PAC Learning with Generalized Samples and an Application to Stochastic Geometry

Sanjeev R. Kulkarni, *Member, IEEE,* Sanjoy K. Mitter, *Fellow, IEEE,* John N. Tsitsiklis, *Member, IEEE,* and Ofer Zeitouni, *Senior Member, IEEE,*

***Abstract*— In this paper, we introduce an extension of the standard probably approximately correct (PAC) learning model, which allows the use of generalized samples. We view a generalized sample as a pair consisting of a functional on the concept class together with the value obtained by the functional operating on the unknown concept. It appears that this model can be applied to a number of problems in signal processing and geometric reconstruction to provide sample size bounds under a PAC criterion. We consider a specific application of the generalized model to a problem of curve reconstruction and discuss some connections with a result from stochastic geometry.**

***Index Terms*— Curves, generalized samples, learning, PAC model, stochastic geometry.**

I. Introduction

THE PROBABLY approximately correct (PAC) learning model is a precise framework attempting to formalize the notion of learning from examples. The earliest work on PAC-like models was done by Vapnik [23], and many fundamental results relevant to the PAC model have been obtained in the probability and statistics literature [21], [22], [6], [13]. Valiant [20] independently proposed a similar model that has resulted in a great deal of work on the PAC model in the machine learning community. More recently, Haussler [7] has formulated a very general framework refining and consolidating much of the previous work on the PAC model.

In the usual PAC model, the information received by the learner consists of random samples of some unknown function. Here, we introduce an extension in which the learner may receive information from much more general types of samples, which we refer to as generalized samples. A generalized sample is essentially a functional assigning a real number to each concept, where the number assigned may not necessarily be the value of the unknown concept at a point but could be some other attribute of the unknown concept (e.g., the integral over a region or the derivative at a given point, etc.). The model is defined for the general case in which the concepts are real valued functions and is applicable to both distribution-free and fixed-distribution learnability. The idea is simply to transform learning with generalized samples to a problem of learning with standard samples over a new instance space and concept class. The PAC learning criteria over the original space is induced by the corresponding standard PAC criteria over the transformed space. Thus, the criteria for learnability and sample size bounds are the usual ones involving metric entropy and a generalization of VC dimension for functions (in the fixed distribution and distribution-free cases, respectively).

We consider a particular example of learning from generalized samples that is related to a classical result from stochastic geometry, namely, we take X to be the unit square in the plane and consider concept classes that are collections of curves contained in X. For example, one simple concept class of interest is the set of straight-line segments contained in X. A much more general concept class we consider is the set of curves in X with bounded length and bounded turn (total absolute curvature). The samples observed by the learner consist of randomly drawn straight lines labeled as to the number of intersections the random line makes with the target concept (i.e., the unknown curve). We consider learnability with respect to a fixed distribution, where the distribution is the uniform distribution on the set of lines intersecting X. A learnability result is obtained by providing metric entropy bounds for the class of curves under consideration.

The example of learning a curve is closely related to a result from stochastic geometry that states that the expected number of intersections a random line makes with an arbitrary rectifiable curve is proportional to the length of the curve. This result suggests that the length of a curve can be estimated (or ''learned'') from a set of generalized samples. In fact, this idea has been studied, although primarily from the point of view of deterministic sampling [19], [12]. The learnability result makes the much stronger statement that for certain classes of curves, from just knowing the number of intersections with a set of random lines, the curve itself can be learned (from which the length can then be estimated). In addition, for these classes of curves, the learning result guarantees uniform convergence of empirical estimates of length to true length, which does not follow directly from the stochastic geometry result.

Manuscript received August 5, 1991; revised June 29, 1992. This work was supported by the U.S. Army Research Office under Contract DAAL03-86-K-0171, by the Department of the Navy under Air Force Contract F19628-90-C-0002, by the National Science Foundation under contract ECS-8552419, and by the Air Force under contract AFOSR-91-0368. Recommended for acceptance by Associate Editor T. Dean.

S.R. Kulkarni was with the Laboratory for Information and Decision Systems and Lincoln Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139. He is now with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544.

S. K. Mitter and J. N. Tsitsiklis are with the Department of Electrical Engineering and Computer Science and the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139.

O. Zeitouni is with the Department of Electrical Engineering, Technion, Haifa, Israel.

IEEE Log Number 9211195.

0162–8828/93$03.00 © 1993 IEEE

Finally, we discuss a number of open problems and directions for further work. We believe that the framework presented here can be applied to a number of problems in signal/image processing, geometric reconstruction, and stereology to provide sample size bounds under a PAC criterion. Some specific problems that may be approachable with these ideas include tomographic reconstruction using random ray or projection sampling and convex set reconstruction from support line or other types of measurements. For such problems, we are interested in estimating some unknown function from data that are not traditional samples of the function. However, questions concerning the amount of data required for a reconstruction of a given quality are still of great interest. The framework presented may provide one approach to addressing such questions.

II. PAC Learning with Generalized Samples

In the original PAC learning model [20], a *concept* is a subset of some *instance space* X, and a *concept class* C is a collection of concepts. The learner knows C and tries to learn a target concept c belonging to C. The information received by the learner consists of points of X (drawn randomly) and labeled as to whether or not they belong to the target concept. The goal of the learner is to produce with high probability (greater than $1-\delta$) a hypothesis that is close (within ϵ) to the target concept (hence, the name PAC for "probably approximately correct"). It is assumed that the distribution is unknown to the learner, and the number of samples needed to learn for fixed ϵ and δ is independent of the unknown concept as well as the unknown distribution (hence, the term "distribution free"). For precise definitions, see e.g., [20] and [5].

Some variations/extensions of the original model that have been studied and are relevant to the present work include learning with respect to a fixed distribution [4], [7] and learning functions as opposed to sets (i.e., binary valued functions) [7]. As the name suggests, learning with respect to a fixed distribution refers to the case in which the distribution with which the samples are being drawn is fixed and known to the learner. A very general framework was formulated by Haussler [7] building on some fundamental work by Vapnik and Chervonenkis [21]–[23], Dudley [6], and Pollard [13]. In this framework, the concept class (hypotheses), which is denoted by F, is a collection of functions from a domain X to a range Y. The samples are drawn according to a distribution on $X \times Y$ from some class of distributions. A loss function is defined on $Y \times Y$, and the goal of the learner is to produce a hypothesis from F that is close to the optimal one in the sense of minimizing the expected loss between the hypothesis and the random samples. The work of Wahba [24] is also related to the model discussed below. In [24], observations of the form $y_i = L_i f + e_i$ are considered where the L_i are bounded linear functionals and e_i is noise. Hence, these observations consist of a particular form of generalized samples in which the functionals are chosen to be linear and bounded.

Learning from generalized samples can be formulated as an extension of the framework in [7], as is briefly described at the end of this section. However, for simplicity of the presentation, we consider a restricted formulation that is sufficiently general to treat the example of learning a curve discussed in this paper. We now define more carefully what we mean by learning from generalized samples. Let X be the original instance space as before, and let the concept class F be a collection of real valued functions on X. In the usual model, the information one gets are samples $(x, f(x))$, where $x \in X$ and where $f \in F$ is the target concept. We can view this as obtaining a functional δ_x and applying this functional to the target concept f to obtain the sample $(\delta_x, \delta_x(f)) = (\delta_x, f(x))$. The functional in this case simply evaluates f at the point x and is chosen randomly from the class of all such "impulse" functionals. Instead, we now assume that we get generalized samples in the sense that we obtain a more general functional $\tilde{x}$, which is some mapping from F to R. The observed labeled sample is then $(\tilde{x}, \tilde{x}(f))$, consisting of the functional and the real number obtained by applying this functional to the target concept f. We assume the functional $\tilde{x}$ is chosen randomly from some collection of functionals $\tilde{X}$. Thus, $\tilde{X}$ is the instance space for the generalized samples, and the distribution P is a probability measure on $\tilde{X}$. Let S_F denote the set of labeled m samples for each $m \geq 1$ for each $\tilde{x} \in \tilde{X}$ and each $f \in F$, i.e.

$$S_F = \{(\tilde{x}_1, \tilde{x}_1(f)), \ldots, (\tilde{x}_m, \tilde{x}_m(f)) \mid m \geq 1, \tilde{x}_i \in \tilde{X}, f \in F\}.$$

Given P, we can define an error criterion (i.e., notion of distance between concepts) with respect to P as

$$d_P(f_1, f_2) = E|\tilde{x}(f_1) - \tilde{x}(f_2)|.$$

This is simply the average absolute difference of real numbers produced by generalized samples on the two concepts. Note that we could define the framework with more general loss criteria as in [7], but for the example considered in this paper, we use the criterion above.

Definition 1: *Learning from Generalized Samples:* Let $\mathcal{P}$ be a fixed and known collection of probability measures. Let F be a collection of functions from the instance space X into R, and let $\tilde{X}$ be the instance space of generalized samples for F. F is said to be *learnable with respect to* $\mathcal{P}$ *from the generalized samples* $\tilde{X}$ if there is a mapping $\mathcal{A} : S_F \to F$ for producing a hypothesis h from a set of labeled samples such that for every $\epsilon, \delta > 0$, there is a $0 < m = m(\epsilon, \delta) < \infty$ such that for every probability measure $P \in \mathcal{P}$ and every $f \in F$, if h is the hypothesis produced from a labeled m sample drawn according to P^m, then the probability that $d_P(f, h) < \epsilon$ is greater than $1 - \delta$.

If $\mathcal{P}$ is the set of all distributions over some σ algebra of $\tilde{X}$, then this corresponds to distribution-free learning from generalized samples. If $\mathcal{P}$ consists of a single distribution P, then this corresponds to fixed distribution learning from generalized samples. This is a direct extension of the usual definition of PAC learnability (see, for example, [5]) to learning functions from generalized samples over a class of distributions. In the definition, we have assumed that there is an underlying target concept f. As with the restrictions mentioned earlier, this could be removed following the framework of [7].

Learning with generalized samples can be easily transformed into an equivalent problem of PAC learning from standard samples. The concept class F on X corresponds naturally to a concept class $\tilde{F}$ on $\tilde{X}$ as follows. For a fixed $f \in F$, each functional $\tilde{x} \in \tilde{X}$ produces a real number when applied to f. Therefore, f induces a real valued function on $\tilde{X}$ in a natural way. The real valued function on $\tilde{X}$ induced by f will be denoted by $\tilde{f}$ and is defined by

$$\tilde{f}(\tilde{x}) = \tilde{x}(f).$$

The concept class $\tilde{F}$ is the collection of all functions on $\tilde{X}$ obtained in this way as f ranges through F.

We are now in the standard PAC framework with instance space $\tilde{X}$, concept class $\tilde{F}$, and distribution P on $\tilde{X}$. Hence, as usual, P induces a learning criterion or metric (actually only a pseudo-metric in general) on $\tilde{F}$, and as a result of the correspondence between F and $\tilde{F}$, this metric is equivalent to the (pseudo-)metric d_P induced by P on F defined above. This metric will be denoted by d_P over both F and $\tilde{F}$ and is given by

$$d_P(\tilde{f}_1, \tilde{f}_2) = E|\tilde{f}_1 - \tilde{f}_2| = E|\tilde{x}(f_1) - \tilde{x}(f_2)| = d_P(f_1, f_2).$$

Distribution-free and fixed distribution learnability are defined in the usual way for $\tilde{X}$ and $\tilde{F}$. Thus, a generalized notion of VC dimension for functions (called pseudo dimension in [7]) and metric entropy of $\tilde{F}$ characterize the learnability of $\tilde{F}$ in the distribution-free and fixed-distribution cases, respectively. These same quantities for $\tilde{F}$ then also characterize the learnability of F with respect to d_P.

Definition 2: *Metric Entropy:* Let (Y, ρ) be a metric space. A set $Y^{(\epsilon)}$ is said to be an ϵ cover (or ϵ approximation) for Y if for every $y \in Y$ there exists $y' \in Y^{(\epsilon)}$ such that $\rho(y, y') \leq \epsilon$. Define $N(\epsilon) \equiv N(\epsilon, Y, \rho)$ to be the smallest integer n such that there exists an ϵ cover for Y with n elements. If no such n exists, then $N(\epsilon, Y, \rho) = \infty$. The *metric entropy* of Y (which is often called the *ϵentropy*) is defined to be $\log_2 N(\epsilon)$.

$N(\epsilon)$ represents the smallest number of balls of radius ϵ that are required to cover Y. For convenience, if P is a distribution, we will use the notation $N(\epsilon, C, P)$ (instead of $N(\epsilon, C, d_P)$), and we will speak of the metric entropy of C with respect to P, with the understanding that the metric being used is $d_P(\cdot, \cdot)$.

Using results from [7] (based on results from [13]), we have the following result for learning from generalized samples with respect to a fixed distribution.

Theorem 1: F is learnable from generalized samples (or equivalently, $\tilde{F}$ is learnable) with respect to a distribution P if for each $\epsilon > 0$, there is a finite ϵ cover $\tilde{F}^{(\epsilon)}$ for $\tilde{F}$ (with respect to d_P) such that $0 \leq f_i \leq M(\epsilon)$ for each $f_i \in \tilde{F}^{(\epsilon)}$. Furthermore, a sample size

$$m(\epsilon, \delta) \geq \frac{2(M(\epsilon/2))^2}{\epsilon^2} \ln \frac{2|\tilde{F}^{(\epsilon/2)}|}{\delta}$$

is sufficient for (ϵ, δ) learnability.

Proof: Let $\tilde{F}^{(\epsilon/2)}$ be an $\frac{\epsilon}{2}$ cover with $0 \leq f_i \leq M(\epsilon/2)$ for each $f_i \in \tilde{F}^{(\epsilon/2)}$. Let $F^{(\epsilon/2)}$ be obtained from $\tilde{F}^{(\epsilon/2)}$ using the correspondence between F and $\tilde{F}$. After seeing $m(\epsilon, d)$ samples, let the learning algorithm output a hypothesis $h \in F^{(\epsilon/2)}$ that is most consistent with the data, i.e., that minimizes

$$\frac{1}{m(\epsilon, \delta)} \sum_{i=1}^{m(\epsilon,\delta)} |\tilde{x}_i(h) - y_i|$$

where $(\tilde{x}_i, y_i)$ are the observed generalized samples. Then, using Theorem 1 of [7], it follows that with probability greater than $1-\delta$, we have $d_P(f, h) \leq \epsilon$. □

Note that in the theorem above, the f_i in the ϵ cover need to be bounded, but this bound is allowed to depend on ϵ. Hence, $M(\epsilon)$ can be unbounded as a function of ϵ. Abe and Warmuth [1] have also considered this approach of using a bound that depends on ϵ. Note also that as in [7], we have taken the range of the functions to be nonnegative. All the results go through in the more general case, where the range of the functions is $[M_1(\epsilon), M_2(\epsilon)]$.

Although we will not use distribution-free learning in the example of learning a curve, for completeness, we give a result for this case.

Definition 3: *Pseudo Dimension:* Let F be a collection of functions from a set Y to R. For any set of points $\overline{y} = (y_1, \ldots, y_d)$ from Y, let $F_{|\overline{y}} = \{(f(y_1), \ldots, f(y_d)) : f \in F\}$. $F_{|\overline{y}}$ is a set of points in R^d. If there is some translation of $F_{|\overline{y}}$ that intersects all of the 2^d orthants of R^d, then $\overline{y}$ is said to be *shattered* by F. Following terminology from [7], the *pseudo dimension* of F, which we denote $\dim(F)$, is the largest integer d such that there exists a set of d points in Y that is shattered by F. If no such largest integer exists, then $\dim(F)$ is infinite.

We have the following result for distribution-free learning from generalized samples, again using results from [7].

Theorem 2: F is distribution-free learnable from generalized samples (or equivalently, $\tilde{F}$ is distribution-free learnable) if for some $M < \infty$, we have $0 \leq \tilde{f} \leq M$ for every $\tilde{f} \in \tilde{F}$ and if $\dim(\tilde{F}) = d$ for some $1 \leq d < \infty$. Furthermore, a sample size

$$m(\epsilon, \delta) \geq \frac{64M^2}{\epsilon^2}\left(2d \ln \frac{16eM}{\epsilon} + \ln \frac{8}{\delta}\right)$$

is sufficient for (ϵ, δ) distribution-free learnability.

Proof: The result follows from a direct application of Corollary 2 from [7], together with the correspondence between F and $\tilde{F}$ and the fact that $d_P(f_1, f_2) = d_P(\tilde{f}_1, \tilde{f}_2)$. □

Note that the metric entropy of $\tilde{F}$ is identical to the metric entropy of F (since both are with respect to d_P) so that the metric entropy of F characterizes learnability for a fixed distribution as well. However, the pseudo dimension of F with respect to X does *not* characterize distribution-free learnability. This quantity can be very different from the pseudo dimension of $\tilde{F}$ with respect to $\tilde{X}$.

As mentioned above, for simplicity, we have defined the concepts to be real valued functions, have chosen the generalized samples to return real values, and have selected a particular form for the learning criterion or metric d_P. Our ideas can easily be formulated in the much more general framework considered by Haussler [7]. Specifically, one could take F to be a family of functions with domain X and range Y.

The generalized samples $\tilde{X}$ would be a collection of mappings from F to $\tilde{Y}$. A family of functions $\tilde{F}$ mapping $\tilde{X}$ to $\tilde{Y}$ would be obtained from F by assigning to each $f \in F$ an $\tilde{f} \in \tilde{F}$ defined by $\tilde{f}(\tilde{x}) = \tilde{x}(f)$. As in [7], the distributions would be defined on $\tilde{X} \times \tilde{Y}$, a loss function L would be defined on $\tilde{Y} \times \tilde{Y}$, and for each $\tilde{f} \in \tilde{F}$, the error of the hypothesis $\tilde{f}$ with respect to a distribution would be $EL(\tilde{f}(\tilde{x}), \tilde{y})$, where the expectation is over the distribution on $(\tilde{x}, \tilde{y})$.

Although learning with generalized samples is, in essence, simply a transformation to a different standard learning problem, it allows the learning framework and results to be applied a broad range of problems. To show the variety in the type of observations that are available, we briefly mention some types of generalized samples that may be of interest in certain applications. In the case where the concepts are subsets of X (i.e., binary valued functions), some interesting generalized samples might be to draw random (parameterized) subsets (e.g., disks, lines, or other parameterized curves) of X labeled as to whether or not the random set intersects or is contained in the target concept. Alternatively, the random set could be labeled as to the number of intersections (or length, area, or volume of the intersection, as appropriate). In the case where the concepts are real valued functions, one might consider generalized samples consisting of certain random sets and returning the integral of the concept over these sets. For example, drawing random lines would correspond to tomographic-type problems with random ray sampling. Other possibilities might be to return weighted integrals of the concept where the weighting function is selected randomly from a suitable set (e.g., an orthonormal basis) or to sample derivatives of the concept at random points.

III. A Result from Stochastic Geometry

In this section, we state an interesting and well-known result from stochastic geometry. This result will be used in the next section in connection with a specific example of learning from generalized samples.

To state the result, we first need to describe the notion of drawing a "random" straight line, i.e., a uniform distribution for the set of straight lines intersecting a bounded domain. A line in the plane will be parameterized by the polar coordinates r, θ of the point on the line closest to the origin, where $r \geq 0$ and $0 \leq \theta \leq 2\pi$. The set (manifold) of all lines in the plane parameterized in this way corresponds to a semi-infinite cylinder.

A well-known result from stochastic geometry states that the unique measure (up to a scale factor) on the set of lines that is invariant to rigid transformations of the plane (translation, rotation) is $dr\, d\theta$, i.e., uniform density in r and θ [16]. This measure is thus independent of the choice of coordinate system and is referred to as the uniform measure (or density) for the set of straight lines in the plane. This measure corresponds precisely to the surface area measure on the cylinder.

From this measure, a uniform probability measure can be obtained for the set of all straight lines intersecting a bounded domain. Specifically, the set of straight lines intersecting a bounded domain X, which we will denote by $\tilde{X}$, is a bounded subset of the cylinder. The uniform probability measure on $\tilde{X}$ is then simply the surface area measure of the cylinder suitably normalized (i.e., by the area of $\tilde{X}$).

We can now state the following classic result from stochastic geometry (see e.g. [16], [3]).

Theorem 3: Let X be a bounded convex subset of R^2, and let $c \subset X$ be a curve of finite length. Suppose lines intersecting X are drawn uniformly, and let $n(\tilde{x}, c)$ denote the number of intersections of the random line $\tilde{x}$ with the curve c. Then,

$$E\, n(\tilde{x}, c) = \frac{2}{A} \mathrm{L}(c)$$

where $\mathrm{L}(c)$ denotes the length of the curve c, and A is the perimeter of X.

In the next section, for simplicity, we will take X to be the unit square. In this case, the theorem reduces simply to $En(\tilde{x}, c) = \frac{1}{2}\mathrm{L}(c)$.

A surprising (and powerful) aspect of this theorem is that the expected number of intersections a random line makes with the curve c depends only on the length of c but is independent of any other geometric properties of c. In fact, the expression on the left-hand side (suitably normalized) can be used as a definition for the length (or 1-D measure) of general sets in the plane [17].

An interesting implication of Theorem 3 is that the length of an unknown curve can be estimated or "learned" if one is told the number of intersections between the unknown curve and a collection of lines drawn randomly (from the uniform distribution). In fact, deterministic versions of this idea have been studied [19], [12].

IV. Learning a Curve by Counting Intersections with Lines

In this section, we consider a particular example of learning from generalized samples. For concreteness, we take X to be the unit square in R^2, although our results easily extend to the case where X is any bounded convex domain in R^2. We will consider concept classes C, which are collections of curves contained in X. For example, one particular concept class of interest will be the set of straight line segments contained in X. Other concept classes will consist of more general curves in X satisfying certain regularity constraints. The samples observed by the learner consist of randomly drawn straight lines that are labeled as to the number of intersections the random line makes with the target concept (i.e., the unknown curve). Recall, that with the r, θ parameterization, the set of lines intersecting X, which is the instance space $\tilde{X}$, is a bounded subset of the semi-infinite cylinder. We consider learnability with respect to a fixed distribution, where the distribution P is the uniform distribution on $\tilde{X}$.

A. Learning a Line Segment

Consider the case where C is the set of straight-line segments in X. In this case, given a concept $c \in C$, every straight line (except for a set of measure zero) intersects c either exactly once or not at all. Thus, $\tilde{C}$ consists of subsets (i.e., binary valued functions) of $\tilde{X}$, where each $\tilde{c} \in \tilde{C}$

contains exactly those straight lines $\tilde{x} \in \tilde{X}$ that intersect the corresponding $c \in C$.

The metric d_P on C and $\tilde{C}$ induced by P is given by

$$d_P(c_1, c_2) = d_P(\tilde{c}_1, \tilde{c}_2) = E|n(\tilde{x}, c_1) - n(\tilde{x}, c_2)|$$

where, as in the previous section, $n(\tilde{x}, c)$ is the number of intersections the line x makes with c. In the case of line segments, $n(\tilde{x}, c)$ is either one or zero, i.e., $\tilde{c}$ is binary valued; therefore

$$d_P(c_1, c_2) = d_P(\tilde{c}_1, \tilde{c}_2) = P(\tilde{c}_1 \Delta \tilde{c}_2)$$

where $\tilde{c}_1 \Delta \tilde{c}_2$ is the usual symmetric difference of $\tilde{c}_1$ and $\tilde{c}_2$.

In the case of line segments, a simple bound on the d_P distance between two segments can be obtained in terms of the distances between the endpoints of the segments.

Lemma 1: Let c_1, c_2 be two line segments, and let a_1, b_1 and a_2, b_2 be the endpoints of c_1 and c_2, respectively. Then

$$d_P(c_1, c_2) \leq \frac{1}{2}(||a_1 - a_2|| + ||b_1 - b_2||).$$

Proof: Since c_1, c_2 are line segments, the distance $d_P(c_1, c_2)$ between c_1 and c_2 is the probability that a random line intersects exactly one of c_1 and c_2. Any line that intersects exactly one of c_1, c_2 must intersect one of the segments $\overline{a_1 a_2}$ or $\overline{b_1 b_2}$ joining the endpoints of c_1 and c_2. Therefore

$$\begin{aligned} d_P(c_1, c_2) &\leq P(\tilde{x} \cap \overline{a_1 a_2} \neq \emptyset \text{ or } \tilde{x} \cap \overline{b_1 b_2} \neq \emptyset) \\ &\leq P(\tilde{x} \cap \overline{a_1 a_2} \neq \emptyset) + P(\tilde{x} \cap \overline{b_1 b_2} \neq \emptyset). \end{aligned}$$

Using Theorem 3, the probability that a random line intersects a line segment in the unit square is simply half the length of the line segment, from which the result follows. □

Using Lemma 1, we can bound the metric entropy of C (and, hence, $\tilde{C}$) with respect to the metric induced by P.

Lemma 2: Let C be the set of line segments contained in the unit square X, and let P be the uniform distribution on the set of lines intersecting X. Then

$$N(\epsilon, \tilde{C}, P) = N(\epsilon, C, P) \leq \frac{1}{4\epsilon^4}.$$

Proof: We construct an ϵ cover for C as follows. Consider a rectangular grid of points in X with spacing $\sqrt{2}\epsilon$. Let $C^{(\epsilon)}$ be the set of all line segments with endpoints on this grid. There are $\frac{1}{2\epsilon^2}$ points in the grid; therefore, there are $\frac{1}{4\epsilon^4}$ line segments in $C^{(\epsilon)}$. (We ignore the fact that some of these segments are actually just points since there are just $\frac{1}{2\epsilon^2}$ of these.) For any $c \in C$, there is a $c' \in C^{(\epsilon)}$ such that each endpoint of c' is within ϵ of an endpoint of c. Hence, from Lemma 1, $d_P(c, c') \leq \frac{1}{2}(\epsilon + \epsilon) = \epsilon$ so that $C^{(\epsilon)}$ is an ϵ cover for C with $\frac{1}{4\epsilon^4}$ elements. □

The construction of the previous lemma allows us to obtain the following learning result for straight line segments.

Theorem 4: Let C be the set of line segments in the unit square X. Then, C is learnable by counting intersections with straight lines drawn uniformly using

$$m(\epsilon, \delta) = \frac{2}{\epsilon^2} \ln \frac{8}{\epsilon^4 \delta}$$

samples.

Proof: Let $\tilde{C}$ be the concept class over $\tilde{X}$ corresponding to C. Then, $\tilde{c} \in \tilde{C}$ is defined by $\tilde{c}(\tilde{x}) = n(\tilde{x}, c)$, i.e., $\tilde{c}(\tilde{x})$ is the number of intersections of the line $\tilde{x}$ with c. Clearly, $0 \leq \tilde{c} \leq 1$ (except for a set of measure zero) for every $\tilde{c} \in \tilde{C}$. Using the construction of Lemma 2, we have an $\frac{\epsilon}{2}$ cover of $\tilde{C}$ with $4/\epsilon^4$ elements. Hence, the result follows from Theorem 1.

B. Learning Curves of Bounded Turn and Length

Now, we consider the learnability of a much more general class of curves. First, we need some preliminary definitions. We will consider rectifiable curves parameterized by arc length s so that a curve c of length L is given by

$$c = \{(x_1(s), x_2(s)) \mid 0 \leq s \leq L\}$$

where $x_1(\cdot)$ and $x_2(\cdot)$ are absolutely continuous functions from $[0, L]$ to R such that $\sqrt{\dot{x}_1^2 + \dot{x}_2^2}$ is defined and equal to unity almost everywhere. If x_1 and x_2 are twice differentiable at s, then the curvature of c at s, denoted $\kappa(s)$, is defined as the rate of change of the direction of the tangent to the curve at s and is given by $\kappa(s) = \ddot{x}_2 \dot{x}_1 - \ddot{x}_1 \dot{x}_2$. The total absolute curvature of c is given by $\int_0^L |\kappa(s)|\, ds$.

Alexandrov and Reshetnyak [2] have developed an interesting theory for irregular curves. Among other things, they study the notion of the "turn" of a curve, which is a generalization of total absolute curvature to curves that are not necessarily twice differentiable. For example, for a piecewise linear curve, the turn is simply the sum of the absolute angles that the tangent turns between adjacent segments. The turn for more general curves can be obtained by piecewise linear approximations. In fact, this is precisely the manner in which turn is defined [2].

Definition 4: *Turn:* Let $\overline{v_0 \cdots v_n}$ denote a piecewise linear curve with vertices $v_0, \ldots, v_n$. Let a_i be the vector $\overline{v_{i-1} v_i}$, and let ϕ_i be the angle between the vector a_i and a_{i+1}, that is, ϕ_i represents the total angle through which the tangent to the curve turns at vertex i (π minus the interior angle at vertex i). The turn of $\overline{v_0 \cdots v_n}$, which is denoted $\kappa(\overline{v_0 \cdots v_n})$, is defined by

$$\kappa(\overline{v_0 \cdots v_n}) = \sum_{i=1}^{n-1} \phi_i.$$

The turn of a general parameterized curve c, which is denoted $\kappa(c)$, is defined as the supremum of the turn over all piecewise linear curves inscribed in c, i.e.

$$\kappa(c) = \sup\{\kappa(\overline{c(s_0) \cdots c(s_n)}) \mid 0 \leq s_0 < s_1 < \cdots < s_n \leq L\}$$

where L is the length of c. As expected, the notion of turn reduces to the total absolute curvature of a curve when the latter quantity is defined [2]. We will use the generalized notion of turn throughout; therefore, our results will apply to curves that are not necessarily twice differentiable (e.g., piecewise linear curves).

We will consider classes of curves of bounded length and bounded turn. Specifically, let $C_{K,L}$ be the set of all curves contained in the unit square whose length is less than or equal to L and whose turn is less than or equal to K. Note that for curves contained in a bounded domain, the length of a curve can be bounded in terms of the turn of the curve and the diameter of the domain (Theorem 5.6.1 from [2]; for differentiable curves, see, for example, p. 35 of [16]). Hence, we really need only consider classes of curves with a bound on the turn. However, for convenience, we will carry both parameters K and L explicitly.

As before, the samples will be random lines drawn according to the uniform distribution P on $\tilde{X}$, which is labeled as to the number of intersections the line makes with the unknown curve c. However, with curves in $C_{K,L}$, the number of intersections with a given line can be any positive integer as opposed to just zero or one for straight line segments. (Note that by Theorem 3, the probability that a random line has an infinite number of intersections with a given curve is zero; therefore, the number of intersections is a well-defined integer-valued function.) Thus, the class $\tilde{C}_{K,L}$ consists of a collection of integer-valued functions on $\tilde{X}$ as opposed to just subsets of $\tilde{X}$, as in the previous section.

In addition, as before, the results on learning for the set of curves will be with respect to the metric d_P induced by the measure P, that is, the d_P distance between two curves c_1 and c_2 or their corresponding functions $\tilde{c}_1$, $\tilde{c}_2$ is given by

$$d_P(c_1, c_2) = d_P(\tilde{c}_1, \tilde{c}_2) = E|n(\tilde{x}, c_1) - n(\tilde{x}, c_2)|$$

where the expectation is taken over the random line $\tilde{x}$ with respect to the uniform measure P. This notion of distance between curves has been studied previously (e.g., see [19] and p. 38 of [16]). For example, on the set of rectifiable curves, d_P satisfies the triangle inequality and is always nonnegative. It is not known whether on the set of rectifiable curves, d_P satisfies the property that $d_P(c_1, c_2) = 0$ implies $c_1 = c_2$. However, on the set of curves of bounded turn, this property does hold [15]. Hence, over the class of curves $C_{K,L}$, d_P is, in fact, a metric. (Note that in [16] and [19], the notion of distance used is actually $\frac{1}{2}d_P$, but this makes no difference in the metric properties.)

To obtain a learning result for $C_{K,L}$, we will show that each curve in $C_{K,L}$ can be approximated (with respect to d_P) by a bounded number of straight-line segments. The metric entropy computation for a single straight-line segment can be extended to provide a metric entropy bound for curves consisting of a bounded number of straight-line segments. Thus, by combining these two ideas, we can obtain a metric entropy bound for $C_{K,L}$ that yields the desired learning result.

First, we need several properties of the d_P metric for curves of bounded turn.

Lemma 3: If c_1, c_2 are curves with a common endpoint (so that $c_1 \cup c_2$ is a curve) and similarly for c'_1, c'_2, then

$$d_P(c_1 \cup c_2, c'_1 \cup c'_2) \le d_P(c_1, c'_1) + d_P(c_2, c'_2).$$

Proof: For any line $\tilde{x}$ (except for a set of measure zero), $n(\tilde{x}, c_1 \cup c_2) = n(\tilde{x}, c_1) + n(\tilde{x}, c_2)$ and similarly for c'_1, c'_2. Therefore

$$\begin{aligned} d_P(c_1 \cup c_2, c'_1 \cup c'_2) &= \frac{1}{2}E|n(\tilde{x}, c_1 \cup c_2) - n(\tilde{x}, c'_1 \cup c'_2)| \\ &= \frac{1}{2}E|n(\tilde{x}, c_1) - n(\tilde{x}, c'_1) + n(\tilde{x}, c_2) - n(\tilde{x}, c'_2)| \\ &\le \frac{1}{2}E|n(\tilde{x}, c_1) - n(\tilde{x}, c'_1)| + \frac{1}{2}E|n(\tilde{x}, c_2) - n(\tilde{x}, c'_2)| \\ &= d_P(c_1, c'_1) + d_P(c_2, c'_2). \end{aligned}$$

□

By induction, this result can clearly be extended to unions of any finite number curves. The case of a finite number of curves will be used later in Lemma 6.

Lemma 4: If c is a curve and $\hat{c}$ is the line segment joining the endpoints of c, then

$$d_P(c, \hat{c}) = \frac{1}{2}(\text{L}(c) - \text{L}(\hat{c})).$$

Proof: Each line can intersect $\hat{c}$ at most once, and every line intersecting $\hat{c}$ also intersects c. Therefore, $n(\tilde{x}, c) \ge n(\tilde{x}, \hat{c})$ so that $|n(\tilde{x}, c) - n(\tilde{x}, \hat{c})| = n(\tilde{x}, c) - n(\tilde{x}, \hat{c})$ for all lines $\tilde{x}$ (except a set of measure zero). Hence

$$\begin{aligned} d_P(c, \hat{c}) &= E|n(\tilde{x}, c) - n(\tilde{x}, \hat{c})| \\ &= E(n(\tilde{x}, c) - n(\tilde{x}, \hat{c})) \\ &= \frac{1}{2}\text{L}(c) - \frac{1}{2}\text{L}(\hat{c}) \end{aligned}$$

where the last equality follows from the stochastic geometry result (Theorem 3). □

We will make use of the following result from [2].

Theorem 5: *Alexandrov and Reshetnyak:* Let c be a curve in R^n with $\kappa(c) < \pi$, and let α be the distance between its endpoints. Then

$$\text{L}(c) \le \frac{\alpha}{\cos \frac{\kappa(c)}{2.}}$$

Equality is obtained iff c consists of two line segments of equal length.

Lemma 5: For $0 \le \alpha \le \pi/6$, $1/\cos\alpha - 1 \le \alpha^2$ so that if c is a curve with turn $\kappa(c) \le \pi/6$ and $\hat{c}$ is the line connecting the endpoints of c, then

$$d_P(c, \hat{c}) \le \frac{\text{L}(\hat{c})}{8}\kappa^2(c).$$

Proof: Let $g(\alpha) = 1/\cos\alpha$ and $h(\alpha) = 1 + \alpha^2$. For $0 \le \alpha \le \pi/6$, $\sin\alpha \le 1/2$, and $\cos\alpha \ge \sqrt{3}/2$ so that $\ddot{g}(\alpha) = 2\sin^2\alpha/\cos^3\alpha + 1/\cos\alpha \le \frac{4}{3\sqrt{3}} + \frac{2}{\sqrt{3}} = \frac{10\sqrt{3}}{9} < 2 = \ddot{h}(\alpha)$. Combining $\ddot{g}(\alpha) < \ddot{h}(\alpha)$ with the fact that $g(0) = h(0)$ and $\dot{g}(0) = \dot{h}(0)$ gives $g(\alpha) \le h(\alpha)$; therefore, $1/\cos\alpha - 1 \le \alpha^2$ for $0 \le \alpha \le \pi/6$.

Now, using the above result, Lemma 4, and Theorem 5, we have

$$\begin{aligned} d_P(c, \hat{c}) &= \frac{1}{2}(\text{L}(c) - \text{L}(\hat{c})) \\ &\le \frac{1}{2}\text{L}(\hat{c})\left(\frac{1}{\cos(\kappa(c)/2)} - 1\right) \le \frac{\text{L}(\hat{c})}{8}\kappa^2(c). \end{aligned}$$

□

Lemma 6: If $c \in C_{K,L}$, then for all sufficiently small $\epsilon > 0$ (e.g., $\epsilon \leq \pi KL/48$), the curve c can be approximated to within ϵ by an inscribed piecewise linear curve with at most $\frac{K^2L}{8\epsilon}$ segments.

Proof: As usual, let s denote arc length along c. Since $\kappa(c) \leq K$ for any $\alpha > 0$, we can find a decomposition of c into at most $\lceil K/\alpha \rceil$ pieces $\ell_1, \ldots, \ell_{\lceil K/\alpha \rceil}$ such that $\kappa(\ell_i) \leq \alpha$ for each i. For example, let $s_0 = 0$, and let

$$s_i = \sup\{s_{i-1} \leq s \leq L | \kappa(c(s_{i-1}, s)) \leq \alpha\}$$

where $c(s_{i-1}, s)$ is the part of the curve c between arc length s_{i-1} and s inclusive. Then, let $\ell_i = c(s_{i-1}, s_i)$. Theorem 5.1.1, which is on p. 120 of [2], states that if a sequence of curves c_m converges to a curve c, then $\kappa(c) \leq \liminf_{m\to\infty} \kappa(c_m)$. Using this result and the definition of the s_i, we have that $\kappa(\ell_i) \leq \alpha$. Another result from [2] (Theorem 5.1.3 on p. 122) states that if $c(s)$, $0 \leq s \leq L$ is a curve of finite turn, then for any $0 < t < L$, we have $\kappa(c) = \kappa(c(0,t)) + \kappa(c(t,L)) + \phi(c(t))$, where $\phi(t)$ is the angle through which the tangent turns at the point $c(t)$ (i.e., the angle between the left-hand and the right-hand tangents). From the definition of s_i and the property that $\kappa(c(s, s')) \to 0$ as $s' \to s$ from the right (see Corollary 3 on p. 121 of [2]), we have that $\kappa(c(s_{i-1}, s_i)) + \phi(c(s_i)) \geq \alpha$. It follows that if $s_i < L$, then for any $\eta > 0$, $\kappa(c(0, s_i + \eta)) \geq i\alpha$. Since $\kappa(c) \leq K$, we must have $s_i = L$ for some $i \leq \lceil K/\alpha \rceil$.

Now, let $\hat{\ell}_i$ be the line segment joining the ends of ℓ_i. Clearly, the union of the $\hat{\ell}_i$ forms a piecewise linear curve inscribed in c (i.e., with endpoints of the segments lying on c). For $\alpha \leq \pi/6$, from Lemmas 3 and 5, and the fact that $\mathrm{L}(\hat{\ell}_i) \leq \mathrm{L}(c) \leq L$, we have

$$d_P(c, \cup_{i=1}^{K/\alpha} \hat{\ell}_i) \leq \sum_{i=1}^{K/\alpha} d_P(\ell_i, \hat{\ell}_i) \leq \frac{K}{\alpha} \cdot \frac{L}{8} \alpha^2 = \frac{KL}{8} \alpha.$$

Thus, if $\alpha \leq \frac{8\epsilon}{KL}$ (with $\epsilon \leq \pi KL/48$), then $d_P(c, \cup_{i=1}^{K/\alpha} \hat{\ell}_i) \leq \epsilon$ so that $\frac{K}{\alpha} = \frac{K^2L}{8\epsilon}$ segments suffice for an ϵ approximation to c by an inscribed piecewise linear curve. □

Theorem 6: Let $C_{K,L}$ be the set of all curves in the unit square with turn bounded by K and length bounded by L. Let P be the uniform distribution on the set of lines intersecting the unit square, and let d_P be the metric on $C_{K,L}$ defined by $d_P(c_1, c_2) = E|n(\tilde{x}, c_1) - n(\tilde{x}, c_2)|$. Then, the metric entropy of $C_{K,L}$ with respect to d_P satisfies

$$N(\epsilon, C_{K,L}, P) \leq \left(\frac{K^4L^2}{8\epsilon^4}\right)^{1+\frac{K^2L}{4\epsilon}}$$

Proof: We construct an ϵ cover for C as follows. Consider a rectangular grid of points in the unit square with spacing $\frac{2\sqrt{2}\epsilon^2}{K^2L}$. Let $C_{K,L}^{(\epsilon)}$ be the set of all piecewise linear curves with at most $\frac{K^2L}{4\epsilon}$ segments, where the endpoints all lie on this grid. There are $K^4L^2/8\epsilon^4$ points in the grid so that there are at most $(K^4L^2/8\epsilon^4)^{1+K^2L/4\epsilon}$ distinct curves in $C_{K,L}^{(\epsilon)}$.

To show that $C_{K,L}^{(\epsilon)}$ is an ϵ cover for $C_{K,L}$, let $c \in C_{K,L}$. By Lemma 6, there is a piecewise linear curve $\hat{c}$ with at most $\frac{K^2L}{4\epsilon}$ segments such that $d_P(c, \hat{c}) < \epsilon/2$. We can find a curve $c' \in C_{K,L}^{(\epsilon)}$ close to $\hat{c}$ by finding a point on the grid within $\frac{2\epsilon^2}{K^2L}$ of each endpoint of a segment in $\hat{c}$. By Lemma 1, each line segment of c' is a distance at most $\frac{2\epsilon^2}{K^2L}$ (with respect to d_P) from the corresponding line segment of $\hat{c}$. Since $\hat{c}, c'$ consist of at most $\frac{K^2L}{4\epsilon}$ segments, applying Lemma 3, we get $d_P(\hat{c}, c') \leq \epsilon/2$. Hence, by the triangle inequality, $d_P(c, c') \leq \epsilon$. □

We can now prove a learning result for curves of bounded turn and length.

Theorem 7: Let $C_{K,L}$ be the set of all curves in the unit square with turn bounded by K and length bounded by L. Then, $C_{K,L}$ is learnable by counting intersections with straight lines drawn uniformly using

$$m(\epsilon, \delta) = \frac{K^4L^2}{2\epsilon^4} \ln \frac{2}{\delta} \left(\frac{2K^4L^2}{\epsilon^4}\right)^{1+\frac{K^2L}{2\epsilon.}}$$

Proof: Let $\tilde{C}_{K,L}$ be the concept class over $\tilde{X}$ corresponding to $C_{K,L}$. Then, $\tilde{c} \in \tilde{C}$ is defined by $\tilde{c}(\tilde{x}) = n(\tilde{x}, c)$, i.e., $\tilde{c}(\tilde{x})$ is the number of intersections of the line $\tilde{x}$ with c.

Using the construction of Theorem 6, we have an $\frac{\epsilon}{2}$ cover $\tilde{C}^{(\epsilon/2)}$ of $\tilde{C}_{K,L}$ with $(2K^4L^2/\epsilon^4)^{1+K^2L/2\epsilon}$ elements. Furthermore, each element of the $\frac{\epsilon}{2}$ cover consists of at most $\frac{K^2L}{2\epsilon}$ line segments. Since a line $\tilde{x}$ can intersect each segment at most once, we have $0 \leq \tilde{c}_i(\tilde{x}) \leq \frac{K^2L}{2\epsilon}$ for every $\tilde{c}_i \in \tilde{C}^{(\epsilon/2)}$. Hence, the result follows from Theorem 1. □

This learning result is in terms of d_P, which is a metric induced by the uniform measure on the set of lines. Although some properties of this metric are known, to better understand the implications of the learning result, it would be useful to obtain further properties of this metric. Alternatively, we could try to obtain a learning result with respect to other more standard metrics. For example, a common measure of distance between curves is Hausdorff metric $d_H(\cdot,\cdot)$ defined as

$$d_H(c_1, c_2) = \inf\{\eta : c_1 \subset c_2^{(\eta)} \text{ and } c_2 \subset c_1^{(\eta)}\}$$

where $c^{(\eta)}$ is the η neighborhood of c i.e.

$$c^{(\eta)} = \{x \ : \ \inf_{y \in c} |x - y| < \eta\}.$$

A sufficient condition for learning with respect to d_H (d_H could actually be replaced by any other metric) is

$$\inf_{\{c_1, c_2 \ | \ d_H(c_1, c_2) > \epsilon\}} d_P(c_1, c_2) > 0.$$

This result combined with the learning result with respect to d_P would immediately imply a learning result with respect to d_H. However, this condition is not satisfied since c_1 and c_2 can can have arbitrarily small lengths (so that $d_P(c_1, c_2)$ is arbitrarily small) but with c_1 and c_2 spatially separated so that $d_H(c_1, c_2)$ is greater than some fixed constant. We can eliminate this problem by considering the class $C_{K,\ell,L}$ consisting of all curves in $C_{K,L}$ whose lengths are larger than some fixed ℓ, that is, $C_{K,\ell,L}$ consists of all curves with length between ℓ and L and with turn less than K. Using results from [2] and [15], it can be shown that the condition above is satisfied for curves in $C_{K,\ell,L}$. Hence, we have a learning result for $C_{K,\ell,L}$ with respect to d_H (and with respect to a stronger metric defined in [2] as well).

It is interesting to note that $\tilde{C}_{K,L}$ has infinite pseudodimension (generalized VC dimension). In general, infinite pseudodimension does not necessarily imply that a class is not distribution-free learnable; however, in the present case, we expect that $C_{K,L}$ is not distribution-free learnable. That the pseudodimension is infinite can be seen as follows. First, assume that $K, L \geq 2\pi$. For each k, let $\tilde{x}_1, \ldots, \tilde{x}_k$ be the set of lines corresponding to the sides of a k-gon inscribed in the unit circle. For any subset G of these k lines, we can find a curve $c_G \in C_{K,L}$ so that $n(\tilde{x}_i, c_G) = 2$ for $\tilde{x}_i \in G$ and $n(\tilde{x}_i, c_G) = 0$ for $x_i \notin G$. Such a curve can be obtained by taking a point on the unit circle in each arc corresponding to $\tilde{x}_i \in G$ and taking c_G to be the boundary of the convex hull of these points. Then, $\kappa(c_G) = 2\pi$ and $L(c_G) < 2\pi$ so that $c_G \in C_{K,L}$. Thus, the set $\tilde{x}_1, \ldots, \tilde{x}_k$ is shattered by $\tilde{C}_{K,L}$, and since k is arbitrary, the pseudodimension of $\tilde{C}_{K,L}$ is infinite. For $K, L < 2\pi$, we can apply essentially the same construction over an arc of the unit circle and without taking c_G to be a closed curve.

C. Connections with the Stochastic Geometry Result

For the class of curves whose length and curvature are bounded by constants, the learnability result of Theorem 7 can be thought of as a refinement of the stochastic geometry result. First, using the expression for the expected number of intersections, one can estimate or "learn" the length of c from a set of generalized samples. The learnability result makes the much stronger statement that the curve c itself can be learned (from which the length can then be estimated). To show that the length can be estimated, we need only note that

$$\begin{aligned}|\mathrm{L}(c_1) - \mathrm{L}(c_2)| &= |\frac{1}{2}E(n(y, c_1) - n(y, c_2))| \\ &\leq \frac{1}{2}E|n(y, c_1) - n(y, c_2)| = \frac{1}{2}d_P(c_1, c_2)\end{aligned}$$

so that if we learn c to within ϵ, then the length of c can be obtained to within $\epsilon/2$.

Second, for the class of curves considered, we have a ***uniform*** learning result. Hence, this refines the stochastic geometry result by guaranteeing uniform convergence of empirical estimates of length to the true length for the class of curves considered.

V. Discussion

We introduced a model of learning from generalized samples and considered an application of this model to a problem of reconstructing a curve by counting intersections with random lines. The curve reconstruction problem is closely related to a well-known result from stochastic geometry. The stochastic geometry result (Theorem 3) suggests that the length of a curve can be estimated by counting the number of intersections with an appropriate set of lines, and this has been studied by others. Our results show that for certain classes of curves, the curve itself can be learned from such information. Furthermore, over these classes of curves, the estimates of length from a random sample converge uniformly to the true length of a curve.

There are a number of interesting questions/possible extensions concerning the problem of learning a curve. We have not considered the question of computational complexity of algorithms for learning a curve from intersections with lines. The results of Lemma 2 and Theorem 6 suggest exhaustive procedures that simply consider all curves in an ϵ cover and select one that is most consistent with the line crossing data. For the case of learning a straight-line segment, this trivial procedure runs in time polynomial in $1/\epsilon$. However, in the case of learning a curve of bounded length and turn, this simplistic approach requires time exponential in $1/\epsilon$, K, and L. We leave open the question of whether there exists a polynomial time algorithm for the problem of learning a curve of bounded length and turn.

The stochastic geometry result holds for any bounded convex subset of the plane, and as we mentioned before, our results can be extended to this case as well. Furthermore, results that are analogous to Theorem 3 can be shown in higher dimensions and in some nonEuclidean spaces [16]. Some results on curves of bounded turn that are analogous to those we needed can also be obtained more generally [2]. Hence, learning results should be obtainable for these cases.

Regarding other possible extensions of the problem of learning a curve, note that the stochastic geometry result is not true for distributions other than the uniform distribution. In addition, we are not aware of any generalizations to cases where parameterized curves other than lines are drawn randomly. However, learnability results likely hold true for some other distributions and perhaps for other randomly drawn parameterized curves, although the metric entropy computations may be difficult.

There is an interesting connection between the problem of learning a curve discussed here and a problem of computing the length of curves from discrete approximations. In particular, it can be shown that computing the length of a curve from its digitization on a rectangular grid requires a nonlocal computation (even for just straight line segments), although computing the length of a line segment from discrete approximations on a random tesselation can be done locally [10]. The construction is essentially a learning problem with intersection samples from random straight lines. Furthermore, the construction provides insight as to why local computation fails for a rectangular digitization and suggests that appropriate deterministic digitizations would still allow local computations. This is related to the work in [12].

We considered here only one particular example of learning from generalized samples. However, we expect that this framework can be applied to a number of problems in signal/image processing, geometric reconstruction, stereology, etc., to provide learnability results and sample size bounds under a PAC criterion. As previously mentioned, learning with generalized samples is, in essence, simply a transformation to a different standard learning problem, although the variety available in choosing this transformation (i.e., the form of the generalized samples) should allow the learning framework and results to be applied to a broad range of problems.

For example, the generalized samples could consist of drawing certain random sets and returning the integral of the concept over these sets. Other possibilities might be to return weighted integrals of the concept where the weighting function is selected randomly from a suitable set (e.g., an orthonormal basis) or to sample derivatives of the concept at random points.

One interesting application would be to problems in tomographic reconstruction. In these problems, one is interested in reconstructing a function from a set of projections of the function onto lower dimensional subspaces. One could have the generalized samples consist of drawing random lines labeled according to the integral of the unknown function along the line. This would correspond to a problem in tomographic reconstruction with random ray sampling. Alternatively, as previously mentioned, one could combine the general framework discussed by Haussler [7] with generalized samples and consider an application to tomography where the generalized samples consist of entire projections. This would be more in line with standard problems in tomography, where the directions of the projections are chosen randomly, however.

For more geometric problems in which the concepts are subsets of X, some interesting generalized samples might be to draw random (parameterized) subsets (e.g., disks, lines, or other parameterized curves) of X labeled as to whether or not the random set intersects or is contained in the target concept. Other possibilities might be to label the random set as to the number of intersections (or length, area, or volume of the intersection, as appropriate) with the unknown concept. One interesting application to consider would be the reconstruction of a convex set from various types of data (e.g., see [8], [18], [11], and [14]). For example, the generalized samples could be random lines labeled as to whether or not they intersect the convex set (which would provide bounds on the support function). This is actually just a special case of learning a curve, which is closed and convex, although tighter bounds should be obtainable due to the added restrictions. Alternatively, the lines could be labeled as to the length of the intersection (which is like the tomography problem with random ray sampling in the case of binary objects). A third possibility (which is actually just learning from standard samples) would be to obtain samples of the support function.

Formulating learning from generalized samples in the general framework of Haussler [7] allows issues such as noisy samples to be treated in a unified framework. Application of the framework to a particular problem reduces the question of estimation/learning under a PAC criterion to a metric entropy (or generalized VC dimension) computation. This is not meant to imply that such a computation is easy. On the contrary, the metric entropy computation is the essence of the problem and can be quite difficult. Another problem that can be difficult is interpreting the learning criterion on the original space induced by the distribution on the generalized samples. The induced metric is a natural one, given the type of information available, but it may be difficult to understand the properties it endows on the original concept class. Finally, although this approach may provide sample size bounds for a variety of problems, it leaves open the question of finding good algorithms.

ACKNOWLEDGMENT

We would like to thank T. Richardson for helpful discussions regarding relationships of the metric d_P to the Hausdorff metric. We would also like to thank the anonymous referees for a number of helpful suggestions.

REFERENCES

[1] N. Abe and M. K. Warmuth, "On the computational complexity of approximating distributions by probabilistic automata," in *Proc. Third Ann. Workshop Comput. Learning Theory*, 1990, pp. 52-66.

[2] A. D. Alexandrov and Yu. G. Reshetnyak, *General Theory of Irregular Curves*, Mathematics and Its Applications (Soviet series). Boston: Kluwer, 1989, vol. 29.

[3] A. J. Baddeley, "Stochastic geometry and image analysis," in *CWI Monographs*. Amsterdam: North Holland, 1986.

[4] G. M. Benedek and A. Itai, "Learnability by fixed distributions," in *Proc. First Workshop Computat. Learning Theory*, 1988, pp. 80-90.

[5] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth, "Classifying learnable geometric concepts with the Vapnik-Chervonenkis dimension," in *Proc. 18th ACM Symp. Theory Comput.* (Berkeley, CA), 1986, pp. 273-282.

[6] R. M. Dudley, "Central limit theorems for empirical measures," *Ann. Probab.*, vol. 6, pp. 899–929, 1978.

[7] D. Haussler, "Decision theoretic generalizations of the PAC learning model for neural net and other learning applications," *Inform. Comput.*, vol. 100, pp. 78–150, 1992.

[8] W. C. Karl, "Reconstructing objects from projections," Ph.D. thesis, Dept. of EECS, Mass. Inst. of Technol., Feb. 1991.

[9] S. R. Kulkarni, "On metric entropy, Vapnik-Chervonenkis dimension and learnability for a class of distributions," Cent. for Intell. Contr. Syst. Rep. CICS-P-160, Mass. Inst. of Technol., 1989.

[10] ———, "Problems of computational and information complexity in machine vision and learning," Ph.D. thesis, Dept. of Elect. Eng. Comput. Sci., Mass. Inst. of Technol., June 1991.

[11] A. S. Lele, S. R. Kulkarni, and A. S. Willsky, "Convex set estimation from support line measurements and applications to target reconstruction from laser radar data," in *SPIE Proc. Laser Radar V*, 1990, pp. 58-82, vol. 1222 (submitted to *J. Opt. Soc. Amer.*).

[12] P. A. P. Moran, "Measuring the length of a curve," *Biometrika*, vol. 53, pp. 359-364, 1966.

[13] D. Pollard, *Convergence of Stochastic Processes*. New York: Springer-Verlag, 1984.

[14] J. L. Prince and A. S. Willsky, "Estimating convex sets from noisy support line measurements," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 12, pp. 377-389, 1990.

[15] T. J. Richardson, forthcoming, 1992.

[16] L. A. Santalo, "Integral geometry and geometric probability," *Encyclopedia of Mathematics and its Applications.* Reading, MA: Addison-Wesley, 1976, vol. 1.

[17] S. Sherman, "A comparison of linear measures in the plane," *Duke Math. J.*, vol. 9, pp. 1–9, 1942.

[18] S. S. Skiena, "Geometric probing," Ph.D. thesis, Dept. of Comput. Sci., Univ. of Illinois, Urbana-Champaign, Rep. no. UIUCDCS-R-88-1425, Apr. 1988.

[19] H. Steinhaus, "Length, shape, and area," *Colloquium Mathematicum*, vol. 3, pp. 1–13, 1954.

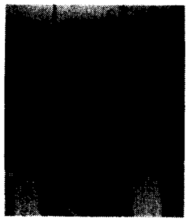
[20] L. G. Valiant, "A theory of the learnable," *Comm. ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[21] V. N. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of relative frequencies to their probabilities," *Theory Probab. Applications*, vol. 16, no. 2, pp. 264-280, 1971.

[22] ———, "On the uniform convergence of relative frequencies to their probabilities," *Theory Probab. Applications*, vol. 16, no. 2, pp. 264-280, 1971.

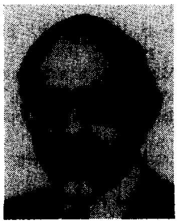
[23] V. N. Vapnik, *Estimation of Dependencies Based on Empirical Data.* Berlin: Springer-Verlag, 1982.

[24] G. Wahba, "Spline models for observational data," in*Series in Applied Mathematics.* New York: SIAM, 1990, vol. 59.



Sanjeev R. Kulkarni (M'91) received the B.S. degree in mathematics, the B.S. degree in electrical engineering, and the M.S. degree in mathematics from Clarkson University in 1983, 1984, and 1985, respectively, the M.S. degree in electrical engineering from Stanford University in 1985, and the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology (MIT) in 1991.

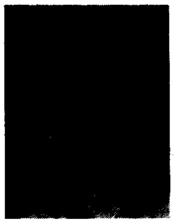
From 1985 to 1991, he was a Member of the Technical Staff at the MIT Lincoln Laboratory, where he worked on the modeling and processing of laser radar measurements. In the spring of 1986, he was a part-time faculty at the University of Massachusetts, Boston. Since 1991, he has been an Assistant Professor of Electrical Engineering at Princeton University. His research interests include signal and image processing, machine learning, and pattern recognition.



Sanjoy K. Mitter (F'79) received the Ph.D. degree from the Imperial College of Science and Technology, University of London, in 1965.

He has previously worked as a research engineer at Brown Boveri & Co., Ltd., Switzerland (now ASEA Brown Boveri), and Battelle Institute, Geneva, Switzerland. He taught at Case Western Reserve University, Cleveland, OH, from 1965–1969 and joined the Massachusetts Institute of Technology (MIT), Cambridge, in 1969, first as a Visiting Professor and then, in 1970, as an Associate Professor in the Department of Electrical Engineering and Computer Science. He is currently Professor of Electrical Engineering and Co-Director of the Laboratory for Information and Decision Systems. He is also Director of the Center for Intelligent Control Systems, which is an interuniversity (Brown-Harvard-MIT) center for research on the foundations of intelligent systems. He has held visiting positions at the Tata Institute of Fundamental Research, Bombay, India; the Scuola Normale Superiore, Pisa, Italy; the Imperial College of Science and Technology; the Institut National de Recherche en Informatique et en Automatique, France; the University of Groningen, the Netherlands; and several universities in the United States. His research has spanned the broad areas of systems, communications, and control. Although his primary contributions have been on the theoretical foundations of the field, he has also contributed to significant engineering applications, notably in the control of interconnected power systems and automatic recognition and classification of electrocardiograms. His current research interests are theory of stochastic dynamical systems; nonlinear filtering, stochastic and adaptive control; mathematical physics and its relationship to system theory; image analysis and computer vision; and structure, function, and organization of complex systems.

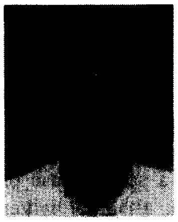
Professor Mitter has served on several advisory committees and editorial boards for IEEE, SIAM, AMS, NSF, and ARO. He is currently Associate Editor of *Acta Applicandae Mathematicae; Circuits, Systems, and Signal Processing; Journal of Applied Mathematics and Optimization; SIAM Review;* and the *ULAM Quarterly*. In 1988, he was elected to the National Academy of Engineering.



John N. Tsitsiklis (M'83) was born in Thessloniki, Greece, in 1958. He received the B.S. degree in mathematics in 1980 and the B.S., M.S., and Ph.D. degrees in electrical engineering in 1980, 1981, and 1984, respectively, from the Massachusetts Institute of Technology (MIT).

During the academic year 1983–1984, he was an acting Assistant Professor of Electrical Engineering at Stanford University. Since 1984, he has been with the Department of Electrical Engineering and Computer Science at MIT, where he is currently a Professor. His research interests are in the areas of parallel and distributed computation, systems and control theory, and operations research.

Dr. Tsitsiklis is coauthor of *Parallel and Distributed Computation: Numerical Methods* (1989). He has been a recipient of an IBM Faculty Development Award (1983), an NSF Presidential Young Investigator Award (1986), an Outstanding Paper Award from the IEEE Control Systems Society, and the Edgerton Faculty Achievement Award from MIT (1989). He is an Associate Editor of *Automatica* and *Applied Mathematics Letters* and has been an Associate Editor of the IEEE TRANSACTIONS ON AUTOMATIC CONTROL.



Ofer Zeitouni (SM'90) was born in Haifa, Israel, in 1960. He received the B.Sc. (summa cum laude), M.Sc., and Ph.D. degrees in electrical engineering in 1980, 1983, and 1986, respectively, all from the Technion-Israel Institute of Technology, Haifa, Israel.

During 1980–1985, he was with the Israel Defense Forces. He held Post-Doctural and visiting appointments at the Division of Applied Mathematics, Brown University (1986–1987), and at the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology (1987–1990). Since 1989, he has been with the Department of Electrical Engineering, Technion, where he is currently an Associate Professor. His research interests are in applied probability and stochastic processes with recent emphasis on large deviation techniques and asymptotic methods, stochastic partial differential equations, and probablistic methods for inference and learning.