Math. Program., Ser. B (2019) 174:225–251
https://doi.org/10.1007/s10107-018-1257-5

FULL LENGTH PAPER

Statistics with set-valued functions: applications to inverse approximate optimization

Anil Aswani[1]

Received: 1 February 2017 / Accepted: 2 March 2018 / Published online: 7 March 2018

Abstract Much of statistics relies upon four key elements: a law of large numbers, a calculus to operationalize stochastic convergence, a central limit theorem, and a framework for constructing local approximations. These elements are well-understood for objects in a vector space (e.g., points or functions); however, much statistical theory does not directly translate to sets because they do not form a vector space. Building on probability theory for random sets, this paper uses variational analysis to develop operational tools for statistics with set-valued functions. These tools are first applied to nonparametric estimation (kernel regression of set-valued functions). The second application is to the problem of inverse approximate optimization, in which approximate solutions (corrupted by noise) to an optimization problem are observed and then used to estimate the amount of suboptimality of the solutions and the parameters of the optimization problem that generated the solutions. We show that previous approaches to this problem are statistically inconsistent when the data is corrupted by noise, whereas our approach is consistent under mild conditions.

Keywords Set-valued functions · Statistics · Inverse optimization

Mathematics Subject Classification 62F12 · 62G08 · 49J53

This work was supported in part by NSF Award CMMI-1450963.

✉ Anil Aswani
aaswani@berkeley.edu

[1] Industrial Engineering and Operations Research, University of California, Berkeley, CA, USA

1 Introduction

While statistical theory is well-developed for problems concerning (single-valued) functions [13,46], there has been less work on statistics with sets or set-valued functions. Most attention in statistics on sets has been focused on the problem of estimating a single set under different measurement models [18,21,23,26,36,37,44]. The problem of estimating set-valued functions is less well studied, though it has potential applications in varied domains including healthcare, robotics, and energy. For instance, we study in this paper the problem of inverse approximate optimization, where approximate solutions (corrupted by noise) to a parametric optimization problem are observed and then used to estimate the amount of suboptimality of the solutions and the parameters that generated the solutions. Inverse approximate optimization can be used to construct predictive models of human behavior and decision-making, where the explicit model is that an individual makes decisions by approximately solving an optimization problem. Statistical estimation in this context could be used to quantify the tradeoffs made by a particular individual between competing objectives, as well as quantify the predictability of the decision-making process. This particular problem of inverse approximate optimization is related to the broader topic of statistics with set-valued functions because the solution mapping of an (even strictly convex) optimization problem becomes a set when suboptimality of solutions is allowed. Thus a framework for statistics with set-valued functions is needed to study such problems.

A substantial impediment to studying such estimation problems is the lack of statistical tools for random sets and set-valued functions, and two technical issues prevent the use of existing tools. The first is that most statistical theory assumes objects belong to a vector space, which is the case for points and functions. But sets do not form a vector space, and so existing statistical theory cannot be used. This is a fundamental difficulty, and even the usual notion of expectation does not apply to sets [31]. The second is that most statistical theory has been developed by using metrics and distance functions to derive results. But analyzing sets using distances is difficult, and most analysis tools and results for sets do not use this approach [10,38].

Arguably the most natural approach to statistics with random sets is to define a family of sets parametrized by a random vector, and then perform standard statistical analysis with respect to this parametrization. However it is not clear without further analysis whether stochastic convergence of the estimated parameters implies stochastic convergence of the corresponding set estimates. We study this question in a more general framework and give a counterexample to demonstrate how parameter convergence does not always imply set convergence. Moreover, the parametrization approach does not lead to a useful definition for the expectation of random sets [31]; the reason is that the expectation of the parameters does not characterize the expectation of the set in a way in that ensures the law of large numbers holds.

One goal of this paper is to establish tools for statistics with set-valued functions, and this requires understanding four main ingredients: a law of large numbers, a calculus to operationalize stochastic convergence, a central limit theorem, and tools for constructing local approximations. Probability theory for random sets [31] provides an expectation for random sets [8,27], a law of large numbers [3], and a central limit theorem [49]. Here we use variational analysis [38] to advocate a notion of local approximation for set-valued functions, and to develop results that allow us to interpret stochastic convergence and expectations of random sets as operators.

The paper begins by describing our notation and providing some useful definitions related to set-valued functions. We focus in this paper on almost sure (a.s.) convergence because the corresponding definitions and approach most clearly demonstrate the tight link between variational analysis and statistics. Defining set convergence in probability requires metrization, which partially obscures the relationship to variational analysis. We also focus on Lipschitz continuity for set-valued functions because we advocate using this concept as a notion of local approximation for set-valued functions. The utility of this approach is displayed later in the paper when we use Lipschitz continuity as a replacement for differentiability when proving a Delta method-like result and proving statistical consistency of a kernel regression estimator.

The next section shows how to interpret stochastic convergence and expectation of random sets as operators. We study the limit of sequences of sets under different set operations, after proving a set-based generalization of the continuous mapping theorem [13] from statistics. Then we study the expectation of random sets under various set operations. Standard proofs about the properties of the expectation of random variables do not extend because the expectation of a random set cannot be computed by integration. This means properties like distribution of expectation under independence of the product of a random matrix with a random set or Jensen's inequality have not been previously established, and we prove such results. We conclude by reviewing a law of large numbers and a central limit theorem for random sets.

Another goal of this paper is to study two problems of estimating set-valued functions, and through the process of analyzing these problems we demonstrate the utility of our tools for statistics with set-valued functions. The first problem we study is estimating a set-valued function using noisy measurements of the set. We propose a kernel regression estimator that can be interpreted as a generalization of methods for functions [4,5,12,33,48]. The key step in proving statistical consistency is using Lipschitz continuity of the set-valued function to construct local approximations. We show that statistical consistency follows by combining our results on stochastic convergence with convergence bounds on (vector-valued) random variables.

The second problem we study is inverse approximate optimization, where noisy measurements of approximate solutions to an optimization problem are used to estimate the suboptimality of the solutions and the parameters of the optimization problem. In contrast, past work on inverse optimization assumes no noise [1,16,20] or exact solutions [7,11,24]. We develop a method for inverse approximate optimization and prove its statistical consistency using stochastic epi-convergence [4,19,22,25,28,42]. Combining with our results on stochastic convergence and results on the continuity of solutions to optimization problems [38,39] shows our method consistently estimates the (set-valued) approximate solution mapping that generates the data.

We conclude by examining extensions of the problem of inverse approximate optimization, as well as discussing related open questions about statistics with set-valued functions. In particular, we describe how some extensions lead to formulations of optimization problems with structures (e.g., objective functions that are integrals whose domain of integration depends on the decision variable) that have not been well-studied from the perspective of numerical optimization. Performing statistics with sets and set-valued functions also leads to questions about the design of numerical representations of sets. We argue that further study of statistics with set-valued functions will require developing new numerical methods and optimization theory.

2 Preliminaries

This section presents the notation used in this paper, as well as several useful concepts from variational analysis. Most of the variational analysis definitions are from [38]. The definition of set-valued set functions is from [30], and we use the definitions of the Minkowski set operations from [43]. We abbreviate *almost surely* using a.s.

2.1 Notation

Let $\mathcal{F}(E)$ be the space of closed subsets of E, and let $\mathcal{K}(E)$ be the space of compact subsets of E. We will focus on cases where E is a Euclidean space, and so will use the notation $\mathcal{F}, \mathcal{K}$ to refer to the corresponding spaces. Clearly $\mathcal{F} \supset \mathcal{K}$ by definition.

Suppose C, D are sets and Ψ is a matrix or scalar. We use the set notation: $C \cup D$ is the union of C, D; $C \cap D$ is the intersection of C, D; $C \subseteq D$ denotes that C is a subset of D; $C \supseteq D$ denotes that C is a superset of D; $\mathrm{cl}(C)$ is the closure of C; $\mathrm{co}(C)$ is the convex hull of C; C^{c} is the complement of C; ∂C is the boundary of C; $C \oplus D = \{c + d : c \in C, d \in D\}$ is the Minkowski sum of C, D; $C \ominus D = \{x : x \oplus D \subseteq C\}$ is the Minkowski difference of C, D; $\Psi \cdot C = \{\Psi \cdot c : c \in C\}$; and $\Psi^{-1}C = \{\Psi^{-1} \cdot c : c \in C\}$.

2.2 Limit definitions and set-valued mappings

The outer limit of the sequence of sets C_n is defined as

$$\limsup_n C_n = \{x : \exists n_k \text{ s.t. } x_{n_k} \to x \text{ with } x_{n_k} \in C_{n_k}\}, \tag{1}$$

and the inner limit of the sequence of sets C_n is defined as

$$\liminf_n C_n = \{x : \exists x_n \to x \text{ with } x_n \in C_n\}. \tag{2}$$

The outer limit consists of all the cluster points of C_n, whereas the inner limit consists of all limit points of C_n. The limit of the sequence of sets C_n exists if the outer and inner limits are equal, and we define that $\lim_n C_n := \limsup_n C_n = \liminf_n C_n$.

Let $\overline{\mathbb{R}} = [-\infty, \infty]$ denote the extended real line. A sequence of extended-real-valued functions $f_n : X \to \overline{\mathbb{R}}$ is said to epi-converge to f if at each $x \in X$ we have

$$\begin{cases} \liminf_n f_n(x_n) \geq f(x) & \text{for every sequence } x_n \to x \\ \limsup_n f_n(x_n) \leq f(x) & \text{for some sequence } x_n \to x \end{cases} \tag{3}$$

The notion of epi-convergence is so-named because it is equivalent to set convergence of the epigraphs of f_n, meaning that epi-convergence is equivalent to the condition $\lim_n\{(x,\alpha)\in X\times\mathbb{R}: f_n(x)\le\alpha\}=\{(x,\alpha)\in X\times\mathbb{R}: f(x)\le\alpha\}$.

A set-valued set function $G: V\Rightarrow U$ assigns to each set $S\subseteq V$ a set $G(S)\subseteq U$. The outer limit of G at the set $\overline{S}\in V$ is defined as

$$\limsup_{S\to\overline{S}} G(S)=\{u:\exists S_n\to\overline{S}\text{ s.t. } u_n\to u \text{ with } S_n\subseteq V, u_n\in G(S_n)\}, \tag{4}$$

and the inner limit of G at the set $\overline{S}\subseteq V$ is defined as

$$\liminf_{S\to\overline{S}} G(S)=\{u:\forall S_n\to\overline{S},\ \exists u_n\to u \text{ with } S_n\subseteq V, u_n\in G(S_n)\}. \tag{5}$$

The intuition is similar to the notions for sequences of sets. The set-valued set function G is outer semicontinuous (osc) at $\overline{S}$ if $\limsup_{S\to\overline{S}} G(S)\subseteq G(\overline{S})$, and G is inner semicontinuous (isc) at $\overline{S}$ if $\liminf_{S\to\overline{S}} G(S)\supseteq G(\overline{S})$. The set-valued set function G is continuous at $\overline{S}$ when it is both osc and isc, that is when $\lim_{S\to\overline{S}} G(S)=G(\overline{S})$.

Variational analysis typically uses set-valued functions, rather than set-valued set functions. A set-valued function $F: X\rightrightarrows U$ assigns to each point $x\in X$ a set $F(x)\subseteq U$. Outer limits, inner limits, outer semicontinuity, inner semicontinuity, and continuity are defined as above but with points replacing sets in the domain. Moreover, a set-valued function applied pointwise to sets is an osc, isc, continuous set-valued set function whenever the set-valued function is osc, isc, continuous, respectively.

2.3 Probability definitions and stochastic convergence

Let $(\Omega,\mathfrak{F},\mathbb{P})$ be a complete probability space, where Ω is the sample space, $\mathfrak{F}$ is the set of events, and $\mathbb{P}$ is the probability measure. A map $S:\Omega\to\mathcal{F}$ is a random set if $\{\omega: S(\omega)\in\mathcal{X}\}\in\mathfrak{F}$ for each $\mathcal{X}$ in the Borel σ-algebra on $\mathcal{F}$ [31]. Like the usual convention for random variables, we notationally drop the argument for a random set.

When discussing samples for estimation, we use the convention that capital letters denote random variables, and lowercase letters denote measured data. Also, we use the notation $U(a,b)$ to specify a uniform distribution with support $[a,b]$.

We next define almost sure stochastic convergence of random sets. The notation as-$\limsup_n C_n\subseteq C$ denotes $\mathbb{P}(\limsup_n C_n\subseteq C)=1$, the notation as-$\liminf_n C_n\supseteq C$ denotes $\mathbb{P}(\liminf_n C_n\supseteq C)=1$, and the notation as-$\lim_n C_n=C$ denotes $\mathbb{P}(\lim_n C_n=C)=1$. Note as-$\limsup_n C_n\subseteq C$ and as-$\liminf_n C_n\supseteq C$ if and only if as-$\lim_n C_n=C$, since a countable intersection of almost sure events occurs almost surely.

2.4 Distances and Lipschitz continuity

Let $d(x,C)=\inf_{y\in C}\|x-y\|$ and $d^2(x,C)=\inf_{y\in C}\|x-y\|^2$ be the distance and squared distance, respectively, from a point x to set C. The support function of C is

$h(x, C) = \sup_{y \in C} x^{\mathsf{T}} y$. We also define the indicator function $\delta(x, C)$ to equal 0 when $x \in C$ and $+\infty$ when $x \notin C$. The (integrated) set distance between C and D is defined as $d\!l(C, D) = \int_0^\infty d\!l_r(C, D)e^{-r}dr$, where the pseudo-distance between sets C and D is given by $d\!l_r(C, D) = \max_{\|x\| \leq r} \left|d(x, C) - d(x, D)\right|$. Note $d\!l_r(\{x\}, C) \neq d(x, C)$ for all r. The integrated set distance $d\!l$ is a metric that characterizes the convergence defined earlier for sets in $\mathcal{F}$, and the Pompeiu–Hausdorff distance $d\!l_\infty$ is a metric that characterizes the convergence defined earlier for sets in $\mathcal{K}$. Since these metrics are complex, the sequence characterization of convergence is arguably more natural for sets.

One exception to this statement is in defining Lipschitz continuity for set-valued functions. A set-valued function $F : X \rightrightarrows U$ is Lipschitz continuous on X with constant $\kappa \in \mathbb{R}_+$ if it is nonempty, closed-valued and such that

$$F(x') \subseteq F(x) + \kappa\|x' - x\|\mathbb{B} \;\text{ for } x, x' \in X, \tag{6}$$

where $\mathbb{B} = \{u : \|u\| \leq 1\}$ is the unit ball. A set-valued set function $G : V \Rightarrow U$ is Lipschitz continuous on V with constant $\kappa \in \mathbb{R}_+$ if it is nonempty, closed-valued and

$$G(S') \subseteq G(S) + \kappa d\!l_\infty(S', S)\mathbb{B} \;\text{ for } S, S' \subseteq V \text{ with } S, S' \in \mathcal{K}. \tag{7}$$

We will make use of Lipschitz continuity as a zeroth-order local approximation.

3 Mathematical tools for statistics with set-valued functions

This section develops mathematical tools that allow us to interpret stochastic convergence and the expectation of random sets as operators. We prove results on the limit of sequences of sets under different set operations, define an expectation for random sets, and then derive results about the behavior of this expectation under different set operations. We conclude this section by briefly summarizing a law of large numbers and a central limit theorem for random sets.

3.1 Stochastic limit theorems

Our reason for considering set-valued set functions is this allows us to more precisely generalize the classical *continuous mapping theorem* of statistics [13] to mappings applied to sequences of sets. Because semicontinuity is an important aspect of set convergence, a generalization that considers semicontinuity leads to a richer set of results than simply considering continuity.

Theorem 1 (Semicontinuous mapping theorem) *Let G be a set-valued set function, and suppose* as-$\lim_n C_n = C$. *There are three cases:*

(a) *If G is osc at C, then* as-$\limsup_n G(C_n) \subseteq G(C)$.
(b) *If G is isc at C, then* as-$\liminf_n G(C_n) \supseteq G(C)$.
(c) *If G is continuous at C, then* as-$\lim_n G(C_n) = G(C)$.

Proof The definition of osc (isc) means $\lim_n C_n = C$ implies $\limsup_n G(C_n) \subseteq G(C)$ ($\liminf_n G(C_n) \subseteq G(C)$). This means $\mathbb{P}(\limsup_n G(C_n) \subseteq G(C)) \geq \mathbb{P}(\lim_n C_n = C) = 1$ ($\mathbb{P}(\liminf_n G(C_n) \supseteq G(C)) \geq \mathbb{P}(\lim_n C_n = C) = 1$), which shows the first two cases. The third case follows from the first two cases by recalling that continuity at C is equivalent to being both osc and isc at C. □

Remark 1 One consequence is that the set-valued function $S(\theta)$ parametrized by θ has the behavior that as-$\lim_n \theta_n = \theta_0$ implies as-$\lim_n S(\theta_n) = S(\theta_0)$ only when the set is continuous with respect to the parametrization. For example, consider $S(\theta) = \{1\}$ if $\theta > 0$, $S(\theta) = \{-1\}$ if $\theta < 0$, and $S(\theta) = [-1, 1]$ if $\theta = 0$. If $\theta_n = 1/n$, then $S(\theta_n) \equiv \{1\}$ and so as-$\lim_n S(\theta_n) = \{1\}$. But as-$\lim_n \theta_n = 0$ and $S(0) = [-1, 1]$.

As is customary in statistics, we immediately get some useful corollaries to our semicontinuous mapping theorem by applying the theorem to specific mappings. Our first corollary applies the semicontinuous mapping theorem to set operations like unions and intersections of sets, the boundary of sets, the convex hull of sets, etc.

Corollary 1 *Let $C_n, D_n \in \mathcal{F}$ be almost surely convergent sequences of sets (i.e.,* as-$\lim_n C_n = C$ *and* as-$\lim_n D_n = D$*). Then we have:*

(a) as-$\lim_n (C_n \cup D_n) = C \cup D$
(b) as-$\limsup_n (C_n \cap D_n) \subseteq C \cap D$
(c) as-$\liminf_n \mathrm{cl}(C_n^{\mathsf{c}}) \supseteq \mathrm{cl}(C^{\mathsf{c}})$
(d) as-$\liminf_n \partial C_n \supseteq \partial C$
(e) as-$\liminf_n \mathrm{co}(C_n) \supseteq \mathrm{co}(C)$
(f) as-$\lim_n \mathrm{co}(C_n) = \mathrm{co}(C)$*, when there is a deterministic $C_0 \in \mathcal{K}$ so $C_n \subseteq C_0$ a.s.*

Proof We interpret $\cup, \cap$ as set-valued set functions with a domain over the product space $\mathcal{F} \times \mathcal{F}$: The function $G_1(S, T) = S \cup T$ is continuous [30], and the function $G_2(S, T) = S \cap T$ is osc [30]. The set complement and boundary operators can be interpreted as set-valued set functions with domain $\mathcal{F}$: The function $G_3(S) = \mathrm{cl}(S^{\mathsf{c}})$ is isc [30], and the function $G_4(S) = \partial S$ is isc [30]. The convex hull operation can be cast as set-valued set functions: $G_5(S) = \mathrm{co}(S)$ is isc when the domain is $\mathcal{F}$, and $G_6(S) = \mathrm{co}(S)$ is continuous when the domain is C_0 [30]. The results now follow from the corresponding parts of the semicontinuous mapping theorem. □

Remark 2 Note the above result states that the stochastic limit of the convex hull operator is sensitive to the domain of the sequence of sets.

We can also apply the semicontinuous mapping theorem to the Minkowski set operations. These results are useful for proving convergence of statistical estimators.

Corollary 2 *Let $C_n, D_n \in \mathcal{F}$ be almost surely convergent sequences of sets (i.e.,* as-$\lim_n C_n = C$ *and* as-$\lim_n D_n = D$*), and let Ψ_n be an almost surely convergent (in the Frobenius norm) sequence of matrices or scalars (i.e.,* as-$\lim_n \Psi_n = \Psi$*). If there exists a deterministic $D_0 \in \mathcal{K}$ so $D_n \subseteq D_0$ a.s., then*

(a) as-$\lim_n (C_n \oplus D_n) = C \oplus D$
(b) as-$\limsup_n (C_n \ominus D_n) \subseteq C \ominus D$*, when $D \neq \emptyset$*

(c) $\text{as-lim}_n\, \Psi_n \cdot D_n = \Psi \cdot D$
(d) $\text{as-lim}_n\, \Psi_n^{-1} D_n^{-1} = \Psi_n^{-1} D$*, when* Ψ *is invertible*

Proof We interpret $\oplus, \ominus$ as set-valued set functions with a domain over the product space $\mathcal{F} \times \mathcal{K}$: The function $G_1(S, T) = S \oplus T$ is continuous [30], and the function $G_2(S, T) = S \ominus T$ is osc if $T \neq \emptyset$ [30]. So the first two results follow from Theorem 1. The multiplication operation can be interpreted as a set-valued set function $G_3(S, T) = T \cdot S$ with domain over the product space $\mathbb{M} \times D_0$, where $\mathbb{M}$ is the space of matrices of appropriate dimension or the space of scalars. We show it is continuous. Suppose G_3 is not osc at $\overline{S} \times \overline{T}$; then there exist $T_n \to \overline{T}$, $S_n \to \overline{S}$, and $u_n \to \overline{u}$ with $T_n \in \mathbb{M}$, $S_n \in \mathcal{F}$, $u_n \in T_n \cdot S_n$, and $\overline{u} \notin \overline{T} \cdot \overline{S}$. But by the definition of matrix-set (or scalar-set) multiplication there exists $v_n \in S_n$ with $u_n = T_n \cdot v_n$, and by the boundedness by assumption of D_0 there exist n_k and $\overline{v}$ such that $v_{n_k} \to \overline{v}$ with $\overline{v} \in \overline{S}$, which is a contradiction since matrix-vector (or scalar-vector) multiplication is osc. Thus G_3 is osc. Next, we show G_3 is isc at $\overline{T} \cdot \overline{S}$: Consider any $\overline{x} \in \overline{S}$ and $u = \overline{T} \cdot \overline{x}$, and let T_n, S_n be any sequences satisfying $T_n \to \overline{T}$ and $S_n \to \overline{S}$. By the inner limit definition there exists $x_n \to \overline{x}$ with $x_n \in S_n$, and so $T_n \cdot x_n \to \overline{T} \cdot \overline{x}$ with $T_n \cdot x_n \in T_n \cdot S_n$. So G_3 satisfies the definition of being isc at $\overline{T} \cdot \overline{S}$, and is continuous since it is also osc. The third result follows from Theorem 1. The fourth result is proved by noting Theorem 1 implies $\text{as-lim}_n\, \Psi_n^{-1} = \Psi_n^{-1}$ since the matrix inverse operation is continuous except at points of singularity, and so $\text{as-lim}_n\, \Psi_n^{-1} C_n^{-1} = \Psi_n^{-1} C$ by the third result. □

Our final results on stochastic limits are not based on the semicontinuous mapping theorem, but are nevertheless useful for writing stochastic convergence proofs.

Lemma 1 (Sandwich lemma) *Let* $L_n \in \mathcal{F}$ *and* $U_n \in \mathcal{F}$ *be almost surely convergent sequences of sets (i.e.,* $\text{as-lim}_n\, L_n = L$ *and* $\text{as-lim}_n\, U_n = U$*), and let* $C_n \in \mathcal{F}$ *be a sequence of sets. Then we have*

(a) $\text{as-lim sup}_n\, C_n \subseteq U$*, when* $C_n \subseteq U_n$ *a.s.*
(b) $\text{as-lim inf}_n\, C_n \supseteq L$*, when* $C_n \supseteq L_n$ *a.s.*
(c) $\text{as-lim}_n\, C_n = L = U$*, when* $L_n \subseteq C_n \subseteq U_n$ *a.s. and* $L = U$

Proof For the first two results, note $\text{as-lim sup}_n\, C_n \subseteq \text{as-lim sup}_n\, U_n = \text{as-lim}_n\, U_n = U$ and $\text{as-lim inf}_n\, C_n \supseteq \text{as-lim inf}_n\, L_n = \text{as-lim}_n\, L_n = L$. The third result follows from the first two results and the definition of limit. □

This sandwich lemma is valuable for statistical analysis, and we next present a convergence result that is helpful in proving statistical consistency.

Corollary 3 *Let* $C_n, D_n \in \mathcal{F}$ *be sequences of sets, with* $D_n \subseteq r_n \mathbb{B}$ *for a sequence* $r_n \in \mathbb{R}_+$*. If* $\text{as-lim}_n\, r_n = 0$ *and* $\text{as-lim}_n\, C_n \oplus D_n$ *exists, then* $\text{as-lim}_n\, C_n = \text{as-lim}_n\, C_n \oplus D_n$*.*

Proof Consider any $\overline{c} \in \text{as-lim sup}_n\, C_n$, and note that by the outer limit definition there exist n_k and $c_{n_k} \in C_{n_k}$ such that $c_{n_k} \to \overline{c}$. Thus $c_{n_k} + d_{n_k} \to \overline{c}$ for any $d_n \in D_n$ since by assumption $d_n \to 0$. This means $\text{as-lim sup}_n\, C_n \subseteq \text{as-lim sup}_n\, C_n \oplus D_n = \text{as-lim}_n\, C_n \oplus D_n$, where the equality holds since $\text{as-lim}_n\, C_n \oplus D_n$ exists. Next, choose any $\overline{u} \in$ as-$\lim_n C_n \oplus D_n$. By the inner limit definition there exists $u_n \in C_n \oplus D_n$ such that $u_n \to \overline{u}$, and so by the Minkowski sum definition there exist $c_n \in C_n$ and $d_n \in D_n$ such that $u_n = c_n + d_n$ or equivalently that $c_n = u_n - d_n$. Since by assumption $d_n \to 0$, this means $c_n \to \overline{u}$. Thus as-$\lim\inf_n C_n \supseteq$ as-$\lim_n C_n \oplus D_n$. The result follows by noting as-$\lim\inf_n C_n \subseteq$ as-$\lim\sup_n C_n$ always holds, and combining with the above. □

3.2 Expectation

Because sets do not form a vector space, defining expectations for random sets is not straightforward. In fact, a number of different definitions have been proposed [31] that capture different features that might be desired for an expectation operation. One particularly useful definition is the *selection expectation* [8,27]. This definition for the expectation of random sets is the most well studied because it leads to a corresponding law of large numbers and central limit theorem [31].

For a random set X, a selection ξ is a (single-valued) random vector that almost surely belongs to X. We say the selection ξ is integrable if $\mathbb{E}\|\xi\|_1$ is finite, where $\|\cdot\|_1$ is the usual ℓ_1-norm. The selection expectation of a random set X is defined as

$$\mathbb{E}(X) = \mathrm{cl}\{\mathbb{E}\xi : \xi \in \mathcal{S}^1(X)\}, \tag{8}$$

where $\mathcal{S}^1(X)$ is the set of all integrable selections of X. The random set X is called integrable if $\mathcal{S}^1(X) \neq \emptyset$, and note this property implies X is almost surely non-empty.

The selection expectation is difficult to use because it cannot be computed by taking an integral, as is the case for expectations for objects in a vector space. But since we assume E is Euclidean space, the definition of the selection expectation simplifies and has a sharp characterization [31]: If the probability space is nonatomic and X is a bounded and closed integrable random set, then $\mathbb{E}(X) = \{\mathbb{E}\xi : \xi \in \mathcal{S}^1(X)\}$ is a compact set, $\mathbb{E}(X)$ is convex, $\mathbb{E}(X) = \mathbb{E}(\mathrm{co}(X))$, and $h(u, \mathbb{E}(X)) = \mathbb{E}(h(u, X))$ for all $u \in E$, where h is the support function. This support function characterization is powerful, and allows us to prove several properties about the selection expectation. More importantly, the following results allow us to operationalize the selection expectation, which is useful from a practical standpoint for performing statistical analysis.

Proposition 1 *Suppose C, D are bounded and closed integrable random sets, and let Ψ be a random matrix or a random scalar. If the probability space is nonatomic, then*

(a) $\mathbb{E}(C) = \mathrm{co}(C)$, *when C is deterministic*
(b) $\mathbb{E}(C \oplus D) = \mathbb{E}(C) \oplus \mathbb{E}(D)$
(c) $\mathbb{E}(\Psi C) = \mathbb{E}(\Psi) \cdot \mathbb{E}(C)$, *when Ψ is independent of C*
(d) $\mathbb{E}(C) \subseteq \mathbb{E}(D)$, *when $C \subseteq D$ a.s.*
(e) $\mathbb{E}(C) \cup \mathbb{E}(D) \subseteq \mathbb{E}(C \cup D)$
(f) $\mathbb{E}(C \cap D) \subseteq \mathbb{E}(C) \cap \mathbb{E}(D)$
(g) $\mathbb{E}(C \ominus D) \subseteq \mathbb{E}(C) \ominus \mathbb{E}(D)$, *when $C \ominus D$ is a.s. non-empty.*

Proof The first result holds since $\mathbb{E}(X) = \mathbb{E}(\mathrm{co}(X))$ and $h(u, \mathbb{E}(C)) = \mathbb{E}(h(u, C)) = h(u, C)$. The next result follows from $h(u, C \oplus D) = h(u, C) + h(u, D)$ [43], since

$h(u, \mathbb{E}(C \oplus D)) = \mathbb{E}(h(u, C \oplus D)) = \mathbb{E}(h(u, C) + h(u, D)) = \mathbb{E}(h(u, C)) + \mathbb{E}(h(u, D)) = h(u, \mathbb{E}(C)) + h(u, \mathbb{E}(D)) = h(u, \mathbb{E}(C) \oplus \mathbb{E}(D))$. The fourth result holds since $h(u, C) \leq h(u, D)$ when $C \subseteq D$ [43], which implies $h(u, \mathbb{E}(C)) = \mathbb{E}(h(u, C)) \leq \mathbb{E}(h(u, D)) = h(u, \mathbb{E}(D))$. For the fifth result, note $C \subseteq C \cup D$ and $D \subseteq C \cup D$. The fourth result gives $\mathbb{E}(C) \subseteq \mathbb{E}(C \cup D)$ and $\mathbb{E}(D) \subseteq \mathbb{E}(C \cup D)$, which implies $\mathbb{E}(C) \cup \mathbb{E}(D) \subseteq \mathbb{E}(C \cup D)$. The sixth result follows since combining $C \cap D \subseteq C$, $C \cap D \subseteq D$, and the fourth result gives: $\mathbb{E}(C \cap D) \subseteq \mathbb{E}(C)$ and $\mathbb{E}(C \cap D) \subseteq \mathbb{E}(D)$, which implies $\mathbb{E}(C \cap D) \subseteq \mathbb{E}(C) \cap \mathbb{E}(D)$. To prove the seventh result, note $(C \ominus D) \oplus D \subseteq C$ [43]. Applying the second and fourth results yields $\mathbb{E}(C \ominus D) \oplus \mathbb{E}(D) \subseteq \mathbb{E}(C)$, and so $\mathbb{E}(C \ominus D) \subseteq \mathbb{E}(C) \ominus \mathbb{E}(D)$ [43].

The third result cannot be proved using support functions since $h(x, \Psi C)$ cannot be written in terms of $h(x, C)$. (If $\Psi = -1$, then $h(x, \Psi C) = \inf_{y \in C} x^\mathsf{T} y$ while $h(x, C) = \sup_{y \in C} x^\mathsf{T} y$.) Our approach is to show $\mathcal{S}^1(\Psi C) = \Psi \mathcal{S}^1(C)$, since this implies $\mathbb{E}(\Psi C) = \{\mathbb{E}(\Psi \xi) : \xi \in \mathcal{S}^1(C)\} = \{\mathbb{E}(\Psi) \cdot \mathbb{E}(\xi) : \xi \in \mathcal{S}^1(C)\} = \mathbb{E}(\Psi) \cdot \{\mathbb{E}(\xi) : \xi \in \mathcal{S}^1(C)\} = \mathbb{E}(\Psi) \cdot \mathbb{E}(C)$. The inclusion $\mathcal{S}^1(\Psi C) \supseteq \Psi \mathcal{S}^1(C)$ is obvious by definition. To prove the reverse inclusion, let $\{\xi_n, n \geq 1\}$ with $\xi_n \in \mathcal{S}^1(C)$ be the Castaing representation [15,31,38] of C. Then $\{\Psi \xi_n, n \geq 1\}$ is the Castaing representation of ΨC. But by Lemma 1.3 of [31], each selection in $\mathcal{S}^1(\Psi C)$ can be approximated arbitrarily well by step functions with arguments from $\{\Psi \xi_n, n \geq 1\}$. Thus $\mathcal{S}^1(\Psi C) \subseteq \Psi \mathcal{S}^1(C)$, and so $\mathcal{S}^1(\Psi C) = \Psi \mathcal{S}^1(C)$ since both inclusions were shown. □

Remark 3 Note the assumptions for part (c) include the cases where: Ψ is deterministic, C is deterministic, or Ψ has positive or negative entries.

Another result used in statistics is Jensen's inequality [13], which bounds changing the order of applying an expectation and a convex function to a random variable. Our next result shows we can generalize Jensen's inequality to set-valued functions.

Proposition 2 *(Jensen's inequality) Let $S(u)$ be a graph-convex set-valued function (i.e., $S((1 - \lambda)u_0 + \lambda u_1) \supseteq (1 - \lambda) \cdot S(u_0) + \lambda \cdot S(u_1)$ for $\lambda \in (0, 1)$), and let X be bounded and closed integrable random set. If $S(\cdot)$ is locally bounded (i.e., $S(B)$ is bounded for every bounded set B) and continuous, then we have $S(\mathbb{E}(X)) \supseteq \mathbb{E}(S(X))$.*

Proof The selection expectation equals the Debreu expectation under our assumptions [31]. This means there exists a sequence of random sets X_n with the distribution

$$X_n = F_{in} \text{ with probability } p_{in}, \quad \text{for } i = 1, \dots, n, \quad \text{with } \sum_{i=1}^{n} p_{in} = 1 \tag{9}$$

such that as-lim $X_n = X$, $\mathbb{E}(X) = \lim_n \mathbb{E}(X_n)$, and $\mathbb{E}(X_n) = \bigoplus_{i=1}^n p_{in} \cdot F_{in}$. Using the semicontinuous mapping theorem implies as-$\lim_n S(X_n) = S(X)$, and so we have equality of the selection expectation and Debreu expectation [31]. This means that $\mathbb{E}(S(X)) = \lim_n \mathbb{E}(S(X_n))$ and $\mathbb{E}(S(X_n)) = \bigoplus_{i=1}^n p_{in} \cdot S(F_{in})$. Next note $S(\bigoplus_{i=1}^n p_{in} \cdot F_{in}) \supseteq \bigoplus_{i=1}^n p_{in} \cdot S(F_{in})$ by the graph-convexity of $S(\cdot)$. Taking the limit of this set relationship gives $S(\mathbb{E}(X)) = \lim S(\bigoplus_{i=1}^n p_{in} \cdot F_{in}) \supseteq \lim_n \bigoplus_{i=1}^n p_{in} \cdot S(F_{in}) = \mathbb{E}(S(X))$, where we have used the fact that $\lim_n S(\bigoplus_{i=1}^n p_{in} \cdot F_{in}) = S(\mathbb{E}(X))$ by definition of the continuity of the set-valued function $S(\cdot)$. □

Remark 4 Jensen's inequality is sometimes stated for concave functions, and such a generalization exists for set-valued mappings. If $S(u)$ is a graph-concave set-valued function (i.e., $S((1-\lambda)u_0 + \lambda u_1) \subseteq (1-\lambda)\cdot S(u_0) + \lambda \cdot S(u_1)$ for $\lambda \in (0, 1)$)) and the other assumptions of the above theorem hold, then we have $S(\mathbb{E}(X)) \subseteq \mathbb{E}(S(X))$.

Lastly, we present a strong law of large numbers (SLLN) for the selection expectation. The key idea is the Minkowski sum takes the role of averaging.

Theorem 2 (Artstein and Vitale [3]) *Suppose the probability space is non-atomic. If* $X, X_i, i \geq 1$*, are i.i.d. bounded and closed integrable random sets, then we have that:* as-$\lim_n \frac{1}{n}\bigoplus_{i=1}^n X_i = \mathbb{E}(X)$.

This particular strong law of large numbers can be generalized in a number of ways, and a survey of the different generalizations possible can be found in [31].

3.3 Central limit theorems

Unlike laws of large numbers that relate convergence of Minkowski sums of i.i.d. random sets $\frac{1}{n}\bigoplus_{i=1}^n X_i$ to their selection expectation $\mathbb{E}(X)$, analogs of the central limit theorem (CLT) relating Minkowski sums and selection expectations are less well-understood. One major impediment is that the $\ominus$ operator does not generally invert the $\oplus$ operator, which means it is generally not possible to normalize (in the sense of having a zero mean) the Minkowski sum $\frac{1}{n}\bigoplus_{i=1}^n X_i$. As a result, the standard approach to generalizing the central limit theorem is to normalize by instead considering the Hausdorff distance between Minkowski sum and the selection expectation.

Theorem 3 (Weil [49]) *Suppose the probability space is nonatomic. If* $X, X_i, i \geq 1$*, are i.i.d. bounded and closed integrable random sets, then we have that:* $\sqrt{n}\cdot d\!\!l_\infty(\frac{1}{n}\bigoplus_{i=1}^n X_i, \mathbb{E}(X)) \rightarrow \sup_u\{\|\zeta(u)\| \mid \|u\| \leq 1\}$ *in distribution, where* $\zeta(u)$ *for* $\|u\| \leq 1$ *is a centered Gaussian random field with covariance given by:* $\mathbb{E}(\zeta(u)\cdot\zeta(v)) = \mathbb{E}(h(u, X)\cdot h(v, X)) - \mathbb{E}(h(u, X))\cdot\mathbb{E}(h(v, X))$.

The difficulty with this central limit theorem is that it lacks a clear geometrical interpretation (in contrast to the classical central limit theorem for random variables) for the limiting distribution, and the question of whether such a geometrical interpretation exists remains open [31]. However, one advantage of this formulation is that it lends itself to a generalization of the Delta method [13] from statistics.

Proposition 3 (Approximate delta method) *Suppose* $r_n d\!\!l_\infty(C_n, C) \rightarrow w$ *in distribution, where* r_n *is a strictly increasing sequence,* $C_n \in \mathcal{K}$ *is a sequence of random sets,* $C \in \mathcal{K}$ *is a deterministic set, and* w *is a random variable. If* S *is a Lipschitz continuous set-valued set function, then*

$$\limsup_n \mathbb{P}(r_n d\!\!l_\infty(S(C_n), S(C)) \geq u) \leq \mathbb{P}(\kappa \cdot w \geq u), \tag{10}$$

where $\kappa \in \mathbb{R}_+$ *is the Lipschitz constant of* S.

Proof Lipschitz continuity of S gives $r_n d\!\!l_\infty(S(C_n), S(C)) \leq \kappa \cdot r_n d\!\!l_\infty(C_n, C)$. Thus $\mathbb{P}(r_n d\!\!l_\infty(S(C_n), S(C)) \geq u) \leq \mathbb{P}(\kappa \cdot r_n d\!\!l_\infty(C_n, C) \geq u)$. The limit superior of both sides gives the result since $r_n d\!\!l_\infty(C_n, C) \rightarrow w$ in distribution. □

Remark 5 The Delta method relates asymptotic distributions of random variables under differentiable functions [13], and the intuition is the derivative is used as a local approximation of the function. The above result demonstrates one instance where Lipschitz continuity can be used as a local approximation for set-valued mappings.

Though $\ominus$ does not generally invert $\oplus$, there is one special case when inversion is possible. If C, D are compact convex sets, then $(C \oplus D) \ominus D = C$ [43]. Using this property, we describe a new central limit theorem for random sets with a particular structure that is useful for statistical applications. Specifically, this result applies to randomly translated sets (RaTS), which are random sets of the form $C = K \oplus \xi$, where K is a deterministic compact convex set, and ξ is a (vector-valued) random variable.

Theorem 4 (Central limit theorem for RaTS) *Suppose the probability space is nonatomic, and that $X, X_i, i \geq 1$, are i.i.d. random sets with $X_i = K \oplus \xi_i$, where K is a deterministic compact convex set and $\xi, \xi_i, i \geq 1$, are i.i.d. (vector-valued) random variables with zero mean and finite variance. Then*

$$\sqrt{n} \cdot ((\tfrac{1}{n} \textstyle\bigoplus_{i=1}^n X_i) \ominus \mathbb{E}(X)) \rightarrow \mathcal{N}(0, \mathbb{E}(\xi\xi^{\mathsf{T}})) \tag{11}$$

in distribution, where $\mathcal{N}(0, \mathbb{E}(\xi\xi^{\mathsf{T}}))$ is a jointly Gaussian random variable with zero mean and covariance matrix given by $\mathbb{E}(\xi\xi^{\mathsf{T}})$.

Proof Since $(\frac{1}{n} \bigoplus_{i=1}^n X_i) \ominus \mathbb{E}(X) = ((\frac{1}{n} \sum_{i=1}^n \xi_i) \oplus K) \ominus K = \frac{1}{n} \sum_{i=1}^n \xi_i$, the result follows by the classical central limit theorem [13]. □

The benefit of this new formulation of the central limit theorem is that it has a clear geometrical interpretation like the classical central limit theorem for random variables, but unfortunately this result only applies to the specific class of RaTS.

4 Kernel regression

We will construct a nonparametric estimator for set-valued functions using an approach that can be viewed as a natural generalization of kernel regression methods for functions [4,5,12,33,48]. These techniques are considered nonparametric because, in contrast to parametric models with a finite number of parameters, the number of parameters in nonparametric models increases as the amount of data increases.

4.1 Problem setup

Consider a Lipschitz continuous set-valued function $S(u) : U \rightrightarrows \mathbb{R}^q$ with random samples $(X_i, S_i) \in U \times \mathbb{R}^q$ for $i = 1, \ldots, n$, where: $U \subseteq \mathbb{R}^d$ is a convex compact set; $S(u)$ is a convex compact set for each $u \in U$; X_i are i.i.d. (vector-valued) random variables with a Lipschitz continuous density function f_X that has the property $f_X(u) > 0$ for $u \in U$; and $S_i = S(X_i) \oplus W_i$ with W_i i.i.d. (vector-valued) random variables that have zero mean $\mathbb{E}(W) = 0$ and finite variance $\|\mathbb{E}(WW^{\mathsf{T}})\| < +\infty$. The problem is to estimate $S(u)$ at any $u \in U$ using the above described samples, and we need convexity of U to ensure its tangent cone is derivable at ∂U [38]; however, our results will hold for all $u \in \mathrm{int}(U)$ unconditional of any such regularity assumptions.

4.2 Kernel functions

Kernel regression is so named because these approaches use kernel functions $\varphi : \mathbb{R} \to \mathbb{R}$, which are functions that are non-negative, bounded, even (i.e., $\varphi(-u) = \varphi(u)$), and have finite support (i.e., there is a constant $\eta \in (0, 1)$ such that $\varphi(u) > 0$ when $|u| \leq \eta$, and $\varphi(u) = 0$ for $|u| \geq 1$). One example of a kernel function is the indicator function $\varphi(u) = (1/2) \cdot \mathbf{1}(|u| < 1)$, and another example is the Epanechnikov kernel $\varphi(u) = (3/4) \cdot (1 - u^2) \cdot \mathbf{1}(u \leq 1)$. Notationally, it is useful to define the family of kernel functions $\varphi_h(u) = h^{-d}\varphi(\|u\|/h)$ and the function $\gamma(u) = \int_{z \in T_U(u)} \varphi_1(z)dz$, where $T_U(u)$ is the tangent cone of U at the point u. (Note $\gamma(u)$ is strictly greater than zero and finite because of the assumptions.) We first prove a lemma about $\varphi_h(u)$:

Lemma 2 *If $h = n^{-1/(d+4)}$, then for $u \in U$ we have*

(a) $\text{as-lim}_n \frac{1}{n} \sum\limits_{i=1}^{n} \varphi_h(X_i - u) = \gamma(u) \cdot f_X(u)$

(b) $\text{as-lim}_n \frac{1}{n} \sum\limits_{i=1}^{n} \varphi_h(X_i - u) \cdot W_i = 0$

(c) $\text{as-lim}_n \frac{1}{n} \sum\limits_{i=1}^{n} \varphi_h(X_i - u) \cdot \|X_i - u\| = 0$

Proof We prove these three results by verifying the hypothesis of Kolmogorov's strong law of large numbers holds in each case, then applying this law of large numbers, and finally computing the expectation of the corresponding quantity in each case. To prove the first result, observe that

$$\lim_n \sum_{i=1}^{n} n^{-2} \cdot \mathrm{var}\big(\varphi_h(X_i - u)\big) \leq \lim_n c/\big(nh^d\big) < \infty, \tag{12}$$

where the first inequality holds for some constant $c \in \mathbb{R}^+$ because $\varphi_h(X_i - u)$ is bounded and nonzero with probability at most $s \cdot h^d$ for some constant $s \in \mathbb{R}^+$; and the second inequality holds because $nh^d = n^{4/(d+4)}$. The finiteness of the above summation means we can apply Kolmogorov's strong law of large numbers, which gives $\text{as-lim}_n \frac{1}{n} \sum_{i=1}^{n} \varphi_h(X_i - u) - \mathbb{E}(\frac{1}{n} \sum_{i=1}^{n} \varphi_h(X_i - u)) = 0$. Our next step is to compute this expectation. Note that

$$\begin{aligned} \mathbb{E}(\tfrac{1}{n} \sum_{i=1}^{n} \varphi_h(X_i - u)) &= \textstyle\int_{x \in \mathbb{R}^d} \varphi_h(x - u) \cdot f_X(x)dx \\ &= \textstyle\int_{z \in \mathbb{B} \cap (U \oplus \{-u\})/h} \varphi_1(z) \cdot f_X(u + hz)dz \end{aligned} \tag{13}$$

where in the last line we made the change of variables $z = (x - u)/h$. Let $R(h) = (\mathbb{B} \cap (U \oplus \{-u\})/h)\backslash T_U(u)$ and $S(h) = (\mathbb{B} \cap T_U(u))\backslash(U \oplus \{-u\})/h$. So we have

$$\begin{aligned}
&\big|\mathbb{E}(\tfrac{1}{n}\sum_{i=1}^{n}\varphi_h(X_i - u)) - f_X(u)\cdot\textstyle\int_{z\in\mathbb{B}\cap T_U(u)}\varphi_1(z)dz\big|\\
&\quad\le \textstyle\int_{z\in\mathbb{B}}\varphi_1(z)\cdot\big|f_X(u+hz) - f_X(u)\big|dz + \int_{z\in R(h)\cup S(h)}|\varphi_1(z)f_X(u+hz)|dz\\
&\quad\le \textstyle\int_{z\in\mathbb{B}}\varphi_1(z)\cdot\kappa h\|z\|dz + s\int_{z\in R(h)}dz + s\int_{z\in S(h)}dz\\
&\quad\le h\cdot\kappa\textstyle\int_{z\in\mathbb{B}}\varphi_1(z)dz + s\int_{z\in R(h)}dz + s\int_{z\in S(h)}dz
\end{aligned} \tag{14}$$

where $\kappa \in \mathbb{R}_+$ is the Lipschitz constant of the density $f_X(u)$, and $s \in \mathbb{R}_+$ is a constant that exists by continuity of $f_X(u)$. Next note $s\int_{z\in R(h)}dz + s\int_{z\in S(h)}dz \to 0$ as $h \to 0$ by Proposition 6.2 and Theorem 4.10 of [38]. Thus taking the limit of (14) gives $\lim_n \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n}\varphi_h(X_i - u)) = f_X(u)\cdot\int_{z\in\mathbb{R}^d}\varphi_1(z)dz$. This proves the first result when combined with the implication of Kolmogorov's strong law of large numbers in our setting, and after noting $\gamma(u) = \int_{z\in\mathbb{B}\cap T_U(u)}\varphi_1(z)dz$ since $\varphi_1(u) = 0$ for $\|u\| > 1$.

For the proof of the second result, let $\langle w\rangle_j$ denote the jth component of the vector w. Next observe that

$$\lim\nolimits_n \sum_{i=1}^{n} n^{-2}\cdot\mathrm{var}\big(\varphi_h(X_i - u)\cdot\langle W_i\rangle_j\big) \le \lim\nolimits_n c\cdot\mathrm{var}\big(\langle W\rangle_j\big)/\big(nh^d\big) < \infty, \tag{15}$$

where the first inequality holds for some constant $c \in \mathbb{R}^+$ because the W_i have zero mean and because $\varphi_h(X_i - u)$ is bounded and nonzero with probability at most $s\cdot h^d$ for some constant $s \in \mathbb{R}^+$; and the second inequality holds because $nh^d = n^{4/(d+4)}$. The finiteness of the above summation means Kolmogorov's strong law of large numbers gives as-$\lim_n \frac{1}{n}\sum_{i=1}^{n}\varphi_h(X_i - u)\cdot\langle W_i\rangle_j - \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n}\varphi_h(X_i - u)\cdot\langle W_i\rangle_j) = 0$. But the W_i are zero mean, and so we have that $\mathbb{E}(\frac{1}{n}\sum_{i=1}^{n}\varphi_h(X_i - u)\cdot\langle W_i\rangle_j) = 0$.

To prove the third result, observe that

$$\lim\nolimits_n \sum_{i=1}^{n} n^{-2}\cdot\mathrm{var}\big(\varphi_h(X_i - u)\cdot\|X_i - u\|\big) \le \lim\nolimits_n c/\big(nh^d\big) < \infty, \tag{16}$$

where the first inequality holds for some constant $c \in \mathbb{R}^+$ because U is a compact set and because $\varphi_h(X_i - u)$ is bounded and nonzero with probability at most $s\cdot h^d$ for some constant $s \in \mathbb{R}^+$; and the second inequality holds because $nh^d = n^{4/(d+4)}$. The finiteness of the above summation means we can apply Kolmogorov's strong law of large numbers, which gives as-$\lim_n \frac{1}{n}\sum_{i=1}^{n}\varphi_h(X_i - u)\cdot\|X_i - u\| - \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n}\varphi_h(X_i - u)\cdot\|X_i - u\|) = 0$. Our next step is to compute this expectation. Note that

$$\begin{aligned}
\mathbb{E}(\tfrac{1}{n}\sum_{i=1}^{n}\varphi_h(X_i - u)\cdot\|X_i - u\|) &\le \textstyle\int_{x\in\mathbb{R}^d}\varphi_h(x - u)\cdot\|x - u\|\cdot f_X(x)dx\\
&\le h\textstyle\int_{z\in\mathbb{R}^d}\varphi_1(z)\cdot z\cdot f_X(u + hz)dz\\
&\le c\cdot h = c\cdot n^{-1/(d+4)}
\end{aligned} \tag{17}$$

where the second line makes the change of variables $z = (x - u)/h$, and the third line holds for some constant $c \in \mathbb{R}^+$ because the kernel has finite support and the density is continuous. The above expectation is non-negative, and so $\lim_n \mathbb{E}(\frac{1}{n}\sum_{i=1}^n \varphi_h(X_i - u) \cdot \|X_i - u\|) = 0$. This proves the third result when combined with the outcome of Kolmogorov's strong law of large numbers. □

4.3 Kernel regression estimator

We define a kernel regression estimate of S at the point u to be

$$\widehat{S}(u) = \Big[\tfrac{1}{n}\bigoplus\nolimits_{i=1}^n \varphi_h(X_i - u) \cdot S_i\Big] \cdot \Big[\tfrac{1}{n}\sum_{i=1}^n \varphi_h(X_i - u)\Big]^{-1} \tag{18}$$

The following theorem proves the strong pointwise consistency of this estimator.

Theorem 5 *If* $h = n^{-1/(d+4)}$, *then* as-lim$_n\, \widehat{S}(u) = S(u)$ *for* $u \in U$.

Proof Let $\kappa \in \mathbb{R}_+$ be the Lipschitz constant of S, and note that by Lipschitz continuity we have

$$\begin{aligned}
&\tfrac{1}{n}\bigoplus\nolimits_{i=1}^n \varphi_h(X_i - u) \cdot \big(S(u) \oplus W_i\big) \subseteq\\
&\tfrac{1}{n}\bigoplus\nolimits_{i=1}^n \varphi_h(X_i - u) \cdot \big(S_i \oplus \kappa\|X_i - u\|\mathbb{B}\big) \subseteq\\
&\tfrac{1}{n}\bigoplus\nolimits_{i=1}^n \varphi_h(X_i - u) \cdot \big(S(u) \oplus 2\kappa\|X_i - u\|\mathbb{B} \oplus W_i\big)
\end{aligned} \tag{19}$$

Corollary 2(c) and Lemma 2(a) give as-lim$_n\, \frac{1}{n}\bigoplus_{i=1}^n \varphi_h(X_i - u) \cdot S(u) = \gamma(u) \cdot f_X(u) \cdot S(u)$, and Corollary 2(a) and Lemma 2(b) yield as-lim$_n\, \frac{1}{n}\bigoplus_{i=1}^n \varphi_h(X_i - u) \cdot (S(u) \oplus W_i) = \gamma(u) \cdot f_X(u) \cdot S(u)$. Corollary 2(c) and Lemma 2(c) give as-lim$_n\, \frac{1}{n}\bigoplus_{i=1}^n \varphi_h(X_i - u) \cdot \kappa\|X_i - u\|\mathbb{B} = 0$, and so Corollary 2(a) implies as-lim$_n\, \frac{1}{n}\bigoplus_{i=1}^n \varphi_h(X_i - u) \cdot \big(S(u) \oplus 2\kappa\|X_i - u\|\mathbb{B} \oplus W_i\big) = \gamma(u) \cdot f_X(u) \cdot S(u)$. So applying the sandwich lemma to (19) yields as-lim$_n\, \frac{1}{n}\bigoplus_{i=1}^n \varphi_h(X_i - u) \cdot \big(S_i \oplus \kappa\|X_i - u\|\mathbb{B}\big) = \gamma(u) \cdot f_X(u) \cdot S(u)$. Corollary 3 gives as-lim$_n\, \frac{1}{n}\bigoplus_{i=1}^n \varphi_h(X_i - u) \cdot S_i = \gamma(u) \cdot f_X(u) \cdot S(u)$. Finally, using Corollary 2(d) and Lemma 2(a) imply that as-lim$_n\, \widehat{S}(u) = S(u)$. □

4.4 Algorithms to compute kernel regression estimator

The statistical consistency of our kernel regression estimator is a theoretical result, and numerical computation of this estimator using the measured data (u_i, s_i) for $i = 1, \ldots, n$ needs some discussion. The key point is that the corresponding algorithm used to compute the estimator depends on the representation of the sets s_i. Since the random sets S_i are RaTS, we only need to consider different representations of convex sets. Moreover, we focus our discussion on polytope representations since any compact convex set can be approximated arbitrarily well by polytopes [43].

If the sets s_i are each represented by polynomial time membership oracles, then

$$\tfrac{1}{n}\bigoplus_{i=1}^{n}\varphi_h(x_i-u)\cdot s_i=\Big\{\tfrac{1}{n}\bigoplus_{i=1}^{n}\varphi_h(x_i-u)\cdot t_i : t_i\in s_i \text{ for } i=1,\dots,n\Big\},\quad(20)$$

and so membership in the Minkowski sum can be determined in polynomial time. Polynomial time membership oracles exist for s_i in a known compact set G, with a self-concordant barrier function for G and the functions defining s_i [32]: the measurement of s_i would consist of the function parameters defining s_i, and set membership is determined by using interior point to solve a feasibility problem. Examples include polytopes $s_i=\{t_i : a_i t_i\le b_i\}$, with measured data a_i, b_i; second-order cone sets $s_i=\{t_i : \|a_{i,j}t_i+b_{i,j}\|_2\le c_{i,j}^{\mathsf{T}}t_i+d_{i,j} \text{ for } j=1,\dots,k\}$, with measured data $a_{i,j}, b_{i,j}, c_{i,j}, d_{i,j}$; and combinations thereof. Other examples can be found in [32].

Next suppose the sets s_i are each represented by the zonotopes $s_i=\bigoplus_{k=1}^{p}w_{ik}\cdot z_k$, where w_{ik} are weights and z_k are vectors, which are polytopes defined as the Minkowski sum of vectors. Restated, the observations are the w_{ik} and z_k. Then

$$\tfrac{1}{n}\bigoplus_{i=1}^{n}\varphi_h(x_i-u)\cdot s_i=\bigoplus_{k=1}^{p}\Big[\tfrac{1}{n}\sum_{i=1}^{n}\varphi_h(x_i-u)\cdot w_{ik}\Big]\cdot z_k,\quad(21)$$

and so the Minkowski sum is polynomial time computable for this representation.

Lastly, suppose the sets s_i are represented by the convex hull of a finite set of p_i vertices, meaning that $s_i=\mathrm{co}(\{v_{i1},\dots,v_{ip_i}\})$. In this setting the measurements are the vertices of each set s_i, and the Minkowski sum is given by

$$\begin{aligned}\tfrac{1}{n}\bigoplus_{i=1}^{n}\varphi_h(x_i-u)\cdot s_i=\mathrm{co}\Big(\Big\{\tfrac{1}{n}\sum_{i=1}^{n}\varphi_h(x_i-u)\cdot v_{ij_i} :\\ \text{for } j_i=1,\dots,p_i \text{ and } i=1,\dots,n\Big\}\Big).\end{aligned}\quad(22)$$

This is a polynomial time computation since the number of vertices is finite.

4.5 Numerical example

We conclude our discussion on kernel regression of set-valued functions with a numerical example to visually demonstrate the estimation problem being solved by our estimator. Consider the set-valued function in the bottom-left of Fig. 1, given by

$$S(u)=\begin{cases}\big[-2,-\frac{2u+1}{u}+2\big], & \text{if } u\in\big[-2,-\frac{1}{4}\big]\\ \big[-2,\ 2\big], & \text{if } u\in\big[-\frac{1}{4},\ \frac{1}{4}\big]\\ \big[\frac{4u-1}{u}-2,\ 2\big], & \text{if } u\in\big[\ \frac{1}{4},\ 2\big]\end{cases}\quad(23)$$

The X_i variables have a $U(-2,2)$ distribution, and each measurement s_i is in a vertex representation. The noise W_i has a $U(-1,1)$ distribution, meaning its variance is $1/6$. The top row of Fig. 1 shows measurements for $n=10^2$, $n=10^3$, and $n=10^4$ data points, respectively; and the bottom row shows estimates computed by (18) and

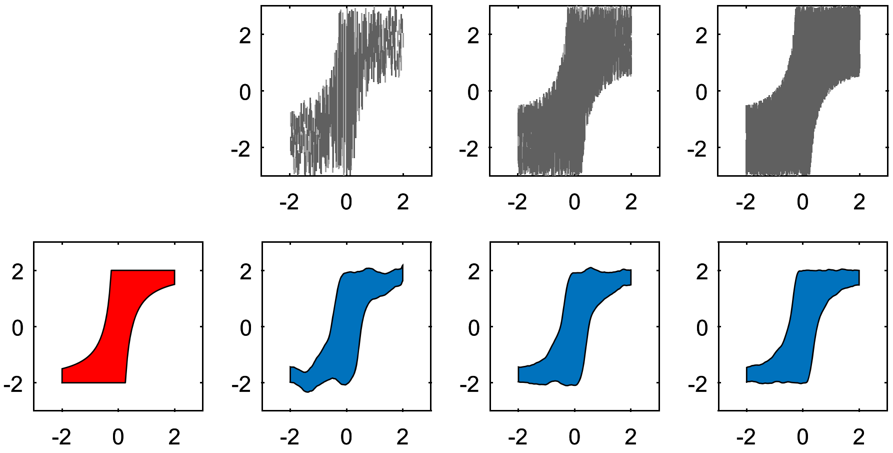

Fig. 1 The x-axes on the plots are u. The top row from left to right shows the noisy measurements (u_i, s_i) for $n = 10^2$, $n = 10^3$, and $n = 10^4$ data points, respectively, and the bottom row shows the set-valued function $S(u)$ being estimated and the corresponding estimates $\hat{S}(u)$ from our kernel regression estimator

(22) with an Epanechnikov kernel.[1] This example shows that as the amount of data increases, the estimates $\hat{S}(u)$ converge pointwise to the actual set-valued function $S(u)$.

5 Inverse approximate optimization

Inverse optimization involves computing parameters that make measured solutions optimal [1,7,11,16,20,24]. In contrast, the inverse approximate optimization problem makes noisy measurements of suboptimal solutions, and the goal is to estimate the amount of suboptimality and to estimate the parameters of the optimization problem generating the data. In principle, the VIA [11] and KKT [24] estimators can provide estimates of the desired quantities; but we show their estimates are statistically inconsistent. As a result, we construct an estimator for inverse approximate optimization, prove its statistical consistency, and then discuss some possible generalizations.

5.1 Problem setup

Consider a parametric convex optimization problem

$$V(u, \theta) = \min_x \left\{ f(x, u, \theta) \;\middle|\; g(x, u, \theta) \leq 0 \right\} \tag{24}$$

in which f, g are continuous functions that are convex in x for each fixed value of u and θ, and assume that for all u, θ the constraint qualification there exists x such that $g(x, u, \theta) < 0$ holds. (Note this constraint representation is fully general since we

[1] Our code http://ieor.berkeley.edu/~aaswani/code/ssvf.zip runs in a few seconds.

can write $g = \max g_i$.) We use the definition that ϵ-optimal solutions are those in the set

$$S(u, \epsilon, \theta) = \epsilon\text{-arg}\min_x \left\{ f(x, u, \theta) \mid g(x, u, \theta) \le 0 \right\} = \\ \{x : f(x, u, \theta) \le V(u, \theta) + \epsilon, g(x, u, \theta) \le 0\}. \tag{25}$$

Our results also apply when ϵ-optimal solutions are defined as in (25) but with $g(x, u, \theta) \le \epsilon$. The difference is (25) does not allow any constraint violation, while the alternative definition allows ϵ constraint violation. Note there are other notions of ϵ-optimal solutions like distance to the KKT graph, but we do not consider these.

Now suppose ϵ-optimal solutions of (24) generate random samples $(U_i, Y_i) \in D \times \mathbb{R}^p$ for $i = 1, \dots, n$, where: U_i are i.i.d. (vector-valued) random variables distributed on the set $D \subseteq \mathbb{R}^d$; $Y_i = X_i + W_i$, where X_i are i.i.d. (vector-valued) random variables distributed on $S(U_i, \epsilon_0, \theta_0)$ with constants $\epsilon_0 \in \mathbb{R}_+$ and $\theta_0 \in \mathbb{R}^p$; and W_i are i.i.d. (vector-valued) random variables with zero mean $\mathbb{E}(W_i) = 0$ and distributed on a known convex set W with finite support (which implies finite variance). We also assume the densities of W_i, X_i are strictly positive on the interior of their supports (i.e., $f_W(u) > 0$ for $u \in \text{int}(W)$ and $f_X(u|U_i) > 0$ for $u \in \text{int}(S(U_i, \epsilon_0, \theta_0))$).

The inverse approximate optimization problem is to estimate (ϵ_0, θ_0) using the (U_i, Y_i) for $i = 1, \dots, n$. Note that we assume the functional forms of f, g are fixed. Let $E \subseteq \mathbb{R}_+$ be a known closed set such that $\epsilon_0 \in E$, and let $\Theta \subseteq \mathbb{R}^p$ be a known compact set such that $\theta_0 \in \Theta$; the intuition is that these sets represent prior knowledge that constrain the parameters and amount of solution suboptimality. The choice $E = \mathbb{R}_+$ corresponds to a situation with no such prior knowledge on ϵ_0, and the compactness assumption on Θ is not restrictive in practice because this set can be made arbitrarily large. (Unbounded Θ can also be used when a compactification with certain technical properties exists [9].) A so-called *identifiability condition* [13] is also needed. We assume that if $(\epsilon_0, \theta_0) \in E \times \Theta$ then $\mathbb{E}(d^2(Y_i, S(U_i, \epsilon, \theta) \oplus W)) > 0$ for all $\epsilon \in [0, \epsilon_0)$ and $\theta \in \Theta \backslash \{\theta_0\}$. An identifiability condition (such as the one we have assumed) intuitively says that different parameters of the model produce different outputs.

5.2 Inconsistency of existing estimators

The VIA [11] (which minimizes the first order suboptimality of the data) and KKT [24] (which minimizes the KKT suboptimality of the data) estimators are statistically inconsistent for $\epsilon_0 = 0$ [7], but since these approaches minimize the amount of suboptimality of the measured data it is initially unclear without further analysis whether these approaches are inconsistent for problem instances with $\epsilon_0 > 0$. The following result provides qualitative insights into the behavior of these estimators.

Proposition 4 *Let $r \in \mathbb{R}_+$ be a constant, and suppose $f = x^2$, $g = [x - 1; -x - 1]$, $\epsilon_0 = 1$, $W = \{w : \|w\| \le r\}$, and W_i, X_i are uniformly distributed. Then estimates $\hat{\epsilon}$ generated by the VIA [11] and KKT [24] methods are such that* as-lim $\inf_n \hat{\epsilon} > r/3$.

Proof The KKT estimate is given by

$$\hat{\epsilon} = \max \Big\{ \frac{1}{n} \sum_{i=1}^{n} \langle g(Y_i) \rangle_1^+, \tfrac{1}{n} \sum_{i=1}^{n} \langle g(Y_i) \rangle_2^+, \tfrac{1}{n} \sum_{i=1}^{n} |2Y_i + \langle \Lambda_i \rangle_1 - \langle \Lambda_i \rangle_2|, \\ \frac{1}{n} \sum_{i=1}^{n} |\langle \Lambda_i \rangle_1 \cdot (Y_i - 1)|, \tfrac{1}{n} \textstyle\sum_{i=1}^{n} |\langle \Lambda_i \rangle_2 \cdot (-Y_i - 1)| \Big\} \tag{26}$$

where these Λ_i are the minimizers of the below optimization problem

$$\min \big\{ \tfrac{1}{n} \sum_{i=1}^{n} |2Y_i + \langle \lambda_i \rangle_1 - \langle \lambda_i \rangle_2| + \langle \lambda_i \rangle_1 \cdot |Y_i - 1| + \\ \langle \lambda_i \rangle_2 \cdot |-Y_i - 1| \;\big|\; \lambda_i \geq 0, i = 1, \dots, n \big\}. \tag{27}$$

Since ϵ-$\arg\min_x \{ f(x, u, \theta) \mid g(x, u, \theta) \leq 0 \} = [-1, 1]$ under the hypothesis of this proposition, it holds the Y_i are i.i.d. and have triangular distribution with lower limit $-r - 1$, upper limit $r + 1$, and mode 0. Hence the density of $C_i = \langle g(Y_i) \rangle_1^+$ is given by

$$f_C(u) = \Big(\tfrac{1}{2} + \tfrac{1}{r+1} - \tfrac{1}{2(r+1)^2} \Big) \cdot \delta(u) + \Big(\tfrac{r}{(r+1)^2} - \tfrac{1}{(r+1)^2} \cdot u \Big) \cdot \mathbf{1}(u \in [0, r]), \tag{28}$$

where $\delta(u)$ is the Dirac delta function. So $\mathbb{E}(C_i) = \frac{r+1}{3}$, and the strong law of large numbers implies $\frac{r+1}{3} = \text{as-lim}_n \frac{1}{n} \sum_{i=1}^{n} C_i = \text{as-lim inf}_n \frac{1}{n} \sum_{i=1}^{n} C_i \leq \text{as-lim inf}_n \hat{\epsilon}$.

The VIA estimate is given by $\hat{\epsilon} = \frac{1}{n} \sum_{i=1}^{n} |\hat{\epsilon}_i|$ where ϵ_i are the minimizers to

$$\min \Big\{ \tfrac{1}{n} \sum_{i=1}^{n} |\hat{\epsilon}_i| \;\Big|\; 2Y_i(x_i - Y_i) \geq -\hat{\epsilon}_i \text{ for } x_i \in [-1, 1], i = 1, \dots, n \Big\} \tag{29}$$

However, observe that $2Y_i(x_i - Y_i) \geq -\hat{\epsilon}_i$ for $x_i \in [-1, 1]$ simplifies to the constraint $-2(|Y_i| + Y_i^2) \geq -\hat{\epsilon}_i$. Since the above optimization is minimizing each $|\hat{\epsilon}_i|$, this means the constraint will be $\hat{\epsilon}_i = 2(|Y_i| + Y_i^2)$ at optimality. Recall that as shown in the proof for KKT, the Y_i have a triangular distribution with lower limit $-r - 1$, upper limit $r + 1$, and mode 0. This means $\mathbb{E}(|\hat{\epsilon}_i|) = \frac{2r+2}{3} + \frac{(r+1)^2}{9}$. Applying the strong law of large numbers gives $\frac{2r+2}{3} + \frac{(r+1)^2}{9} = \text{as-lim}_n \frac{1}{n} \sum_{i=1}^{n} |\hat{\epsilon}_i| = \text{as-lim}_n \hat{\epsilon}$. □

This proposition shows that existing approaches cannot distinguish between noise in measurements versus suboptimality of the solutions. The reason is that these approaches are minimizing an incorrect error metric: They minimize the amount of suboptimality of the measured data, and this is an incorrect error metric when the measured data is noisy because the noise increases the suboptimality of the measured data. Moreover, this indistinguishability of existing approaches is unbounded in the sense that as the noise variance increases then their estimates of suboptimality increase in an unbounded way. Such behavior is undesirable, and in fact the above result gives the following corollary on the statistical properties of VIA and KKT.

Corollary 4 *The VIA* [11] *and KKT* [24] *estimators are statistically inconsistent.*

Proof By definition an estimator is consistent for a class of models if and only if it is consistent for each model in that class. Thus to show inconsistency of VIA and KKT it suffices to show inconsistency for a single model. The above proposition establishes inconsistency of VIA and KKT for a particular model because $\epsilon_0 = 1$ while as-lim $\inf_n \hat{\epsilon} > r/3$, meaning these approaches are inconsistent when $r > 3$. □

5.3 Approximate bilevel programming (ABP) estimator

To correct the indistinguishability (between suboptimality of solutions and noise in measurements) problem faced by existing approaches, we instead propose an estimator that explicitly models the measured data as consisting of a suboptimal solution added to noise. More specifically, we propose the following statistical estimator

$$(\breve{\epsilon}, \breve{\theta}) \in \arg\min \left\{ \tfrac{1}{n} \sum_{i=1}^{n} d^2(Y_i, S(U_i, \epsilon, \theta) \oplus W) + \lambda \cdot \epsilon \;\middle|\; \epsilon \in E, \theta \in \Theta \right\} \tag{30}$$

where $\lambda \in \mathbb{R}_+$ and d^2 is the squared distance function defined in the preliminaries. It is also useful to consider estimators defined as approximate solutions to the above optimization problem. Let $z \in \mathbb{R}_+$ be a nonnegative value, and define the estimates

$$(\hat{\epsilon}, \hat{\theta}) \in \Big\{ \epsilon \in E, \theta \in \Theta : \tfrac{1}{n} \sum_{i=1}^{n} d^2(Y_i, S(U_i, \epsilon, \theta) \oplus W) + \lambda \cdot \epsilon \leq \tfrac{1}{n} \sum_{i=1}^{n} d^2(Y_i, S(U_i, \breve{\epsilon}, \breve{\theta}) \oplus W) + \lambda \cdot \breve{\epsilon} + z \Big\}. \tag{31}$$

For notational convenience, we will call this estimator the ABP estimator. Note these estimates are defined as being any z -arg min of the optimization problem (30).

Theorem 6 *The ABP estimator is strongly statistically consistent, meaning we have* as-$\lim_n(\hat{\epsilon}, \hat{\theta}) = (\epsilon_0, \theta_0)$ *whenever* $\lambda = 1/n$ *and* $\lim_n(n \cdot z) = 0$.

Proof Our first step is to show $d^2(y, S(u, \epsilon, \theta) \oplus W)$ satisfies certain continuity properties. Note $\{x : g(x, u, \theta) \leq 0\}$ is continuous by Example 5.10 of [38], and so $V(u, \theta)$ is continuous by the Berge maximum theorem [10]. Noting $d^2(y, S(u, \epsilon, \theta) \oplus W) = \min\{\|y - \hat{y}\|^2 \mid \hat{y} = \hat{x} + \hat{\epsilon}, \hat{\epsilon} \in W, f(\hat{x}, u, \theta) \leq V(u, \theta) + \epsilon, g(\hat{x}, u, \theta) \leq 0\}$, we can apply the Berge maximum theorem [10] since this feasible set is osc by Example 5.8 of [38]: This implies $d^2(y, S(u, \epsilon, \theta) \oplus W)$ is lower semicontinuous in (ϵ, θ), and so $\mathbb{E}(d^2(Y_i, S(U_i, \epsilon, \theta) \oplus W))$ is lower semicontinuous in (ϵ, θ) by Fatou's lemma.

Next note that the estimate $(\breve{\epsilon}, \breve{\theta})$ also minimizes the optimization problem

$$\min \left\{ \sum_{i=1}^{n} d^2(Y_i, S(U_i, \epsilon, \theta) \oplus W) + n\lambda \cdot \epsilon \;\middle|\; \epsilon \in E, \theta \in \Theta \right\} \tag{32}$$

But $n\lambda = 1$ by assumption, and so the objective of (32) is nondecreasing in n. Hence $\mathbb{P}(\text{e-lim}_n \sum_{i=1}^n d^2(Y_i, S(U_i, \epsilon, \theta) \oplus W) + \epsilon = \sup_n \sum_{i=1}^n d^2(Y_i, S(U_i, \epsilon, \theta) \oplus W) + \epsilon) = 1$ by Proposition 7.4 of [38], where e-lim is the epi-limit [38]. We next prove that

$$\begin{aligned}\mathbb{P}(\sup_n \textstyle\sum_{i=1}^n d^2(Y_i, S(U_i, \epsilon, \theta) \oplus W) > 0 \\ \text{for } (\epsilon, \theta) \in ([0, \epsilon_0] \times \Theta)\backslash\{(\epsilon_0, \theta_0)\}) = 1,\end{aligned} \tag{33}$$

and our approach is to use a well-known covering argument originally due to Wald [47]. Let S_k be a decreasing sequence (i.e., $S_k \supseteq S_{k+1} \supseteq \cdots$) of open neighborhoods of (ϵ_0, θ_0), with $\lim_k S_k = \{(\epsilon_0, \theta_0)\}$. Since $\mathbb{E}(d^2(Y_i, S(U_i, \epsilon, \theta) \oplus W))$ is lower semicontinuous in (ϵ, θ), this means $\min\{\mathbb{E}(d^2(Y_i, S(U_i, \epsilon, \theta) \oplus W)) \mid (\epsilon, \theta) \in ([0, \epsilon_0] \times \Theta)\backslash S_k\} > 0$ by the identifiability condition. Thus there exists $\nu_k > 0$ such that $\mathbb{E}(d^2(Y_i, S(U_i, \epsilon, \theta) \oplus W)) > 2\nu_k$ for $(\epsilon, \theta) \in ([0, \epsilon_0] \times \Theta)\backslash S_k$. By lower semicontinuity of $d^2(y, S(u, \epsilon, \theta) \oplus W)$ and the monontone convergence theorem, there exists an open neighborhood $T_k(\epsilon, \theta)$ for each $(\epsilon, \theta) \in ([0, \epsilon_0] \times \Theta)\backslash S_k$ so that we have $\mathbb{E}(\inf\{d^2(Y_i, S(U_i, \epsilon', \theta') \oplus W) \mid (\epsilon', \theta') \in T_k(\epsilon, \theta)) > \mathbb{E}(d^2(Y_i, S(U_i, \epsilon, \theta) \oplus W)) - \nu_k$. Since $([0, \epsilon_0] \times \Theta)\backslash S_k$ is compact, there exists a finite set $F_k \in ([0, \epsilon_0] \times \Theta)\backslash S_k$ such that $T_k(\epsilon, \theta)$ for $(\epsilon, \theta) \in F_k$ forms a finite subcover of $([0, \epsilon_0] \times \Theta)\backslash S_k$. Combining the above with the Borel-Cantelli lemma implies $\mathbb{P}(\inf\{\sup_n \sum_{i=1}^n d^2(Y_i, S(U_i, \epsilon', \theta') \oplus W) \mid (\epsilon', \theta') \in T_k(\epsilon, \theta)\} > 0) = 1$ for each $(\epsilon, \theta) \in F_k$, which by the finiteness of F_k implies that $\mathbb{P}(\sup_n \sum_{i=1}^n d^2(Y_i, S(U_i, \epsilon, \theta) \oplus W) > 0 \text{ for } (\epsilon, \theta) \in ([0, \epsilon_0] \times \Theta)\backslash S_k)) = 1$. The desired (33) follows since we choose the sequence S_k such that $S_k \downarrow \{(\epsilon_0, \theta_0)\}$.

Next consider the optimization problem

$$\min\left\{ \sup_n \sum_{i=1}^n d^2(Y_i, S(U_i, \epsilon, \theta) \oplus W) + \epsilon \;\middle|\; \epsilon \in E, \theta \in \Theta \right\} \tag{34}$$

Note (ϵ_0, θ_0) is feasible for both (32) and (34), and so the minimums of (32) and (34) are both less than or equal to ϵ_0. This means $\epsilon > \epsilon_0$ cannot minimize (34). Furthermore, using (33) implies that almost surely the (unique) minimizer of (34) is (ϵ_0, θ_0), and almost surely the minimum value of (34) is ϵ_0. But from the argument in the preceding paragraph, (32) epi-converges almost surely to (34) since E, Θ are fixed. The result now follows from Theorem 7.33 of [38]. □

The above result concerns almost sure convergence of the ABP estimates $(\hat{\epsilon}, \hat{\theta})$ to the actual parameters (ϵ_0, θ_0), but a related question is whether the corresponding solution set estimates $S(u, \hat{\epsilon}, \hat{\theta})$ converge to the actual solution sets $S(u, \epsilon_0, \theta_0)$. Our semicontinuous mapping theorem can be used to establish almost sure convergence of the solution set estimates, and this argument leads to the the following corollary.

Corollary 5 *We have that* as-lim $\sup_n S(u, \hat{\epsilon}, \hat{\theta}) \subseteq S(u, \epsilon_0, \theta_0)$ *for* $u \in D$. *If* $\epsilon_0 > 0$ *or* $f(\cdot, u, \theta)$ *is strictly convex in* x, *then* as-$\lim_n S(u, \hat{\epsilon}, \hat{\theta}) = S(u, \epsilon_0, \theta_0)$ *for* $u \in D$.

Proof The above proof established that $S(u, \epsilon, \theta)$ is osc in ϵ, θ. And so the first part of the corollary follows by the semicontinuous mapping theorem. If $\epsilon_0 > 0$ then

$S(u, \epsilon_0, \theta_0)$ is continuous at (ϵ_0, θ_0) by Example 5.10 of [38]. If $f(\cdot, u, \theta)$ is strictly convex in x, then $S(u, 0, \theta_0)$ is single-valued [38]. Hence $S(u, 0, \theta_0)$ is continuous because a single-valued, osc, and locally bounded function is continuous [38]. Thus the second part of the corollary follows from the semicontinuous mapping theorem.□

5.4 Algorithms to compute ABP estimator

We next discuss numerical computation of ABP using the data (u_i, y_i) for $i = 1, \dots, n$. The ABP estimator is an approximate (i.e., the solution sets have ϵ possibly greater than zero) bilevel program, which are optimization problems where some decision variables are solutions to optimization problems that are called the lower level problem. One approach to solve bilevel programs replaces the lower level problem with its KKT conditions [2,17], and this can sometimes be rewritten as mixed-integer programs that may be numerically solved quickly [6]. Another approach upper bounds the objective function of the lower level problem by its value function [35,50].

Here we describe how a third approach that upper bounds the objective function of the lower level problem by its dual function [7,34] can be used to compute the ABP estimator. If $h(u, \theta, \lambda)$ is the Lagrangian dual function corresponding to (24), then under mild conditions ensuring zero duality gap the ABP estimator is given by

$$\begin{aligned}(\breve{\epsilon}, \breve{\theta}) \in \arg\min\ & \tfrac{1}{n}\sum_{i=1}^{n} \|y_i - \hat{x}_i\|^2 + \lambda \cdot \epsilon \\ \text{s.t. } & f(\hat{x}_i, u_i, \theta) \leq h(u_i, \theta, \lambda_i) + \epsilon \\ & g(\hat{x}_i, u_i, \theta) \leq 0 \\ & \lambda_i \geq 0, \epsilon \in E, \theta \in \Theta \end{aligned} \tag{35}$$

This duality-based reformulation can be numerically solved by two different algorithms [7,34], which we briefly describe here. More details can be found in the corresponding references, and both algorithms assume the sets E, Θ are compact.

Since the reformulation (35) is a convex optimization problem for fixed (ϵ, θ), one algorithm [7] for computing ABP is to: discretize the set $E \times \Theta$ into a finite set $\Delta = \{(\epsilon_1, \theta_1), \dots, (\epsilon_k, \theta_k)\}$ such that it forms a set covering with balls of a prescribed radius, compute the minimum objective function value of (35) for $(\epsilon, \theta) \in \Delta$ (which we call $Q(\epsilon, \theta)$), and then choose estimates $(\hat{\epsilon}, \hat{\theta}) = \arg\min\{Q(\epsilon, \theta) \mid (\epsilon, \theta) \in \Delta\}$. A result from [7] implies that estimates chosen using this *enumeration algorithm* satisfy the assumptions of Theorem 6, which is sufficient for statistical consistency.

A second algorithm [34] replaces the Lagrangian dual by a numerically computed dual. Partial dualization is used to define a regularized dual function (RDF)

$$h_\mu(u, \theta, \lambda) = \min_x \left\{\mu \cdot \|x\|^2 + f(x, u, \theta) + \lambda^\mathsf{T} g(x, u, \theta) \mid x \in X\right\}. \tag{36}$$

Here, X is any compact set defined such that $\{x : \exists(u, \theta) \in U \times \Theta \text{ s.t. } g(x, u, \theta) \leq 0\} \subseteq \text{int}(X)$. The intuition is that X is a set that contains all the feasible sets of (24) within its interior. When g does not depend on (u, θ), we can choose $X = \{x : l_i - 1 \leq x_i \leq u_i + 1\}$ with $u_i = \max\{x_i \mid g(x) \leq 0\}$ and $l_i = \min\{x_i \mid g(x) \leq 0\}$

that are computed by solving convex optimization problems. Many applications of inverse approximate optimization consist of such a setting where the feasible set is independent of the inputs u or the parameters θ. The benefit of the RDF is it can be numerically computed because it is a convex optimization problem, and that its gradient

$$\begin{aligned}\nabla_\theta h_\mu(u,\theta,\lambda) &= \nabla_\theta f(x,u,\theta) + \lambda^\mathsf{T}\cdot\nabla_\theta g(x,u,\theta)\\ \nabla_\lambda h_\mu(u,\theta,\lambda) &= g(x,u,\theta)\end{aligned} \tag{37}$$

always exists when $\mu > 0$. In contrast, the Lagrangian dual is usually only directionally differentiable but not differentiable. The algorithm proceeds by using a nonlinear numerical solver to solve a sequence of optimization problems in which μ goes to 0.

A third possibility is a polynomial time approximation algorithm with the property that statistical consistency holds as the amount of samples n increases to infinity. Such an algorithm has been constructed, when f is affine in θ and g does not depend on θ, for inverse optimization with noisy data [7]; it uses kernel regression to pre-smooth the data and then solves a convex problem corresponding to inverse optimization assuming no noise in the pre-smoothed data. Here we sketch a similar algorithm for inverse approximate optimization, and we leave its analysis for future work. Define $\hat{S}(u) = \mathrm{co}(\{y_i : \|u_i - u\| \leq h\}) \ominus W$ for $h \in \mathbb{R}_+$, and choose the data $\hat{x}_i$ by sampling from the uniform distribution on $\hat{S}(u_i)$. The estimate $(\breve{\epsilon}, \breve{\theta})$ is computed by solving (35) with the change that the $g(\hat{x}_i, u_i, \theta) \leq 0$ constraints are removed.

5.5 Numerical example

We next consider a numerical example to visually compare estimates of $S(u, \epsilon_0, \theta_0)$ produced by our ABP estimator and the VIA [11] and KKT [24] estimators. Suppose $x \in \mathbb{R}$, $f = -(\theta + u)\cdot x$, $g = [x-2; -x-2]$, $\epsilon_0 = 1$, $\theta_0 = 0$, $W = \{w : \|w\| \leq 1\}$, U_i has a uniform distribution $U(-2, 2)$, X_i is uniformly distributed on $S(U_i, \epsilon_0, \theta_0)$, $E = \{\epsilon : 0.1 \leq \epsilon \leq 10\}$, and $\Theta = \{\theta : -2 \leq \theta \leq 2\}$. The solution set $S(u, \epsilon_0, \theta_0)$ in this setting is shown in the left column of Fig. 2. Each measurement (u_i, y_i) for this example is a point, and the top row of Fig. 2 shows the measurements for $n = 10^1$, $n = 10^2$, and $n = 10^3$ data points, respectively. The rows below show the estimated (using the measurements shown above) solution set as computed by ABP, KKT, and VIA, respectively.[2] This example shows that as the number of measurements increases, the solution set estimated by ABP (KKT and VIA) converges (does not converge) to the actual solution set. This statistical behavior is expected given our theoretical results on the strong consistency of ABP and the statistical inconsistency of KKT and VIA.

[2] Our code http://ieor.berkeley.edu/~aaswani/code/ssvf.zip runs in about five minutes.

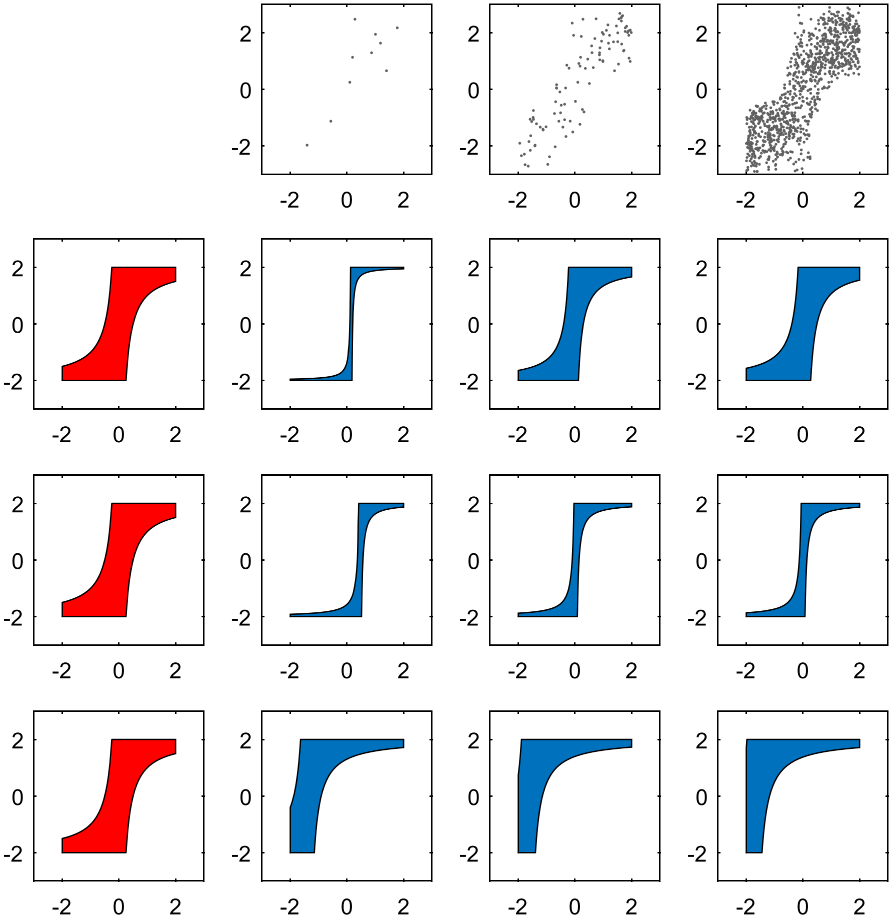

Fig. 2 The x-axes on the plots are u. For the solution set $S(u, \epsilon, \theta) = \epsilon\text{-arg}\min_x\{-(\theta+u)\cdot x \mid -2 \leq x \leq 2\}$, the left column shows the actual $S(u, \epsilon_0, \theta_0)$ when $\epsilon_0 = 1$ and $\theta_0 = 0$. The top row from left to right shows the noisy measurements (u_i, y_i) for $n = 10^1$, $n = 10^2$, and $n = 10^3$ data points, respectively, and the rows below show the estimated solution set as computed by ABP, KKT, and VIA, respectively

5.6 Related inverse optimization problems

In our problem setup, the measurement noise W_i had a distribution with a finite support. However, noise models commonly used in statistics include distributions with unbounded support but finite variance. The canonical example is W_i that are jointly Gaussian with zero mean and finite covariance. A heuristic approach for distributions with unbounded support is to use our ABP estimator with the choices of $W = (2\log n)^{1/2} \cdot \Sigma$ for sub-Gaussian distributions (i.e., distributions bounded from above by a jointly Gaussian random variable) and $W = ((2\log n)^{1/2} + \log n) \cdot \Sigma$ for sub-exponential distributions (i.e., distributions with exponentially decaying tails), where $\Sigma = \mathbb{E}(W_i^{\mathsf{T}} W_i^{\mathsf{T}})$ is the covariance matrix of W_i. The reason for this suggested heuristic is these choices of W are analogous to bounds on the maximum expected values of sub-Gaussian and sub-exponential random variables [14].

Since the ABP estimator is a heuristic in this setting, an obvious topic is to design a statistically consistent estimator for inverse approximate optimization problems with unbounded noise. Maximum likelihood estimation is arguably the most natural approach because otherwise it is difficult to distinguish between noise and suboptimality of solutions. Specifically, consider the original problem setup but with the changes that the random sample is (u_i, X_i), the X_i are uniformly distributed within $S(u_i, \epsilon_0, \theta_0)$, and that W_i is distributed according to some known density $f_W(u)$. Then the maximum likelihood estimator (MLE) for this modified problem setup is given by

$$\begin{aligned}(\epsilon_{\text{mle}}, \theta_{\text{mle}}) \in \arg\min \Big\{ -\tfrac{1}{n}\sum_{i=1}^{n} \log \textstyle\int_{x\in S(u_i,\epsilon,\theta)} f_W(Y_i - x)dx \\ +\tfrac{1}{n}\sum_{i=1}^{n} \log \textstyle\int_{x\in S(u_i,\epsilon,\theta)} dx \;\Big|\; \epsilon \in E, \theta \in \Theta \Big\}. \end{aligned} \tag{38}$$

This optimization problem has a challenging structure in which the domains of integration depend upon the decision variables [41], and presents an opportunity for the further study of designing numerical algorithms to solve such optimization problems. We do note that for fixed (ϵ, θ), the integrals in the objective can be numerically computed in polynomial time using hit-and-run techniques for sampling from convex sets [29,45]. And so the enumeration algorithm we described earlier for the ABP estimator could be easily modified to solve this MLE problem.

Remark 6 The ABP and MLE estimators are actually qualitatively the same. The $\frac{1}{n}\sum_{i=1}^{n} d^2(Y_i, S(U_i, \epsilon, \theta) \oplus W)$ term in ABP and the $-\frac{1}{n}\sum_{i=1}^{n} \log \int_{x\in S(u_i,\epsilon,\theta)} f_W(Y_i - x)dx$ term in MLE both penalize estimates in which the solutions Y_i are far from the solution sets $S(\cdot, \epsilon, \theta)$, and the $\frac{1}{n}\sum_{i=1}^{n} \log \int_{x\in S(u_i,\epsilon,\theta)} dx$ term in MLE and the $\lambda \cdot \epsilon$ term in ABP both penalize estimates that generate large solution sets.

In the two inverse approximate optimization problem setups considered above, we assumed the approximate solutions X_i were drawn from the solution sets $S(U_i, \epsilon_0, \theta_0)$ according to some distribution. However, another modified problem setup would be to assume the X_i were chosen from the solution sets by solution of another optimization problem. This kind of setup corresponds to a scenario in which the X_i are solutions to an optimistic bilevel optimization problem with unique solutions:

$$x = \arg\min \big\{ F(u, x, z) \;\big|\; x \in S(u, \epsilon_0, \theta_0), G(u, x, z) \le 0 \big\}. \tag{39}$$

In this case, the estimation procedure can be posed as a least squares problem

$$\begin{aligned}(\epsilon_{\text{ble}}, \theta_{\text{ble}}) \in \arg\min\ & \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \|Y_i - x_i\|^2 \\ \text{s.t. } & x_i = \arg\min \big\{ F(U_i, x, z) \;\big|\; x \in S(U_i, \epsilon, \theta), G(U_i, x, z) \le 0 \big\} \\ & \epsilon \in E, \theta \in \Theta. \end{aligned} \tag{40}$$

This is a challenging multi-level optimization problem and presents an opportunity for the further study of designing numerical algorithms to solve such optimization problems. We do note that for fixed (ϵ, θ), this becomes a convex optimization problem. And so the enumeration algorithm we described earlier for the ABP estimator could be easily modified to solve this least squares problem.

6 Conclusion

In this paper, we used variational analysis to develop tools for statistics with set-valued functions, and then applied these tools to two estimation problems. We constructed and studied a kernel regression estimator for set-valued functions and an estimator for the inverse approximate optimization problem. The area of statistics with set-valued functions remains largely unexplored with many remaining problems. One question is the design of numerical representations of sets and set-valued functions. Though constraint representations of sets are pervasive, numerical machinery like epi-splines [40] may offer greater representational flexibility. Another question is the development of numerical algorithms to solve optimization problems that arise in statistical estimation for set-valued functions. Related inverse optimization problems lead to formulations (38) and (40) with structures that are not well-studied from the perspective of numerical optimization. Further study of statistics with set-valued functions will require developing new numerical methods and optimization theory.

References

1. Ahuja, R., Orlin, J.: Inverse optimization. Oper. Res. **49**(5), 771–783 (2001)
2. Allende, G., Still, G.: Solving bilevel programs with the KKT-approach. Math. Program. **138**(1–2), 309 (2013)
3. Artstein, Z., Vitale, R.: A strong law of large numbers for random compact sets. Ann. Probab. **3**(5), 879–882 (1975)
4. Aswani, A., Bickel, P., Tomlin, C.: Regression on manifolds: estimation of the exterior derivative. Ann. Stat. **39**(1), 48–81 (2011)
5. Aswani, A., Gonzalez, H., Sastry, S., Tomlin, C.: Provably safe and robust learning-based model predictive control. Automatica **49**(5), 1216–1226 (2013)
6. Aswani, A., Kaminsky, P., Mintz, Y., Flowers, E., Fukuoka, Y.: Behavioral modeling in weight loss interventions. Available at SSRN: https://ssrn.com/abstract=2838443. Accessed 22 Feb 2017 (2016)
7. Aswani, A., Shen, Z.J., Siddiq, A.: Inverse optimization with noisy data. Oper. Res. (2017) **(to appear)**
8. Aumann, R.J.: Integrals of set-valued functions. J. Math. Anal. Appl. **12**(1), 1–12 (1965)
9. Bahadur, R.R.: Some Limit Theorems in Statistics. SIAM, Philadelphia (1971)
10. Berge, C.: Topological Spaces: Including a Treatment of Multi-valued Functions, Vector Spaces, and Convexity. Courier Corporation, North Chelmsford (1963)
11. Bertsimas, D., Gupta, V., Paschalidis, I.: Data-driven estimation in equilibrium using inverse optimization. Math. Program. **153**(2), 595–633 (2015)
12. Bickel, P.: On adaptive estimation. Ann. Stat. **10**(3), 647–671 (1982)
13. Bickel, P., Doksum, K.: Mathematical Statistics: Basic Ideas And Selected Topics, vol. 1, 2nd edn. Pearson Prentice Hall, Upper Saddle River (2006)
14. Boucheron, S., Lugosi, G., Massart, P.: Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford University Press, Oxford (2013)
15. Castaing, C.: Sur les multi-applications mesurables. Revue française d'informatique et de recherche opérationnelle **1**(1), 91–126 (1967)
16. Chan, T., Craig, T., Lee, T., Sharpe, M.: Generalized inverse multiobjective optimization with application to cancer therapy. Oper. Res. **62**(3), 680–695 (2014)
17. Dempe, S., Dutta, J.: Is bilevel programming a special case of a mathematical program with complementarity constraints? Math. Program. **131**(1–2), 37–48 (2012)

18. Devroye, L., Wise, G.: Detection of abnormal behavior via nonparametric estimation of the support. SIAM J. Appl. Math. **38**(3), 480–488 (1980)
19. Dupacová, J., Wets, R.: Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. Ann. Stat. **16**(4), 1517–1549 (1988)
20. Esfahani, P.M., Shafieezadeh-Abadeh, S., Hanasusanto, G.A., Kuhn, D.: Data-driven inverse optimization with incomplete information. arXiv preprint arXiv:1512.05489 (2015)
21. Geffroy, J.: Sur un probleme destimation géométrique. Publ. Inst. Statist. Univ. Paris **13**, 191–210 (1964)
22. Geyer, C.J.: On the asymptotics of constrained M-estimation. Ann. Stat. **22**(4), 1993–2010 (1994)
23. Guntuboyina, A.: Optimal rates of convergence for convex set estimation from support functions. Ann. Stat. **40**(1), 385–411 (2012)
24. Keshavarz, A., Wang, Y., Boyd, S.: Imputing a convex objective function. In: IEEE Multi-Conference on Systems and Control, pp. 613–619 (2011)
25. Knight, K., Fu, W.: Asymptotics for lasso-type estimators. Ann. Stat. **28**(5), 1356–1378 (2000)
26. Korostelëv, A., Simar, L., Tsybakov, A.: Efficient estimation of monotone boundaries. Ann. Stat. 476–489 (1995)
27. Kudo, H.: Dependent experiments and sufficient statistics, vol. 4, pp. 905–927. Natural Science Report. Ochanomizu University (1953)
28. Lachout, P., Liebscher, E., Vogel, S.: Strong convergence of estimators as εn-minimisers of optimisation problemsof optimisation problems. Ann. Inst. Stat. Math. **57**(2), 291–313 (2005)
29. Lovász, L., Vempala, S.: Hit-and-run from a corner. SIAM J. Comput. **35**(4), 985–1005 (2006)
30. Matheron, G.: Random Sets and Integral Geometry. Wiley, Hoboken (1975)
31. Molchanov, I.: Theory of Random Sets. Springer, Berlin (2006)
32. Nesterov, Y., Nemirovskii, A.: Interior-Point Polynomial Algorithms in Convex Programming. SIAM, Philadelphia (1994)
33. Noda, K.: Estimation of a regression function by the Parzen kernel-type density estimators. Ann. Inst. Stat. Math. **28**(1), 221–234 (1976)
34. Ouattara, A., Aswani, A.: Duality approach to bilevel programs with a convex lower level. arXiv preprint arXiv:1608.03260 (2016)
35. Outrata, J.: On the numerical solution of a class of Stackelberg problems. Z. Oper. Res. **34**(4), 255–277 (1990)
36. Patschkowski, T., Rohde, A., et al.: Adaptation to lowest density regions with application to support recovery. Ann. Stat. **44**(1), 255–287 (2016)
37. Rényi, A., Sulanke, R.: Über die konvexe hülle von n zufällig gewählten punkten. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **2**(1), 75–84 (1963)
38. Rockafellar, R., Wets, R.: Variational Analysis, 3rd edn. Springer, Berlin (2009)
39. Royset, J.O.: Approximations and solution estimates in optimization. Math. Program. (2017)
40. Royset, J.O., Wets, R.: Multivariate epi-splines and evolving function identification problems. Set Valued Var. Anal. **24**(4), 517–545 (2016)
41. Royset, J.O., Wets, R.J.B.: Variational theory for optimization under stochastic ambiguity. SIAM J. Optim. **27**(2), 1118–1149 (2017)
42. Salinetti, G., Wets, R.: On the convergence in distribution of measurable multifunctions (random sets) normal integrands, stochastic processes and stochastic infima. Math. Oper. Res. **11**(3), 385–419 (1986)
43. Schneider, R.: Convex Bodies: The Brunn–Minkowski Theory 151. Cambridge University Press, Cambridge (1993)
44. Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., Williamson, R.: Estimating the support of a high-dimensional distribution. Neural Comput. **13**(7), 1443–1471 (2001)
45. Smith, R.: Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. Oper. Res. **32**(6), 1296–1308 (1984)
46. van der Vaart, A.: Asymptotic Statistics. Cambridge University Press, Cambridge (2000)
47. Wald, A.: Note on the consistency of the maximum likelihood estimate. Ann. Math. Stat. **20**(4), 595–601 (1949)
48. Wand, M., Jones, M.: Kernel Smoothing. Taylor & Francis, Abingdon (1994)
49. Weil, W.: An application of the central limit theorem for Banach-space-valued random variables to the theory of random sets. Probab. Theory Relat. Fields **60**(2), 203–208 (1982)
50. Ye, J., Zhu, D.: Optimality conditions for bilevel programming problems. Optimization **33**(1), 9–27 (1995)