclass logistic regression. As a consequence of our stochastic Tverberg theorems, Theorems 3.1, 3.2, 3.3, 3.4, and 3.5, we provide a sufficient condition for all these MLEs to exist with high probability. It is worth remarking our results apply to multi-class logistic regression, only in settings where the labels and covariates are independent, but we expect further generalizations could be possible.

Here a sequence of events $X_n, n \geq 1$, occurs with *with high probability* if $\lim_{n\to\infty} P(X_n) = 1$. We say $f(x) \gg g(x)$ if $\lim_{x\to\infty} \frac{f(x)}{g(x)} \geq 1$.

Corollary 1.1 (stochastic Tverberg theorems applied to multinomial logistic regression). *Fix $\epsilon > 0$, and let F be a centrally symmetric continuous probability distribution.*
Then we present the following:

1. *Let D_m, $m \in \mathbb{N}$, denote a sequence of labeled point sets generated by drawing $f(m)$ points from F, and then independently assigning each point a label in $[m]$ with uniform probability. Let X_m denote the event that the MLE exists between the data corresponding to every pair of labels in D_m. Then X_m occurs with high probability as long as*
$$f(m) \gg (1+\epsilon) m \log_2(m) \ln(\ln(m)).$$
2. *Fix m, $\delta > 0$, and let G_t, $t \in \mathbb{N}$, denote a sequence of labeled point sets generated by drawing $g(t)$ points from F, and then independently assigning each point a label in $[m]$ with uniform probability. Let Y_t denote the event that the MLE exists between the data corresponding to every pair of labels in G_t, even after any t points (which can be thought of as outliers) are removed. Then Y_t occurs with high probability as long as $g(t) \gg (1+\delta) m t$.*

The same bound applies to "one-versus-rest" logistic regression, since MLE existence in that case is a weaker condition.

Data separability and MLE tolerance. The second application of our results relates measuring data separability and the answer to the question: *How robust is the existence of MLEs for multiclass logistic models under removal of outliers or corrupted data points?*

From the research of [1], we know nonseparability of data classes is important for the existence of MLE. More recently, when studying binomial logistic regression, Freund, Grigas, and Mazumder [11] introduced the following notion to quantify the extent that a dataset is nonseparable (where $a^- := -\min\{a, 0\}$ denotes the negative part of a):

$$\text{DegNSEP}^* := \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} [y_i \boldsymbol{\beta}^T \boldsymbol{x}_i]^- \quad \text{s.t. } \|\boldsymbol{\beta}\| = 1.$$

DegNSEP* is thus the smallest (over all normalized vectors $\boldsymbol{\beta}$) average misclassification error of the model $\boldsymbol{\beta}$ over the n observations. Freund, Grigas, and Mazumder showed that the condition number DegNSEP* informs the computational properties and guarantees of the standard deterministic first-order steepest descent solution method for logistic regression.

Denote by $\Delta\boldsymbol{X}$ a linear perturbation of the data set $\boldsymbol{X}$ by adding, a possibly different perturbation vector, to each data point. Thus $(\boldsymbol{X}+\Delta\boldsymbol{X}, y)$ effects a translation of the points in $\boldsymbol{X}$. Define PertSEP* as the smallest (or more precisely, the infimum thereof) perturbation $\Delta\boldsymbol{X}$ of the feature data $\boldsymbol{X}$ which will render the perturbed problem instance $(\boldsymbol{X}+\Delta\boldsymbol{X}, y)$ separable. To formalize this notion, we can view the perturbation $\Delta\boldsymbol{X}$ as a linear operator, and define the operator norm notation $\|\Delta\boldsymbol{X}\|_{\cdot,q} := \max_{\beta:\|\beta\|\leq 1}\|\Delta\boldsymbol{X}\beta\|_q$ on the space $\mathbb{R}^{n\times p}$. Then we can write

$$\begin{aligned} \text{PertSEP}^* &:= \inf_{\Delta\boldsymbol{X}} \frac{1}{n}\|\Delta\|_{\cdot,1} \\ \text{s.t. } &(\boldsymbol{X}+\Delta\boldsymbol{X}, y) \text{ is separable.} \end{aligned}$$

In Proposition 2.4 of [11] it is shown that DegNSEP* = PertNSEP*.

In this paper we introduce a new parameter PertSEP*_0 simply defined as the L_0 norm of the smallest perturbation of the feature data $\boldsymbol{X}$ which will render the perturbed problem instance $(\boldsymbol{X}+\Delta\boldsymbol{X}, y)$ separable. In other words, it is the minimal number of data points we could move to make the data set separable, normalized by the total number of data points. Namely,

$$\begin{aligned} \text{PertSEP*}_0 &:= \inf_{\Delta\boldsymbol{X}} \frac{1}{n}\|\Delta\|_{\cdot,0} \\ \text{s.t. } &(\boldsymbol{X}+\Delta\boldsymbol{X}, y) \text{ is separable.} \end{aligned}$$

Now we connect our separability parameter PertSEP*_0 to the robustness of the existence of MLEs. A partition of data is *t-tolerant* when the intersection pattern of the convex hulls of the color classes or subsets remains the same, even after any t points are removed. See Figure 1.1 for an example of a 1-tolerant Tverberg partition.

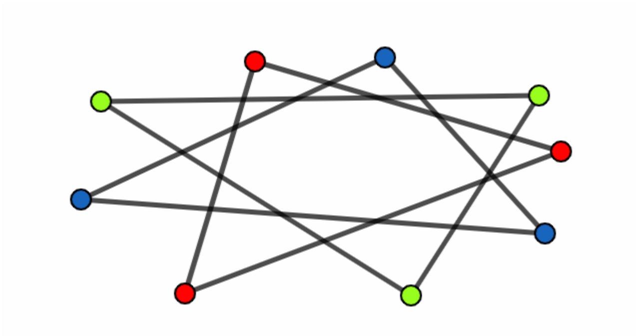

Figure 1.1. *A Tverberg partition in three data classes with tolerance one. All three convex hulls intersect, even after any point is removed.*

It is our observation that t-tolerant partitions correspond to robust MLE existence. Any t points, possibly corrupted or outlier data, can be removed and still the convex hulls of the data of each class intersect. The following theorem shows that the tolerance of a Radon partition is given by PertSEP*_0.

Theorem 1.2. *Suppose that $\boldsymbol{X} = \boldsymbol{X}_1 \cup \boldsymbol{X}_2$, $|\boldsymbol{X}| = n$ is a Radon partition with tolerance precisely equal to t. Then viewing $\boldsymbol{X}$ as a labeled dataset (with $(\boldsymbol{X}, y) = \{(\boldsymbol{x}, -1) : \boldsymbol{x} \in$*

$\boldsymbol{X}_1\} \cup \{(\boldsymbol{x}, 1,) : \boldsymbol{x} \in \boldsymbol{X}_2\})$, *we have that*

$$PertSEP^*{}_0 = t/n.$$

Theorem 1.2, combined with a result of Soberón, has a corollary, stated precisely later, which roughly says that PertSEP*$_0$ of a randomly bipartitioned point set asymptotically approaches 1/2. This is the highest possible value one could hope for since, by definition, PertSEP*$_0$ of *any* two class data set is bounded above by 1/2. In fact, this result extends easily to the multiclass setting. In other words, for a large randomly m-partitioned data set, we expect PertSEP*$_0$ of every pair of data classes to be close to 1/2—independent of both the dimension of the covariates, as well as the number of classes m. One can guarantee that, with sufficiently many samples, the t tolerance can be guaranteed; thus the MLE exists even with removal of outliers.

Centerpoints of data sets. In 1975, Tukey [24] introduced *centerpoints*, sometimes called *Tukey medians*, as a generalization of the notion of median for high-dimensional data. Centerpoints are useful in a variety of applications in nonparametric multivariate data analysis (see e.g., [9, 19, 17] and the references there for more details). To define centerpoints, we first define the *depth* for a point p, relative to a data cloud S, as the smallest number of data points in a closed half-space with boundary through p. We say a point p has *half-space depth k in S* if that every half-space containing p contains at least k points in S. A *centerpoint* of an n point data set $S \subset \mathbb{R}^d$ is a point p such that every half-space containing p has at least $\frac{n}{d+1}$ points in S, thus it is a point of depth at least $\frac{n}{d+1}$. Unfortunately, computing a centerpoint is difficult, and the current best randomized algorithm constructs a centerpoint in time $O(n^{d+1} + n \log n)$ [6, 15]. Thus, finding an approximate centerpoint of a data set is of interest.

Tverberg's theorem shows a way to find a centerpoint. This is because the intersection point of a Tverberg partition, called a *Tverberg point*, must be a point of half-space depth one in each of the $m = \lceil \frac{n}{d+1} \rceil$ color classes. Hence an effective way to compute Tverberg partitions is desirable as a method to obtain centerpoints. The proof of Radon's lemma is constructive and, in fact, one of the most notable randomized algorithms for computing approximate centerpoints works by repeatedly replacing subsets of $d + 2$ points by their Radon point. In contrast, no known polynomial time algorithm exists for computing exact Tverberg points. Thus, fast algorithms for approximate Tverberg points have been introduced in [7, 16, 18]. Our third main application is a simple algorithm for finding a Tverberg partition among a set of i.i.d. points drawn from a distribution which is balanced about a point $\boldsymbol{p}$. Here and in the rest of the paper, balanced about a point p means every hyperplane through $\boldsymbol{p}$ partitions the distribution into two sets of equal measure.

As a result one can compute approximate centerpoints. The random models for sampling points will be explained in subsection 2.1.

Corollary 1.3 (randomized generation of approximate centerpoints). *If we fix a $\delta > 0$, then sample and label n random points from the Random equipartition model, then with probability approaching one as $n \to \infty$, a random balanced labeling of the points into $(1 - \delta)(\frac{n}{\log_2(n)})$ classes gives a Tverberg partition.*

Similarly, if we sample n random points from the random allocation model, then with probability approaching one as $n \to \infty$, a random uniform labeling of the points into $(1-\delta)(\frac{n}{\log_2(n)(\ln(\ln(n)))})$ classes yields Tverberg partition.

2. Our methods: Tverberg-type theorems.

We start with formal definitions and background. We remember the famous original version of Tverberg's theorem [25] which generalizes Radon's lemma to m-partitions (see [4, 9] for references and the central importance of this theorem in convex geometry).

Definition 2.1. *Given a set $S \subset \mathbb{R}^d$, a* Tverberg m-partition of S *is a partition of S into m subsets $S_1, \dots, S_m$ with the property that all m convex hulls of the S_i intersect. In other words, we have*

$$\bigcap_{i \in [m]} \operatorname{conv}(S_i) \neq \varnothing.$$

The case $m = 2$ is called a Radon partition.

Theorem: (H. Tverberg 1966). *Every set S with at least $(d+1)(m-1)+1$ points in Euclidean d-space has at least one m-Tverberg partition (with tolerance zero).*

The notion of *tolerant Tverberg theorems* was pioneered by Larman [12]. Here is the definition.

Definition 2.2. *Given a set $S \subset \mathbb{R}^d$, a* Tverberg m-partition of S with tolerance t *is a partition of S into m subsets $S_1, \dots, S_m$ with the property that all m convex hulls of the S_i intersect after any t-points are removed. In other words, for all $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_t\} \in S$, we have*

$$\bigcap_{i \in [m]} \operatorname{conv}(S_i \setminus \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_t\}) \neq \varnothing.$$

The following result is due to Soberón and Strausz [23].

Theorem: (Soberón, Strausz 2012). *Every set S with at least $(t+1)(m-1)(d+1)+1$ points in $\mathbb{R}^d$ has at least one Tverberg m-partition with tolerance t. In other words, S can be partitioned into m parts $S_1, \dots, S_m$ so that for all $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_t\} \in S$, we have*

$$\bigcap_{i \in [m]} \operatorname{conv}(S_i \setminus \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_t\}) \neq \varnothing.$$

More recently, Soberón proved the following bound [22]. Let N denote the smallest positive integer such that a Tverberg m-partition with tolerance t exists among any N points in dimension d. Then $N = mt + O(\sqrt{t})$ for fixed m and d. The proof of this result relies on the probabilistic method and, as Soberón remarked, can, in fact, be used to prove a stochastic Tverberg-type result, which we will revisit later.

2.1. Two random models for stochastic Tverberg theorems.

Before stating our main results, we introduce two models for random data point sets. In both models we will use the term colors instead of subsets or subclass. Hereafter, when we refer to a continuous distribution on $\mathbb{R}^d$, we mean continuous with respect to the Lebesgue measure on $\mathbb{R}^d$. We defer proofs of the new results stated until the next section.

Our first model is a so-called *random equipartition model* i.e., we ensure that every color has the same number of points. More specifically, given integers m and n and a continuous probability distribution D on $\mathbb{R}^d$, we let $\mathcal{E}_{m,n,D}$ denote a random equipartitioned point set with mn points, consisting of m colors, and n points of each color, distributed independently according to D.

Our second model is a *random allocation model.* Given integers k and m and a continuous probability distribution D on $\mathbb{R}^d$, we let $\mathcal{R}_{m,k,D}$ denote a random point set with k points i.i.d. according to D, which are randomly colored with one of m colors with uniform probability (i.e., probability $1/m$ for choosing a color).

Results for the equipartition model can often be extended to the random allocation model via the following.

Observation 2.3. *The probability that a random allocation of k points into m colors is an m-Tverberg partition with tolerance t is bounded below by the probability that a random allocation of k points into m colors has at least n points per color, times the probability that an equipartition of nm points into m colors is Tverberg with tolerance t.*

Using these two probabilistic models we can state all the stochastic versions of Radon and Tverberg's theorem. To begin we can restate Cover's result as follows.

Theorem: (T. Cover 1965). *If D is a continuous probability distribution on $\mathbb{R}^d$, then*

$$\mathbb{P}(\mathcal{R}_{2,n,D} \text{ is Radon}) = 1 - 2^{-n+1} \sum_{k=0}^{d} \binom{n-1}{k}.$$

In particular, we have

$$\mathbb{P}(\mathcal{R}_{2,2(d+1),D} \text{ is Radon}) = 1/2.$$

Furthermore, for any $\epsilon > 0$ and any sequence of continuous probability distributions $\{D_i\}, i \in \mathbb{Z}_+$ where each D_d is a distribution on $\mathbb{R}^d$, we have

$$\lim_{i\to\infty} \mathbb{P}(\mathcal{R}_{2,(1+\epsilon)2i,D_i} \text{ is Radon}) = 1$$

and

$$\lim_{i\to\infty} \mathbb{P}(\mathcal{R}_{2,(1-\epsilon)2i,D_i} \text{ is Radon}) = 0.$$

To the best of the authors' knowledge, the first generalization of Cover's 1964 result to more than two colors appeared only recently in Soberón's paper [22].

Theorem: P. Soberón 2018. *Let N, t, d, m be positive integers, and let $\epsilon > 0$ be a real number. Given N points in $\mathbb{R}^d$, a random allocation of them into m parts is a Tverberg partition with tolerance t with probability at least $1 - \epsilon$, as long as*

$$t + 1 \le N/m - \sqrt{\frac{1}{2}\left[(d+1)(m-1)N\ln(Nm) + N\ln\left(\frac{1}{\epsilon}\right)\right]}.$$

This result is quite remarkable. For any fixed m, d, and δ, it shows that the probability of a random allocation of N points in $\mathbb{R}^d$ in m colors having tolerance at least $(1-\delta)N/m$ approaches one as N goes to infinity. On the other hand, by the pigeonhole principle, any partition of N points into m colors must have one color with at most N/m points. Thus, for a fixed number of colors m, the tolerance of a random partition is asymptotically as high as it could possibly be! By Theorem 1.2, this result yields the following corollary.

Corollary 2.4. *For any sequence* $\{\mathcal{R}_{(2,k,D)}\}, k \in \mathbb{N}$ *of partitioned point sets with* D *a distribution on* $\mathbb{R}^d$*, and any* $\epsilon > 0$*, we have* $|PertSEP^*{}_0(\mathcal{R}_{(2,k,d)}) - 1/2| < \epsilon$ *with high probability.*

3. Our new stochastic Tverberg theorems.

Now we present all our geometric results.

First, we extend Corollary 2.4 to the multiclass setting. More generally, for fixed d and m, for any large randomly m-partitioned data set, we expect PertSEP$^*{}_0$ of every pair of data points to be close to $1/2$.

Theorem 3.1. *Fix* $\epsilon > 0$*. For any distribution* D *on* $\mathbb{R}^d$ *and any sequence* $\{\mathcal{R}_{(m,k,D)}\}, k \in \mathbb{N}$ *of* m*-partitioned point sets* $\mathcal{R}_{(m,k,D)} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m\}$*, we have*

$$\lim_{k\to\infty}\left(\min_{\boldsymbol{X}_i,\boldsymbol{X}_j\in\mathcal{R}_{(m,k,D)}} PertSEP^*{}_0(\boldsymbol{X}_i \cup \boldsymbol{X}_j) = 1/2\right)$$

with high probability.

Our second theorem is a stochastic Tverberg result similar to Soberón's and Cover's but for equipartitions (without considering tolerance).

Theorem 3.2 (stochastic Tverberg theorem for equipartitions). *Suppose* D *is a probability distribution on* $\mathbb{R}^d$ *that is balanced about some point* $\boldsymbol{p} \in \mathbb{R}^d$*, in the sense that every hyperplane through* $\boldsymbol{p}$ *partitions* D *into two sets of equal measure. Then*

$$\left(1-\left(\frac{1}{2^{n-1}}\sum_{k=0}^{d-1}\binom{n-1}{k}\right)\right)^m \le \mathbb{P}(\mathcal{E}_{m,n,D} \text{ is Tverberg }) \le \left(2(1-2^{-n})^m - (1-2^{-n+1})^m\right)^d.$$

In fact, the previous theorem is asymptotically tight in the number of colors m. This is shown by our next theorem, which establishes an interesting threshold phenomenon for Tverberg partitions.

Theorem 3.3 (Tverberg threshold phenomena for equipartitions). *Let* D *be a continuous probability distribution in* $\mathbb{R}^d$ *balanced about some point* $\boldsymbol{p} \in \mathbb{R}^d$*. Consider the sequence of random equipartitioned point sets* $\mathcal{E}_{m,f(m),D}$*, where* $m \in \mathbb{N}$*, and* $n = f(m)$ *depends on* m*. Then* $\mathcal{E}_{m,f(m),D}$ *is Tverberg with high probability if* $f(m) \gg \log_2(m)$*, and* $\mathcal{E}_{m,f(m),D}$ *is not Tverberg with high probability if* $f(m) \ll \log_2(m)$*.*

We note that the number of points needed to reach the conclusion in Theorem 3.3 is independent of the dimension, as in the aforementioned result of Soberón [22].

The next two theorems adapt both Cover's result and Theorem 3.2 to the setting of tolerance.

Theorem 3.4 (stochastic Tverberg with tolerance for equipartition). *Suppose D is a probability distribution on $\mathbb{R}^d$ that is balanced about some point $\boldsymbol{p} \in \mathbb{R}^d$, then*

$$\mathbb{P}(\mathcal{E}_{m,n,D} \textit{ is Tverberg with tolerance } t) \geq \left(1 - 2^{-\lfloor n/2d \rfloor} \sum_{i=1}^{t} \binom{\lfloor n/2d \rfloor}{i}\right)^m .$$

For the case of random bipartitions, we can adapt Cover's result to obtain a stochastic Radon theorem with tolerance.

Theorem 3.5 (stochastic Radon with tolerance for random allocation). *If D is a continuous probability distribution on $\mathbb{R}^d$, then*

$$\mathbb{P}(\mathcal{R}_{2,k,D} \textit{ is Radon with tolerance } t) \geq 1 - \left(2^{-\lfloor k/(2d+2) \rfloor} \sum_{i=0}^{t} \binom{\lfloor k/(2d+2) \rfloor}{i}\right) .$$

In particular, we have

$$\mathbb{P}(\mathcal{R}_{2,k,D} \textit{ is Radon with tolerance } \lfloor k/(4d+4) \rfloor) \geq 1/2.$$

For random allocations with more than two colors, we will use some developments on random allocation problems, including the following notation. If balls are thrown into m urns uniformly and independently, let $N_m(n)$ equal the number of throws necessary to obtain at least m balls in each urn.

Corollary 3.6 (stochastic Tverberg for random allocation). *Suppose D is a probability distribution on $\mathbb{R}^d$ that is balanced about some point $\boldsymbol{p} \in \mathbb{R}^d$.*

1. *Then*

$$\mathbb{P}(\mathcal{R}_{m,k,D} \textit{ is Tverberg with tolerance } t) \geq \mathbb{P}(N_n(m) \leq k) \left(1 - 2^{-\lfloor n/2d \rfloor} \sum_{i=1}^{k} \binom{\lfloor n/2d \rfloor}{i}\right)^m .$$

2. *For the case of Tverberg without tolerance, we also have*

$$\mathbb{P}(\mathcal{R}_{m,k,D} \textit{ is Tverberg}) \geq \mathbb{P}(N_n(m) \leq k) \left(1 - \left(2^{-n+1} \sum_{k=0}^{d-1} \binom{n-1}{k}\right)\right)^m .$$

3. *Suppose $\mathcal{R}_{m,f(m),D}$, $m \in \mathbb{N}$ is a sequence of random partitioned point sets, where $n = f(m)$ depends on m.*
   Then $\mathcal{R}_{m,f(m),D}$ is Tverberg with high probability if $f(m) \gg m \log_2(m) \ln(\ln(m))$.

4. Proofs of our results.

Proof of Theorem 1.2. Let M denote the minimal number of points perturbed among any perturbation that makes $(\boldsymbol{X}, y)$ separable, and let N denote the minimal number of points needing to be removed from $(\boldsymbol{X}, y)$ to make $(\boldsymbol{X}, y)$ separable. Then $\text{PertSEP*}_0(\boldsymbol{X}, y)$ is equal to M/n, and the tolerance t of $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ is equal to N. It suffices to show that $M = N$. To see that $M \geq N$, note if $\boldsymbol{x}_1, \ldots \boldsymbol{x}_M$ in $\boldsymbol{X}$ are moved so that the resulting set $(\boldsymbol{X}', y')$ is separable, then $(\boldsymbol{X} \setminus \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}, y \setminus \{y_1, \dots y_N\})$ is also separable. To see that $M \le N$, suppose that $(\boldsymbol{X} \setminus \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_M\}, y \setminus \{y_1, \dots y_M\})$ is separable by a hyperplane h. Then moving $\boldsymbol{x}_1, \dots, \boldsymbol{x}_M$ to the appropriate sides of the hyperplane determined by h, we can construct a separable dataset $(\boldsymbol{X}', \boldsymbol{y}')$, obtained from moving M points from $(\boldsymbol{X}, \boldsymbol{y})$. ■

Proof of Theorem 3.1. For fixed m, d, and δ, let $T(N)$ denote the event that a random allocation of N points in $\mathbb{R}^d$ in m colors has tolerance at least $(1-\delta)N/m$. By Soberón's theorem above, $\mathbb{P}(T(N))$ asymptotically approaches one as N goes to infinity. In particular, for fixed $\zeta > 0$, we can pick $\tilde{N}$ such that $\mathbb{P}(T(\tilde{N})) > 1 - \zeta$. Now, for fixed m and ϵ, let $E_i(N)$ denote the event that a random allocation of N points into m colors has between $(1-\epsilon)N/m$ and $(1+\epsilon)N/m$ points of color i, where $i \in [m]$. By the law of large numbers, $\mathbb{P}(E_i(N))$ approaches one as N goes to infinity. As the $m+1$ events $E_j(N)$, where $j = \{0, 1, \dots, m\}$, all have probability approaching one, the probability of the intersection of all these events also approaches one. This can be seen by applying the union bound to their complements. Thus, for any fixed $\epsilon_2 > 0$, there exists $N' \in \mathbb{N}$ such that the $E_j(N')$, where $j = \{0, 1, \dots, m\}$, simultaneously occur with probability $(1 - \epsilon_2)$. Thus, if we pick an $N \ge \max(\tilde{N}, N')$, by the union bound, we have with probability $1 - \zeta - \epsilon_2$, that $T(N)$ and $E_j(N')$, where $j = \{0, 1, \dots, m\}$, simultaneously occur.

Therefore, with probability $1 - \zeta - \epsilon_2$, each pair of colors has at most $(1+\epsilon)2N'/m$ points and is a Radon partition of tolerance at least $(1-\delta)N'/m$ (the tolerance of each bipartition is a priori bounded below by the tolerance of the m-partition). By Theorem 1.2, PertSEP$^*_0$ of each pair is at least $(1-\delta)/2(1+\epsilon)$ with probability $(1-\epsilon_2)$. Since δ, ζ, ϵ, and ϵ_2 were arbitrary, this completes the proof.

Proof of the lower bound in Theorem 3.2. After a possible translation, we can assume without loss of generality that D is balanced about the origin. We will prove that

$$\left(1 - \left(2^{-n+1} \sum_{k=0}^{d-1} \binom{n-1}{k}\right)\right)^m \le \mathbb{P}(\mathcal{E}_{m,n,D} \text{ is Tverberg})$$

by bounding from below the probability that the origin is a Tverberg point. We may assume without loss of generality that none of the randomly selected points are the origin. Furthermore, we can radially project the points onto a sphere of radius smaller than the minimal norm of the projected points, since that will not affect whether the origin is a Tverberg point. After this projection, we may assume the points are uniformly sampled on a small sphere centered at the origin. The origin is then a Tverberg point as long as the points from each color contain the origin in their convex hull. This is equivalent to showing no color has all of its points contained in one hemisphere. For a fixed color, the probability of the n points of that color being contained in one hemisphere was computed by Wagner and Welzl [26] (generalizing the celebrated result of Wendel [27] addressing the case when D is rotationally invariant about the origin) as

$$\left(2^{-n+1} \sum_{k=0}^{d-1} \binom{n-1}{k}\right). \tag{4.1}$$

■

Using this to compute the probability that none of the m color classes is contained in one hemisphere, we obtain the desired bound above.

Proof of the upper bound in Theorem 3.2. Again, we assume without loss of generality that D is balanced about the origin. We will first treat the case $d = 1$, and then explain how to obtain the bound for arbitrary d. To bound the probability of a Tverberg partition from above, we bound the probability of the complement below. We let E denote the event that the convex hulls have empty intersection. In dimension one, E is contained in the event that there is at least one color class with all points less than zero, and at least one color class with all points greater than zero. Since we assume that the origin equipartitions D, we can rephrase this as the probability that among m people each flipping n fair coins, there is at least one person with all heads and at least one person with all tails, that is, denoting by H and T the events that at least one person gets all heads or tails respectively, we have $\mathbb{P}(E) \geq \mathbb{P}(H \cap T)$. We have

$$\mathbb{P}(H \cap T) = \mathbb{P}(H) + \mathbb{P}(T) - \mathbb{P}(H \cup T) = \mathbb{P}(H) + \mathbb{P}(T) - (1 - \mathbb{P}(H \cup T)^c).$$

Since $\mathbb{P}(H) = \mathbb{P}(T) = (1 - 2^{-n})^m$ and $\mathbb{P}((H \cup T)^c) = 1 - (1 - 2^{-n+1})^m$, this yields

$$\mathbb{P}(H \cap T) = 1 + (1 - 2^{-n+1})^m - 2(1 - 2^{-n})^m.$$

The probability of a Tverberg partition is thus bounded as follows:

$$\mathbb{P}(\mathcal{E}_{m,n,D} \text{ is Tverberg}) \leq 1 - \mathbb{P}(E) \leq 1 - \mathbb{P}(H \cap T) = 2(1 - 2^{-n})^m - (1 - 2^{-n+1})^m.$$

This proves the desired bound for dimension one. For higher dimensions, we note that if we let p_i, denote the projection onto the ith axis for $i \leq d$, we have that the signs of $p_1(x), \ldots, p_d(x)$ are independent Bernoulli random variables with probability $1/2$ (as the hyperplane orthogonal to the ith axis equipartitions D by the assumption that D is balanced about the origin). Thus to have a Tverberg partition, we must have that no pair of the color classes are separated by the origin after projecting onto the d coordinate axes. Since these d events are independent, the probability of this happening is bounded as follows:

$$\mathbb{P}(\mathcal{E}_{m,n,D} \text{ is Tverberg}) \leq \left(2(1 - 2^{-n})^m - (1 - 2^{-n+1})^m\right)^d.$$

Proof of Theorem 3.3. We will show that $\mathcal{E}_{m,f(m),D}$ is Tverberg with high probability if $f(m) > \log_2(m)$. Fix an $\epsilon > 0$. We set $n = (1 + \epsilon)\log_2(m)$ and apply the lower bound in Theorem 3.2 to deduce that

$$\mathbb{P}(\mathcal{E}_{m,n,D} \text{ is Tverberg }) \geq \left(1 - \left(2^{-(1+\epsilon)*\log_2(m)+1} \sum_{k=0}^{d-1} \binom{n-1}{k}\right)\right)^m$$
$$= \left(1 - \left(2m^{-(1+\epsilon)} \sum_{k=0}^{d-1} \binom{n-1}{k}\right)\right)^m.$$

Choosing a constant K so that $Kn^d \geq 2\sum_{k=0}^{d-1}\binom{n-1}{k}$, we have

$$(1 - Kn^d m^{-(1+\epsilon)})^m \leq \mathbb{P}(\mathcal{E}_{m,n,D} \text{ is Tverberg}).$$

We will show that the limit as m approaches infinity of the left-hand side is bigger than $e^{-\delta}$ for any $\delta > 0$. Fix $\delta > 0$. As $n^d \sim O(\ln(m)^d)$, there exists an M such that $Kn^d m^{-\epsilon} < \delta$ for all $m \geq M$. Consequently, $(1 - Kn^d m^{-(1+\epsilon)})^m > (1 - \delta m^{-1})^m$ for all $m \geq M$. Thus

$$\lim_{m\to\infty} (1 - Kn^d m^{-(1+\epsilon)})^m \geq \lim_{m\to\infty} (1 - \delta m^{-1})^m = e^{-\delta}.$$

Since δ was arbitrary, we see that the probability of a Tverberg partition tends to one.

Now we show that $\mathcal{E}_{m,f(m),D}$ is not Tverberg with high probability if $f(m) < \log_2(m)$. As before, we fix an ϵ greater than zero apply the upper bound in Theorem 3.2 with $n = (1-\epsilon)log_2(m)$ to obtain

$$\mathbb{P}(\mathcal{E}_{m,n,D} \text{ is Tverberg}) \leq \left(2(1 - m^{-1+\epsilon})^m - (1 - 2m^{-1+\epsilon})^m\right)^d.$$

For any $\gamma > 0$, when m is large, both terms inside the parentheses are smaller than $(1 - \gamma m^{-1})^m$. Since $\lim_{m\to\infty}(1-\gamma m^{-1})^m = e^{-\gamma}$, the probability of a Tverberg partition converges to zero as m approaches infinity.

Proof of Theorem 3.4. Again, we assume without loss of generality that D is balanced about 0. Let S denote the set of points of some fixed color. Then we assume that $|S| = n$, and we can partition S into $\lfloor n/2d \rfloor$ subsets $S_1, \ldots, S_{\lfloor n/2d \rfloor}$ with $S_i \geq 2d$ for each i. By Wagner and Welzl's result (see (4.1) above), for each i, $\text{conv}(S_i)$ contains the origin with probability at least 1/2. By independence, the probability that less than $t+1$ of the S_i contain the origin is less than $2^{-\lfloor n/2d \rfloor} \sum_{i=1}^{t} \binom{\lfloor n/2d \rfloor}{i}$. On the other hand, if at least $t+1$ of the $\text{conv}(S_i)$ contain the origin, then by pigeonhole principle $\text{conv}(S \setminus \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t\})$ contains the origin for any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t \in S$. Thus, with probability at least $1 - 2^{-\lfloor n/2d \rfloor} \sum_{i=1}^{t} \binom{\lfloor n/2d \rfloor}{i}$, we have that $\text{conv}(S) \setminus \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t\})$ contains the origin. Since this probability is independent for each of the m colors, the result follows. ■

Using a similar strategy combined with Cover's result, we give the proof of Theorem 3.5 below.

Proof of Theorem 3.5. Given k points in $\mathbb{R}^d$ colored red and blue by random allocation, we arbitrarily partition them into $\lfloor k/(2d+2) \rfloor$ groups of size at least $2d+2$. By Cover's result, for each fixed group, the convex hulls (of the red and blue points) in that group intersect with probability at least 1/2. For each of the $\lfloor k/(2d+2) \rfloor$ groups, we think of the event that the convex hulls in that group intersect as a "success." Then the probability that at least $t+1$ groups have intersecting convex hulls is bounded below by the probability that a binomial process with $\lfloor k/(2d+2) \rfloor$ trials and success probability 1/2 has at least $t+1$ total successes. Computing this binomial probability yields the theorem. (If at least $t+1$ groups have intersecting convex hulls, then removing at most t points leaves at least one group with intersecting convex hulls.) ■

Proof of Corollary 3.6. We split the proof according to the three respective parts of the statement.

1. This follows from Observation 2.3 combined with Theorem 3.4.
2. The result for the special case of Tverberg without tolerance then follows the same reasoning as part (1), except using Theorem 3.2 in place of Theorem 3.4.
3. To show the asymptotic result, we use a result on urn models due to Erdős and Renyi [10] saying that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{N_m(n)}{n} < \log(n) + (m-1)\log(\log(n)) + x\right) = exp\left(-\frac{e^{-x}}{(m-1)!}\right).$$

This implies that for any $\epsilon > 0$ and sequence of $\log(\log(m))log_2(m)(1+\epsilon)$ points allocated into m urns, we have at least $log_2(m)(1+\epsilon/2)$ points in each urn with high probability. Then we apply Theorem 3.3, which says that any equipartition of a point set into m colors and $log_2(m)(1+\epsilon/2)$ points per color is Tverberg with high probability. ■

Proof of Corollary 1.1. For part one, note that a direct application of Corollary 3.6, part 3 implies that the D_m are Tverberg m-partitions with high probability. Since the MLE exists for any two classes in every Tverberg partition, we therefore have that the MLE exists between any two classes in D_m with high probability.

For part two, consider F and $G_t = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m\}$ as hypothesized, and apply Theorem 3.1 with $F = D$, and $G_t = \{\mathcal{R}_{(m,k,D)}\}, k \in \mathbb{N}$. Then we have that

$$\lim_{k\to\infty}\left(\min_{\boldsymbol{X}_i,\boldsymbol{X}_j\in G_t} \text{PertSEP*}_0(\boldsymbol{X}_i \cup \boldsymbol{X}_j) = 1/2\right)$$

with high probability. In particular, for fixed δ and ϵ, we can find N such that

$$\min_{\boldsymbol{X}_i,\boldsymbol{X}_j\in G_N} \text{PertSEP*}_0(\boldsymbol{X}_i \cup \boldsymbol{X}_j) = (1-\delta)/2 \tag{4.2}$$

holds with probability $1-\epsilon$. Then, applying Theorem 1.2 to (4.2), for each pair $\boldsymbol{X}_i, \boldsymbol{X}_j$, we have a Radon partition with tolerance $(|\boldsymbol{X}_i| + |\boldsymbol{X}_j|)(1-\delta)/2$ with probability $1-\epsilon$. This implies that $\min_{i\in[m]}(|\boldsymbol{X}_i|) \geq (\min_{i\in[m]}(|\boldsymbol{X}_i|) + \max_{i\in[m]}(|\boldsymbol{X}_i|))/(1-\delta)/2$, and thus

$$\min_{i\in[m]}(|\boldsymbol{X}_i|) \geq \max_{i\in[m]}(|\boldsymbol{X}_i|)(1-\delta)/(1+\delta) \geq \frac{|G_t|(1-\delta)}{m(1+\delta)}.$$

In light of this and (4.2), for each pair $\boldsymbol{X}_i, \boldsymbol{X}_j$, we actually have a Radon partition with tolerance $\frac{|G_t|(1-\delta)}{m(1+\delta)^2}$ with probability $1-\epsilon$. As ϵ and δ were arbitrary, the theorem follows. ■

Proof of Corollary 1.3. According to our Theorem 3.3, a random equipartition of n such points into $(1-\delta)n/\log_2(n)$ sets should produce a Tverberg partition with probability approaching one as $n \to \infty$. Similarly, for the random allocation model, we can use Corollary 3.6, part 3, which shows that a random partition of n such points into $(1-\delta)n/(\log_2(n)\ln(ln(n)))$ sets should produce a Tverberg partition with probability approaching one as $n \to \infty$. ■

5. Conclusions. This paper presented applications of Tverberg-type results to data science. We end with a summary and a few natural open questions.

The majority of our applications were to the existence of the maximum-likelihood estimator. Table 5.1 summarizes our geometric theorems (middle column) as well as their corresponding consequences to the existence of the maximum-likelihood estimator in terms of the size of the data set (right column). We included the tolerant existence of MLE.

Table 5.1

Stochastic analogues of Tverberg's theorem and their implications for existence of MLEs. By "Likely MLE Existence," we mean that one can bound below the probability of MLE existence as a function of the number of input data points, according to the corresponding theorems in the "Stochastic" column.

Deterministic version	Stochastic version	Likely MLE existence
Radon	Cover's theorem [8]	pair of data classes (mentioned above)
Tverberg	Thm 3.2, 3.3	all pairs of data classes (Corollary 1.1 part 1)
Radon with tolerance	Thm 3.5	pair of data classes with outliers removed (Corollary 1.1 part 2)
Tverberg with tolerance	Thm 3.6, 3.4, [22]	all pairs of classes with outliers removed (Corollary 1.1 part 2)

An important question remains open: *How do these results generalize when the class labels are not independent of the features?* Our results hold only in the cases of independence. It would be desirable to obtain stochastic Tverberg theorems for collections of points that are not i.i.d., similar to the generalization of [5] in the case of two colors. We were unable to complete the technical details for such a result.

As we pointed out, PertSEP* is related to computational performance. We do not know of a connection of PertSEP*0 to computational properties related to logistic regression, so this is another interesting open question.

Finally, we summarize the performance and time complexity of various algorithms for obtaining Tverberg partitions, including our own (last two rows), in Table 5.2.

Table 5.2

Approximate Tverberg partitions for balanced distributions, where n is the total number of points. Below is our prior work and results.

Method	Number of colors	Time complexity
Tverberg [25]	$\lfloor (n+1)/(d+1) \rfloor$	PPAD (unknown if polynomial)
Mulzer, Werner [16]	$n/(4d+1)^3$	$d^{O(\log d)}n$
Rolnick, Soberón [18]	$n/d(d+1)^2$ with error prob. ϵ	weakly poly. in n,d and $\log(1/\epsilon)$
Random equipartition	$O(\frac{n}{\log_2(n)})$	$O(n)$
Random allocation	$O(\frac{n}{\log_2(n)(\ln(\ln(n)))})$	$O(n)$

We gave the first probabilistic construction for Tverberg partitions (without tolerance) that is asymptotically optimal in the number of colors. For the case of Tverberg with tolerance, Soberón gave a probabilistic bound which is asymptotically optimal for obtaining large tolerance in a random partition. Corollary 3.6 part 3 presents improvements on Soberón's bound when the number of colors is large relative to the desired tolerance. In fact, manipulating Soberón's result we could have obtained a corollary of the format, but ours is stronger.

On the other hand, our Theorem 3.5 yields a weaker expected tolerance than Soberóns result, but our proof is shorter and more elementary. A natural open question is whether one could obtain a bound that encapsulates all of these results as special cases.

Acknowledgments. We are grateful to Paul Grigas, Frédéric Meunier, Nabil Mustafa, David Rolnick, and Pablo Soberón for their comments. The recommendations and corrections by two anonymous referees and the journal editors greatly improved the presentation of this paper and we thank them for their time and dedication.

REFERENCES

[1] A. Albert and J. A. Anderson, *On the existence of maximum likelihood estimates in logistic regression models*, Biometrika, 71 (1984), pp. 1–10, https://doi.org/10.1093/biomet/71.1.1.

[2] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, *Living on the edge: phase transitions in convex programs with random data*, Inf. Inference, 3 (2014), pp. 224–294, https://doi.org/10.1093/imaiai/iau005.

[3] F. Baccelli and E. O'Reilly, *The stochastic geometry of unconstrained one-bit data compression*, Electron. J. Probab., 24 (2019), 138, https://doi.org/10.1214/19-EJP389.

[4] I. Bárány and P. Soberón, *Tverberg's theorem is* 50 *years old: A survey*, Bull. Amer. Math. Soc., 55 (2018), pp. 459–492, https://doi.org/https://doi.org/10.1090/bull/1634.

[5] E. J. Candès and P. Sur, *The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression*, Ann. Statist., 48 (2020), pp. 27–42, https://doi.org/10.1214/18-AOS1789.

[6] T. M. Chan, *An optimal randomized algorithm for maximum Tukey depth*, in Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms, ACM, New York, 2004, pp. 430–436.

[7] K. L. Clarkson, D. Eppstein, G. Miller, C. Sturtivant, and S. Teng, *Approximating center points with iterative Radon points*, Internat. J. Comput. Geom. Appl., 6 (1996), pp. 357–377.

[8] T. M. Cover, *Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition*, IEEE Transactions on Electronic Computers, EC-14 (1965), pp. 326–334, https://doi.org/10.1109/PGEC.1965.264137.

[9] J. A. De Loera, X. Goaoc, F. Meunier, and N. H. Mustafa, *The discrete yet ubiquitous theorems of Carathéodory, Helly, Sperner, Tucker, and Tverberg*, Bull. Amer. Math. Soc., 56 (2019), pp. 415–511, https://doi.org/10.1090/bull/1653.

[10] P. Erdős and A. Rényi, *On a classical problem of probability theory*, Magyar Tud. Akad. Mat. Kutató Int. Közl., 6 (1961), pp. 215–220.

[11] R. M. Freund, P. Grigas, and R. Mazumder, *Condition Number Analysis of Logistic Regression, and its Implications for Standard First-Order Solution Methods*, preprint, https://arxiv.org/abs/1810.08727, 2018.

[12] D. G. Larman, *On sets projectively equivalent to the vertices of a convex polytope*, Bull. London Math. Soc., 4 (1972), pp. 6–12.

[13] O. Mangoubi and A. Edelman, *Integral Geometry for Markov Chain Monte Carlo: Overcoming the Curse of Search-subspace Dimensionality*, preprint, https://arxiv.org/abs/1503.03626, 2015.

[14] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd ed., Monographs on Statistics and Applied Probability, Chapman & Hall, London, 1989.

[15] G. L. Miller and D. R. Sheehy, *Approximate centerpoints with proofs*, Comput. Geom., 43 (2010), pp. 647–654, https://doi.org/10.1016/j.comgeo.2010.04.006.

[16] W. Mulzer and D. Werner, *Approximating Tverberg points in linear time for any fixed dimension*, Discrete Comput. Geom., 50 (2013), pp. 520–535.

[17] E. Rafalin and D. L. Souvaine, *Computational geometry and statistical depth measures*, in Theory and Applications of Recent Robust Methods, M. Hubert, G. Pison, A. Struyf, and S. Van Aelst, eds., Birkhäuser Basel, Basel, 2004, pp. 283–295.

[18] D. Rolnick and P. Soberón, *Algorithmic Aspects of Tverberg's Theorem*, preprint, https://arxiv.org/

abs/1601.03083, 2016.

[19] P. J. Rousseeuw and A. Struyf, *Computing location depth and regression depth in higher dimensions*, Statistics and Computing, 8 (1998), pp. 193–203, https://doi.org/10.1023/A:1008945009397.

[20] L. Schläfli, *Gesammelte mathematische Abhandlungen. Band* I, Birkhäuser Verlag, Basel, 1950.

[21] M. J. Silvapulle, *On the existence of maximum likelihood estimators for the binomial response models*, J. Roy. Statist. Soc. Ser. B, 43 (1981), pp. 310–313, http://www.jstor.org/stable/2984941.

[22] P. Soberón, *Robust Tverberg and colourful Carathéodory results via random choice*, Combin. Probab. Comput., 27 (2018), pp. 427–440, https://doi.org/10.1017/S0963548317000591.

[23] P. Soberón and R. Strausz, *A generalisation of Tverberg's theorem*, Discrete Comput. Geom., 47 (2012), pp. 455–460, https://doi.org/10.1007/s00454-011-9379-z.

[24] J. W. Tukey, *Mathematics and the picturing of data*, in Proceedings of the International Congress of Mathematicians, Vancouver, 1975, 2 (1975), pp. 523–531, https://ci.nii.ac.jp/naid/10029477185/en/.

[25] H. Tverberg, *A generalization of Radon's theorem*, J. London Math. Soc., 41 (1966), pp. 123–128.

[26] U. Wagner and E. Welzl, *A continuous analogue of the upper bound theorem*, in Proceedings of the Sixteenth Annual Symposium on Computational Geometry (Hong Kong, 2000), ACM, New York, 2000, pp. 50–56, https://doi.org/10.1145/336154.336176.

[27] J. G. Wendel, *A problem in geometric probability*, Math. Scand., 11 (1962), pp. 109–112, https://doi.org/10.7146/math.scand.a-10655.