Subsampling in Smoothed Range Spaces*

Jeff M. Phillips
University of Utah, SLC, USA
jeffp@cs.utah.edu

Yan Zheng
University of Utah, SLC, USA
yanzheng@cs.utah.edu

Abstract

We consider smoothed versions of geometric range spaces, so an element of the ground set (e.g. a point) can be contained in a range with a non-binary value in $[0, 1]$. Similar notions have been considered for kernels; we extend them to more general types of ranges. We then consider approximation of these range spaces through ε-nets and ε-samples (aka ε-approximations). We characterize when size bounds for ε-samples on kernels can be extended to these more general smoothed range spaces. We also describe new generalizations for ε-nets to these range spaces and show when results from binary range spaces can carry over to the smoothed ones.

1 Introduction

Combinatorial range spaces play a central role in geometry and have important connections to many areas, notably learning theory [6, 3], data structures, and recently differential privacy. We will focus on geometric range spaces where the ground set P is a point set in $\mathbb{R}^d$. The family of ranges $\mathcal{A}$ are typically defined by sets of subsets contained in some geometric objects, e.g., a disk, or a halfspace. The pair $(P, \mathcal{A})$ is called a *range space*.

An important consideration is how well we can approximate these objects through a subset $Q \subset P$, formalized as an ε-sample (aka ε-approximation, which preserves density) and an ε-net (which perverse the existence of large subsets). Formally, an *ε-sample* for a range space $(P, \mathcal{A})$ is a subset $Q \subset P$ s.t.

$$\max_{A \in \mathcal{A}} \left| \frac{|A \cap P|}{|P|} - \frac{|Q \cap A|}{|Q|} \right| \leq \varepsilon.$$

An *ε-net* of a range space $(P, \mathcal{A})$ is a subset $Q \subset P$ s.t. for all $A \in \mathcal{A}$ such that $\frac{|P \cap A|}{|P|} \geq \varepsilon$ then $A \cap Q \neq \emptyset$.

Through techniques ranging from discrepancy theory to Fourier analysis to basic combinatorics, we now largely understand these relationship of these bounds to the size of the subsets Q, for geometrically described ranges and with constructions; see a pair of great books [4, 1]. However, at least at a high-level, many of these size lower bounds are constructed with sets P so that problematic subsets $A \in \mathcal{A}$ have many elements near the boundary. This leads to the question, *what if we smoothed out this boundary?*

Background on Kernels and Kernel Range Spaces. This question was studied in the context of ε-samples for statistical kernels (e.g. Gaussians). A *kernel* is a bivariate similarity function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$, which can be normalized so $K(x, x) = 1$ (which we assume through this paper). We focus on symmetric, shift invariant kernels which depend only on $\|x - p\|$, and can be written as a single parameter function $K(x, p) = k(\|x - p\|)$, so it usually decreases as $\|x - p\|$ increases; these can be parameterized by a single bandwidth (or just width) parameter w so $K_w(x, p) = k_w(\|x-p\|) = k(\|x-p\|/w)$. Most commonly used kernels are Gaussian, Laplace, Triangular, Epanechnikov, and Ball kernels.

A *kernel range space* [2, 5] $(P, \mathcal{K})$ is an extension of the combinatorial concept of a range space $(P, \mathcal{A})$ (or to distinguish it we refer to the classic notion as a *binary range space*). It is defined by a point set $P \subset \mathbb{R}^d$ and a set of kernels $\mathcal{K}$. An element of $\mathcal{K}$ is a kernel $K(x, \cdot)$ applied at point $x \in \mathbb{R}^d$; it assigns a value in $[0, 1]$ to each point $p \in P$ as $K(x, p)$.

Given a point set P of size n and a kernel K, a *kernel density estimate* $\textsc{kde}_P$ is the convolution of that point set with K. For any $x \in \mathbb{R}^d$ we define $\textsc{kde}_P(x) = \frac{1}{n} \sum_{p \in P} K(x, p)$. The notion of ε-kernel sample [2] extends the definition of ε-sample. It is a subset $Q \subset P$ such that $\max_{x \in \mathbb{R}^d} |\textsc{kde}_P(x) - \textsc{kde}_Q(x)| \leq \varepsilon$.

A binary range space $(P, \mathcal{A})$ is *linked* to a kernel range space $(P, \mathcal{K})$ if the set $\{p \in P \mid K(x, p) \geq \tau\}$ is equal to $P \cap A$ for some $A \in \mathcal{A}$, for any threshold value τ.

Two main observations have been made in the kernel range spaces. (1) An ε-sample for a (linked) range space defined by balls, is also an ε-sample for kernels [2]. (2) Using a careful discrepancy-based approach, smaller ε-samples (sometimes significantly smaller) can be constructed for kernels than for balls [5]. In this article we extend this line of work in a few interesting directions.

Contributions.

- We define a general class of *smoothed range spaces*, with application to density estimation.
- We define a notion of an (ε, τ)-net for a smoothed range space. We show how this can inherit sam-

*Thanks to support by NSF CCF-1350888, IIS-1251019, and ACI-1443046.

pling complexity bounds from *linked* non-smooth range spaces. We also relate this concept to a smoothed hitting set problem.

- We provide discrepancy-based bounds and constructions for ε-samples on smooth range spaces requiring significantly fewer points than uniform sampling approaches and discrepancy-based approaches on the linked binary range spaces.

2 Smoothed Range Spaces

Let $\mathcal{H}_w$ denote the family of *smoothed halfspaces with width parameter* w, and let $(P, \mathcal{H}_w)$ be the associated smoothed range space where $P \subset \mathbb{R}^d$. Given a point $p \in P$, the smoothed halfspace $h \in \mathcal{H}_w$ maps p to a value $v_h(p) \in [0,1]$ (rather than the traditional $\{0,1\}$ in a binary range space).

We first describe a specific mapping to the function value $v_h(p)$. Let F be the $(d-1)$-flat defining the boundary of halfspace h. Given a point $p \in \mathbb{R}^d$, let $p_F = \arg\min_{q \in F} \|p - q\|$ describe a point on F closest to p. We make the definition more general using a shift-invariant kernel $k_w(\|p - x\|) = k(\|p - x\|/w)$ such that we define $v_{h,w}(p)$ as follows.

$$v_{h,w}(p) = \begin{cases} \frac{1}{2} + \frac{1}{2} k_w(\|p - p_F\|) & p \in h \\ \frac{1}{2} - \frac{1}{2} k_w(\|p - p_F\|) & p \notin h. \end{cases}$$

For brevity, we will omit the w and just use $v_h(p)$ when clear. We can also further generalize this by replacing the flat F at the boundary of h with a polynomial surface G. The point $p_G = \arg\min_{q \in G} \|p - q\|$ replaces p_F in the above definitions. Then the slab of width $2w$ is replaced with a more curved volume in $\mathbb{R}^d$; see Figure 1. For concreteness and simplicity, the remainder of this note will focus on halfspaces.

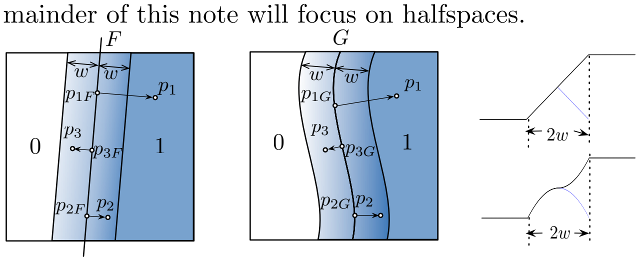

Figure 1: Illustration of the smoothed halfspace F (left), and smoothed polynomial surface G (middle).

We extend the notion of a kernel density estimate to these smoothed range spaces. A *smoothed density estimate* SDE_P is defined for any $h \in \mathcal{H}_w$ as

$$\mathrm{SDE}_P(h) = \frac{1}{|P|} \sum_{p \in P} v_h(p).$$

Then an ε-sample Q of a smoothed range space $(P, \mathcal{H}_w)$ is a subset $Q \subset P$ such that

$$\max_{h \in \mathcal{H}_w} |\mathrm{SDE}_P(h) - \mathrm{SDE}_Q(h)| \leq \varepsilon.$$

(ε, τ)-Net for smoothed range spaces. We introduce two new definitions to generalize the definition of hitting and ε-net. A subset $Q \subset P$ is an *(ε, τ)-net of smoothed range space* $(P, \mathcal{H}_w)$ if for any $h \in \mathcal{H}_w$ such that $\mathrm{SDE}_P(h) \geq \varepsilon$, there exists a point $q \in Q$ such that $v_h(q) \geq \tau$. A subset $Q \subset P$ is an *(ε, τ)-hitting set of smoothed range space* $(P, \mathcal{H}_w)$ if for any $h \in \mathcal{H}_w$ such that $\mathrm{SDE}_P(h) \geq \varepsilon$, then $\mathrm{SDE}_Q(h) \geq \tau$. We can show that both of these notions are implied by an $(\varepsilon - \tau)$-sample.

Theorem 1 *An $(\varepsilon - \tau)$-sample Q in a smoothed range space $(P, \mathcal{H}_w)$ is an (ε, τ)-hitting set in $(P, \mathcal{H}_w)$, and thus also an (ε, τ)-net of $(P, \mathcal{H}_w)$.*

Consider a smoothed range space $(P, \mathcal{H}_w)$, a linked binary range space $(P, \mathcal{A})$, and an ε-sample Q of $(P, \mathcal{A})$. Prior results for kernels [2] can be generalized to show Q is an ε-sample of $(P, \mathcal{H}_w)$. We can further extend this relation for (ε, τ)-nets; thus they can require significantly smaller size sets Q to satisfy.

Theorem 2 *Consider a smoothed range space $(P, \mathcal{H}_w)$, a linked binary range space $(P, \mathcal{A})$, and an $(\varepsilon - \tau)$-net Q of $(P, \mathcal{A})$. Then Q is an (ε, τ)-net of $(P, \mathcal{H}_w)$.*

Discrepancy-based approaches. We improve on random sample bounds using discrepancy [4, 1]. These results are restricted to when points P are contained in a d-dimensional cube $\mathsf{C}_{\ell,d}$ of side length ℓ.

Theorem 3 *In $\mathbb{R}^2$, for any $P \subset C_{\ell,2}$, we can construct an ε-sample of $(P, \mathcal{H}_w)$ of size $O(\frac{1}{\varepsilon}\sqrt{\frac{\ell}{w}\log\frac{\ell}{w\varepsilon\delta}})$ with probability at least $1 - \delta$.*

Theorem 4 *In $\mathbb{R}^d$, for any $P \subset \mathsf{C}_{\ell,d}$ with d is constant, we can construct an ε-sample of $(P, \mathcal{H}_w)$ of size $O\Big((\ell/w)^{2(d-1)/(d+2)} \cdot \Big(\frac{1}{\varepsilon}\sqrt{\log\frac{\ell}{w\varepsilon\delta}}\Big)^{2d/(d+2)}\Big)$ with probability at least $1 - \delta$,*

We can improve some results if the data is "well-clustered" under other specific conditions.

References

[1] B. Chazelle. *The Discrepancy Method*. Camb., 2000.

[2] S. Joshi, R. V. Kommaraju, J. M. Phillips, and S. Venkatasubramanian. Comparing distributions and shapes using the kernel distance. *SOCG*, 2011.

[3] Y. Li, P. M. Long, and A. Srinivasan. Improved bounds on the samples complexity of learning. *J. Comp. and Sys. Sci.*, 62:516–527, 2001.

[4] J. Matoušek. *Geometric Discrepancy*. Sprgr., 1999.

[5] J. M. Phillips. eps-samples for kernels. *SODA*, 2013.

[6] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Th of Prob., Apps.*, 16:264–280, 1971.