Robust Out-of-Distribution Stochastic Optimization with Heterogeneous Inclusive and Exclusive Distribution Clusters

Biao Han

hanb16@tsinghua.org.cn

Abstract

Decision scenarios far from rare in practice often place data-driven decision-making in an awkward position, where the decision maker may have access to neither the target distribution itself nor any empirical samples drawn from it, especially when decisions must be made in novel or highly uncertain environments. In response, *robust out-of-distribution stochastic optimization* (RooDSO) has emerged as a new data-driven decision-making framework of significant practical relevance, which views the unknown target distribution as an independent realization of a meta-distribution and makes robust decisions using observations from the distribution clusters related to the unknown distribution. Notably, human knowledge about the *unknown* typically involves both *positive* and *negative* aspects: beyond identifying the inclusive distribution cluster with which the target distribution is compatible, one can often also recognize the exclusive clusters with which it is incompatible, and the latter are sometimes even easier to obtain than the former, and thus should likewise be exploited to inform decision-making. This paper formalizes this novel RooDSO decision scenario by incorporating both inclusive and exclusive distribution clusters and establishes a complete mathematical model for it. Since these clusters are typically multi-source and heterogeneous, we integrate multiple kernel learning into the decision framework to deeply extract useful information from both sides, and establish statistical guarantees for the proposed model together with efficient learning and optimization strategies. Moreover, the proposed model allows decision makers, by tuning hyperparameters, to flexibly incorporate into the decision process their emphasis on positive versus negative information, the degree of data heterogeneity, and their risk preferences.

1 Introduction

Real-world decision-making is inherently subject to uncertainty, and stochastic programming provides a standard framework for hedging against it. In its basic form, an (unconstrained) stochastic program reads

$$\min_{\mathbf{x}\in\mathcal{X}} \ \mathbb{E}_{\boldsymbol{\xi}\sim\mathbb{P}}\{f(\mathbf{x},\boldsymbol{\xi})\}, \tag{1.1}$$

where $\mathbf{x}\in\mathcal{X}$ is the decision, $\boldsymbol{\xi}$ the random uncertainty, f the cost, and $\mathbb{P}$ the distribution governing the uncertainty [22]. In most applications $\mathbb{P}$ is unknown and only empirical samples are available. The sample average approximation then replaces $\mathbb{P}$ by its empirical measure, whose in-sample

optimum is notoriously over-optimistic. Distributionally robust optimization (DRO) instead hedges against a family of distributions [12],

$$\min_{\mathbf{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{U}} \mathbb{E}_{\boldsymbol{\xi}\sim\mathbb{P}}\{f(\mathbf{x},\boldsymbol{\xi})\}, \tag{1.2}$$

where the ambiguity set $\mathcal{U}$ gathers distributions deemed close to the empirical one under, e.g., ϕ-divergences [2] or the Wasserstein metric [17]. Common to these paradigms, however, is the presumption that data drawn from the target distribution $\mathbb{P}$ are at hand.

Even where such data exist, an ambiguity set is only as good as its coverage: if the true distribution falls *outside* it, an instance of *misspecification* conceptually distinct from the ambiguity *within* the set, robustness guarantees can evaporate [15, 8]. Distinguishing these two layers of distributional uncertainty, and guarding against the second, has recently received explicit attention in data-driven optimization [15]. This distinction is especially acute in the present paper, because the distributions available to the decision maker are not sampled from the target but elicited from human judgment, so the ambiguity set constructed from them can systematically fail to cover the target. We return to this point in Section 2.1.

This paper concerns a strictly harder situation in which the target distribution $\mathbb{P}$ is *completely unobserved*: when the decision is to be made, neither $\mathbb{P}$ itself nor any empirical sample from it is available. Such *cold-start* situations are common in retail, manufacturing, and public health. For instance, to set the stock of a newly launched seasonal coffee product, a shop has no sales history of that product and can only refer to the sales records of related products. As a way out, one naturally wishes to leverage data from other, relevant distributions to inform decision-making under the unseen $\mathbb{P}$.

How should the "unknown" character of $\mathbb{P}$ be described? Alternative formalisms have been proposed. Fuzzy set theory models imprecision through membership functions [26], and Liu's uncertainty theory re-axiomatizes belief via subadditive uncertainty measures [14]. These frameworks are well suited to human-subjective indeterminacy, but the unknown in our problem is not merely a subjective quantity: it is a data-generating randomness that could in principle be observed, and about which we wish to learn from related sources. What is needed is a statistical model of *how the unknown distribution is generated*, together with out-of-sample guarantees for the resulting decisions, semantics that purely non-probabilistic calculi do not offer. We therefore stay within the orthodox probabilistic framework and instead endow the *unknownness itself* with a probabilistic description.

This is precisely the route taken by robust out-of-distribution stochastic optimization (RooDSO) [13], which views the unobserved target distribution $\mathbb{P}$ and each observed source distribution $\mathbb{P}_i$ as independent realizations of an unknown *meta-distribution* $\Psi \in \mathcal{P}(\mathcal{P}(\Xi))$, i.e., a distribution over distributions, a second-order belief in the terminology of the economics of ambiguity [6, 11]. Sampling from Ψ first draws a distribution $\mathbb{P}$, and then draws observations from $\mathbb{P}$. This *two-layer* randomness underpins all subsequent analysis. Embedding the realized source distributions into a

reproducing kernel Hilbert space (RKHS) through their kernel mean embeddings [1, 18], RooDSO learns, by support measure clustering [19], an ambiguity set that approximates the support of Ψ, and plugs it into (1.2) to obtain robust decisions, endowed with finite-sample out-of-distribution guarantees.

In an age increasingly driven by AI and automation, the experiential judgments that human intelligence forms about the world become all the more precious: they encode knowledge that no amount of historical data on the target itself can supply, precisely because that data does not exist. This paper takes the view that such judgments should be fully elicited, mined, and exploited rather than set aside. Human knowledge about the *unknown* typically has both a positive and a negative side: experts can often state what an unseen phenomenon "is like", but they are frequently even better at stating what it is "unlike" or "cannot be". Research in cognitive psychology has shown that negative and contrastive thinking is valuable in characterizing and defining unknown concepts: negative instances help learners reduce overgeneralization and delineate conceptual boundaries, and negative information tends to exert a deeper impact on cognition than positive information [23, 9]. In the contex of the paper, we use the term *inclusive* and *exclusive* to refer to the two sides. We note that exclusive knowledge also matters in RooDSO: unlike anomaly detection, where "normal" data are abundant and anomalies scarce, in the present distribution-level setting it is often difficult for experts to name enough inclusive clusters, since the unknown is after all rarely encountered, whereas exclusive examples are comparatively easy to produce. Therefore, in this paper we push the problem framework of RooDSO [13] one step further and study the following formalized problem.

Problem 1. *Let $\mathbb{P} \in \mathcal{P}(\Xi)$ be an unobserved target distribution. Available to the decision maker are heterogeneous empirical data from N^+* inclusive *distribution clusters $\{\mathbb{P}_i^+\}_{i=1}^{N^+}$, judged compatible with $\mathbb{P}$, and from N^-* exclusive *distribution clusters $\{\mathbb{P}_j^-\}_{j=1}^{N^-}$, judged incompatible with $\mathbb{P}$. Using* both *families of data, construct an ambiguity set $\mathcal{U} \subset \mathcal{P}(\Xi)$ that concentrates near the inclusive clusters while staying away from the exclusive ones, and return a decision $\mathbf{x}^\star \in \mathcal{X}$ that minimizes the worst-case expected cost*

$$\sup_{\mathbb{P}\in\mathcal{U}} \mathbb{E}_{\boldsymbol{\xi}\sim\mathbb{P}}\{f(\mathbf{x},\boldsymbol{\xi})\},$$

so that $\mathbf{x}^\star$ performs robustly when deployed under the unseen target distribution $\mathbb{P}$.

Motivation

The following observations motivate the model developed in this paper.

Inclusive-only information is insufficient. Consider again the coffee shop launching a new product in the high season. The demand distribution of the product is unknown, yet it is reasonable to assert that the demand will resemble that of similar coffees in the high season (an inclusive cluster), and that it will *not* resemble the demand of desserts in the high season, nor that of coffees and desserts in the low season (exclusive clusters). Both kinds of assertion constrain where the target distribution can plausibly lie, and should therefore both inform the ambiguity set. RooDSO [13] uses only the inclusive side.

Single-kernel embedding ignores heterogeneity. The relevant source distributions are typically heterogeneous. Dessert demand and coffee demand, and even different styles of coffee, may be intrinsically different objects, and viewing them as draws from one meta-distribution inside a *single* RKHS is questionable. A more faithful model should compose several feature spaces and let the data determine how they are weighted, that is, the representation should itself be *learned* via multiple kernel learning (MKL).

The SVDD-type geometry is ill posed under finite samples. RooDSO learns its ambiguity set with a classical SVDD-type ball [24]. Although accompanied by statistical guarantees, SVDD carries a subtle ill-posedness that is easy to overlook: under finite samples the learned optimal radius is not unique, in that different radii attain the same optimal value [25]. When such a ball is subsequently used for distributionally robust optimization, the conservatism of the resulting solution is therefore ambiguous. The convex SSLM reformulation [16] removes this ambiguity by admitting a provably unique optimal solution for the hyperparameters of interest.

Taken together, these observations motivate the systematic framework developed in this paper: we learn, in a *convex*, *well-posed*, and *decision-oriented* manner, an ambiguity set that simultaneously exploits inclusive and exclusive distribution clusters (Section 2.1) and adapts to their heterogeneity through multiple kernels (Section 2.2), and we then plug this set into the robust program (1.2) to make decisions under the unseen target distribution.

Contributions

Our main contributions are summarized as follows.

- **A new decision scenario and model (RooDSO-IEDC).** We formalize Problem 1 and instantiate it by learning the ambiguity set through a convex small-sphere–large-margin (SSLM) problem [16] on kernel mean embeddings, so that inclusive clusters are encircled while exclusive clusters are kept out. The model generalizes RooDSO [13], which is recovered in the inclusive-only case $N^- = 0$.

- **Well-posed uncertainty-set learning.** By adopting the convex SSLM geometry [16], the learned ambiguity set admits a provably *unique* optimal description (center and radius) under finite samples. This removes the ill-posedness of the SVDD-type ball [24] underlying RooDSO and makes the conservatism of the downstream robust solution well defined.

- **Heterogeneous clusters via multiple kernel learning (RooDSO-HIEDC).** To account for the intrinsic heterogeneity of multi-source clusters, we compose multiple RKHSs and learn their combination jointly with the ambiguity set, with RooDSO-IEDC emerging as the single-kernel special case.

- **Flexible risk posture.** The hyperparameters of the proposed model transparently encode the emphasis on positive versus negative information, the degree of data heterogeneity, and

the conservatism (risk preference) of the decision maker, providing a simple calibration handle in practice.

- **Learning and optimization.** We derive a dual formulation and present scalable learning and optimization strategies in which the GPU and parallel computing resources can be fully utilized when dealing with large-scale scenarios.

- **Statistical guarantees for the two-sided ambiguity set.** We analyse the properties of the proposed model in two steps: first quantifying the finite-sample stability of the learned description, and then establishing two-sided out-of-distribution guarantees, under which a new compatible distribution is covered while a new incompatible one is rejected.

Structure. The layout of this paper is structured as follows. Section 1 motivates the problem and summarizes our contributions. Section 2 develops the proposed methodology in two steps. Section 2.1 formalizes robust out-of-distribution stochastic optimization with *inclusive and exclusive* distribution clusters (RooDSO-IEDC) under a single kernel, warming up with the convex small-sphere–large-margin geometry and then deriving the uncertainty-set learning problem and the resultant robust program. Section 2.2 then extends the model to *heterogeneous* source distributions via multiple kernel learning (RooDSO-HIEDC). Section 3 presents the strategies of multiple kernel learning and the induced distributionally robust optimization. Section 4 provides statistical analyses of the proposed model. Section 5 closes the paper with concluding remarks. All technical proofs are relegated to Appendix A.

Notation. Let Ξ be a non-empty observation space and let $\mathcal{P}(\Xi)$ denote the set of all probability measures on it. We write $\xi_1, \ldots, \xi_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ to indicate that $\xi_1, \ldots, \xi_n$ are independent and identically distributed (i.i.d.) according to $\mathbb{P}$. For a positive integer N, we use $[N] := \{1, \ldots, N\}$ for the index set. For an event A, we write $\Pr_{\mathbb{P}}\{A\}$ for the probability of A under $\mathbb{P}$. A function $f : \mathcal{X} \to [-\infty, \infty]$ is *upper semicontinuous* on $\mathcal{X}$ if, for every $x_0 \in \mathcal{X}$, $\limsup_{x \to x_0} f(x) \leq f(x_0)$, and *lower semicontinuous* if the corresponding statement holds with lim inf. We denote by $\mathbf{1}_J \in \mathbb{R}^J$ the vector whose entries are all equal to one. For a symmetric matrix $\mathbf{K}$, $\lambda_{\max}(\mathbf{K})$ and $\lambda_{\min}(\mathbf{K})$ denote its largest and smallest eigenvalues, $\operatorname{diag}(\mathbf{K})$ the vector of its diagonal entries, and $\|\mathbf{K}\|_2 := \lambda_{\max}(\mathbf{K})$ its spectral (operator) norm. For a non-empty closed convex subset $\mathcal{M}$ of a Hilbert space $\mathcal{H}$, its support function is $h^*_{\mathcal{M}}(g) := \sup_{\mu \in \mathcal{M}} \langle g, \mu \rangle_{\mathcal{H}}$.

2 Proposed methodology

In this chapter we develop the proposed methodology. Following [13], we represent each probability distribution by a single point in a Hilbert space through its *kernel mean embedding* (KME). We first recall this embedding and its basic properties, since they underlie everything that follows.

Let $k : \Xi \times \Xi \to \mathbb{R}$ be a symmetric positive definite kernel, and let $\mathcal{H}$ be the associated reproducing kernel Hilbert space (RKHS) with canonical feature map $\phi : \Xi \to \mathcal{H}$, so that $k(\boldsymbol{\xi}, \boldsymbol{\xi}') =$

$\langle \phi(\boldsymbol{\xi}), \phi(\boldsymbol{\xi}') \rangle_{\mathcal{H}}$ and the reproducing property $\langle f, \phi(\boldsymbol{\xi}) \rangle_{\mathcal{H}} = f(\boldsymbol{\xi})$ holds for all $f \in \mathcal{H}$; see [1, 18] for a detailed exposition. Because point evaluation is a bounded linear functional, taking expectations lifts pointwise evaluation to distributions: whenever the integrability condition $\mathbb{E}_{\boldsymbol{\xi}\sim\mathbb{P}}\{\sqrt{k(\boldsymbol{\xi},\boldsymbol{\xi})}\} < \infty$ holds, the Bochner integral

$$\mu_{\mathbb{P}} \;:=\; \int_{\Xi} \phi(\boldsymbol{\xi})\, \mathrm{d}\mathbb{P}(\boldsymbol{\xi}) \in \mathcal{H} \tag{2.1}$$

is well defined, and it satisfies $\mathbb{E}_{\boldsymbol{\xi}\sim\mathbb{P}}\{f(\boldsymbol{\xi})\} = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$. In particular, inner products between distributions reduce to inner products in $\mathcal{H}$:

$$\langle \mu_{\mathbb{P}_1}, \mu_{\mathbb{P}_2} \rangle_{\mathcal{H}} = \int\!\!\int k(\boldsymbol{\xi}, \boldsymbol{\zeta})\, \mathrm{d}\mathbb{P}_1(\boldsymbol{\xi})\, \mathrm{d}\mathbb{P}_2(\boldsymbol{\zeta}), \tag{2.2}$$

so that the KME turns distributions into points of $\mathcal{H}$ while their cross-similarities are faithfully encoded by Hilbert-space inner products. If the kernel is *characteristic*, as are the Gaussian, Laplace and Matérn kernels, the map $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective and the representation is lossless [18]. Throughout, we impose only the mild regularity conditions collected in the following assumption, which are satisfied by the kernels above.

Assumption 1. *The kernel $k : \Xi \times \Xi \to \mathbb{R}$ is continuous, symmetric, and integrally strictly positive definite. Moreover, it is uniformly bounded on the diagonal, i.e.,* $\sup_{\boldsymbol{\xi}\in\Xi} k(\boldsymbol{\xi}, \boldsymbol{\xi}) = C < \infty$.

Whenever μ itself denotes a distribution under scrutiny, expressions such as $\langle \mu, \mu_{\mathbb{P}_i} \rangle_{\mathcal{H}}$ remain computable through (2.2), which is what ultimately renders everything kernelized. In the geometry below every object of interest is a *distribution*, viewed through (2.1) as a single point of $\mathcal{H}$. This development is deliberately general: the clusters are treated as (unknown) probability measures, and their data-driven, finite-sample discretization is deferred to Section 3, where the learning problem is actually solved.

2.1 RooDSO with inclusive and exclusive distribution clusters (RooDSO-IEDC)

We now attach a geometry to Problem 1. The target distribution $\mathbb{P}$ is unobserved, but it is believed to be *compatible* with the inclusive clusters $\{\mathbb{P}_i^+\}_{i=1}^{N^+}$ and *incompatible* with the exclusive clusters $\{\mathbb{P}_j^-\}_{j=1}^{N^-}$. Embedding every cluster by (2.1), we obtain two point clouds in $\mathcal{H}$,

$$\mathcal{V}^+ = \{\mu_{\mathbb{P}_i^+}\}_{i=1}^{N^+} \subset \mathcal{H}, \qquad \mathcal{V}^- = \{\mu_{\mathbb{P}_j^-}\}_{j=1}^{N^-} \subset \mathcal{H}, \tag{2.3}$$

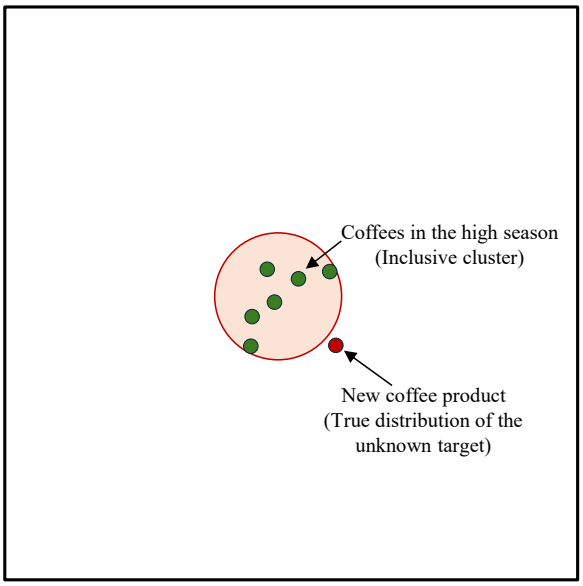
and the unknown embedding $\mu_{\mathbb{P}}$ of the target.

Our first modeling postulate is that, beyond the proximity of $\mu_{\mathbb{P}}$ to $\mathcal{V}^+$, no *directional* information about the target is available. It is then natural to describe the set of plausible locations of $\mu_{\mathbb{P}}$ as an *isotropic* perturbation region: a ball

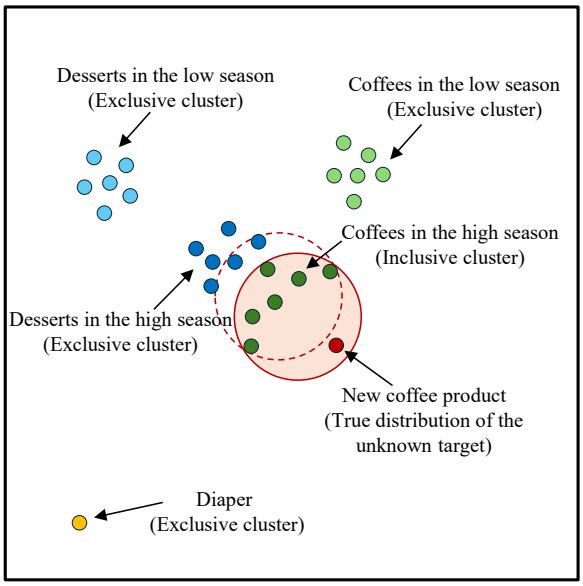
$$\mathcal{B}(\mu_c, R) = \{\mu \in \mathcal{H} : \|\mu - \mu_c\|_{\mathcal{H}} \le R\}$$

centered at some $\mu_c \in \mathcal{H}$, whose radius R gauges how far the target may deviate from the inclusive clusters in *any* direction.

This isotropic ansatz, however, collides with the way the data are produced. The clusters available to the decision maker are not the output of a balanced sampling scheme, but are *elicited from human judgment*. And human judgment of *similarity* is known to be subject to an effect called *cognitive* (or functional) fixedness: when asked what an unfamiliar object "is like", people tend to anchor on one salient dimension, or on a few strongly correlated dimensions, and to name instances that differ mostly along those dimensions. The inclusive cluster $\mathcal{V}^+$ therefore approaches the true location $\mu_{\mathbb{P}}$ from only a small set of (roughly monotone) directions and need not surround it evenly. Figure 1 illustrates the failure mode that results: a ball that merely *shrink-wraps* $\mathcal{V}^+$, that is, the classical SVDD-type description used in RooDSO [13, 24], leaves the directions along which no inclusive example was volunteered (the "angular blind spot" of the cluster) outside its boundary, and the true $\mu_{\mathbb{P}}$ may well fall there. In the terminology of the preceding section this is precisely an *ambiguity-set misspecification*: the ball fails to cover the truth [15, 8]. The exclusive clusters developed next act, in that language, as the correction mechanism, keeping the learned set from being misspecified on exactly the directions that the inclusive evidence leaves uncovered.



(a) Inclusive clusters alone: a ball that merely encloses them can leave the target distribution outside.



(b) Adding exclusive clusters: the geometry of the ball is corrected by repelling the exclusive distribution clusters from various directions.

Figure 1: Cognitive fixedness on the inclusive side versus multi-directional exclusive evidence. A ball enclosing the inclusive cluster alone can miss the unknown target location, whereas jointly repelling the exclusive clusters fence the ambiguous region from many sides.

By contrast, judgments of *dissimilarity* are far less constrained by fixedness: to articulate why an object is "unlike" the target, one is forced to scan across features and directions, so the resulting examples tend to be spread widely and multi-dimensionally. The exclusive clusters $\mathcal{V}^-$ therefore

carries complementary geometric content: it fences off, from many sides, the region in which $\mu_{\mathbb{P}}$ *cannot* lie, and thereby anchors the plausible region along the very directions that the inclusive cluster leaves unmeasured.

Combining the two sides, we seek a ball $\mathcal{B}(\mu_c, R)$ that is *as small as possible while enclosing the inclusive cluster*, so that it still covers the target's neighborhood, and, simultaneously, *as far as possible from the exclusive clusters*, so that it does not spill into regions already known to be incompatible with the target. This "enclose the positives tightly, keep away from the negatives" geometry is precisely the small-sphere–large-margin program of [16]. It brings a crucial extra dividend for our purpose: unlike the classical SVDD ball, the description it produces is convex and admits a provably *unique* optimal solution over the hyperparameter range of interest, removing the ill-posedness of the ball geometry discussed in the Motivation of Section 1.

To turn the geometry above into a concrete model, we instantiate the small-sphere–large-margin formulation (IEDC-$\mathcal{F}$) of [16] on the distribution embeddings (2.3): the inclusive embeddings act as the *positive* examples to be enclosed by the ball, and the exclusive embeddings as the *negative* examples to be repelled from it. Let $N := N^+ + N^-$ denote the total number of clusters, let $R \geq 0$ be the radius of the ball $\mathcal{B}(\mu_c, R)$, and let $t \geq 0$ be the margin measured in squared distance. We look for a ball whose interior is a plausible isotropic perturbation region for the target:

$$\inf_{\substack{\mu_c \in \mathcal{H} \\ R \geq 0,\ t \geq 0}} \quad \ell(\mu_c, R, t), \tag{IEDC-$\mathcal{F}$}$$

where the objective $\ell : \mathcal{H} \times \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ reads

$$\ell(\mu_c, R, t) = \frac{1}{2}(\nu R^2 - \eta\, t) + \frac{1}{2N} \sum_{i=1}^{N^+} \left(\|\mu_{\mathbb{P}_i^+} - \mu_c\|_{\mathcal{H}}^2 - R^2 \right)_+ + \frac{b}{2N} \sum_{j=1}^{N^-} \left(R^2 + t - \|\mu_{\mathbb{P}_j^-} - \mu_c\|_{\mathcal{H}}^2 \right)_+,$$

with $(z)_+ := \max\{z, 0\}$. The first hinge penalizes an inclusive embedding that escapes the ball, which should be enclosed since the target is believed to be compatible with the inclusive clusters, whereas the second hinge penalizes an exclusive embedding that intrudes into the ball inflated by the margin, which should be repelled since the target is believed to be incompatible with the exclusive clusters.

The hyperparameters have transparent roles. $\nu > 0$ trades the ball volume R^2 against the mis-enclosure of inclusive clusters (small sphere). $b > 0$ weighs the exclusive-side loss relative to the inclusive-side loss, which matters because the two sides carry asymmetric information in our setting. And $\eta \geq 0$ trades the margin against the volume (large margin). We write η, rather than the μ of [16], for the margin weight, since the letter μ is reserved throughout for kernel mean embeddings. Its admissible range is

$$\boxed{\eta \leq \min\left\{N^+/N,\ bN^-/N\right\},} \tag{2.4}$$

beyond which the objective of (IEDC-$\mathcal{F}$) is unbounded below [16].

Remark 1 (On the admissible range of η, and the choice of exclusive clusters)**.** *The bound* (2.4) *is intuitive: a decision maker should not repel the exclusive clusters she has selected too strongly, for if an exclusive cluster lies extremely far from the target it may in fact be so unrelated that it should arguably not have been used for training at all. This observation also suggests preferring, as training exclusive clusters, distributions that are* very similar yet clearly distinct *from the target, so as to carve out the boundary between "possible" and "impossible", or "certain" and "uncertain".*

As stated, (IEDC-$\mathcal{F}$) is nonconvex whenever exclusive clusters are present. [16] shows that, under the boundedness condition (2.4) and within the admissible regime

$$\boxed{\nu + \eta < N^+/N,} \tag{2.5}$$

by the variable change $s := \|\mu_c\|_{\mathcal{H}}^2 - R^2$ introduced in [16], the *global* optimum of (IEDC-$\mathcal{F}$) coincides with that of the following explicitly convex quadratic program

$$\begin{aligned} \min_{\substack{\mu_c \in \mathcal{H},\ s \in \mathbb{R},\ t \geq 0 \\ \boldsymbol{\omega}^+ \in \mathbb{R}_+^{N^+},\ \boldsymbol{\omega}^- \in \mathbb{R}_+^{N^-}}} \quad & \frac{N\nu}{2}\left(\|\mu_c\|_{\mathcal{H}}^2 - s\right) - \frac{N\eta}{2}t + \sum_{i=1}^{N^+} \omega_i^+ + b\sum_{j=1}^{N^-} \omega_j^- \\ \text{s.t.} \quad & e_i^+ \leq \omega_i^+, \quad i \in [N^+]; \qquad \frac{t}{2} - e_j^- \leq \omega_j^-, \quad j \in [N^-], \end{aligned} \tag{IEDC-$\mathcal{P}$}$$

where the hinge inputs are the *linear* functionals

$$e_i^+ := \frac{\|\mu_{\mathbb{P}_i^+}\|_{\mathcal{H}}^2 + s}{2} - \langle \mu_{\mathbb{P}_i^+}, \mu_c \rangle_{\mathcal{H}}, \qquad e_j^- := \frac{\|\mu_{\mathbb{P}_j^-}\|_{\mathcal{H}}^2 + s}{2} - \langle \mu_{\mathbb{P}_j^-}, \mu_c \rangle_{\mathcal{H}},$$

which has a provably unique solution on the hyperparameter range of interest, exactly the well-posedness we asked for, and in contrast to the non-uniqueness of the SVDD-type solution analyzed in [25].

By the representer theorem [10, 21, 16], the optimal center of (IEDC-$\mathcal{P}$) can be expanded over the distribution embeddings,

$$\mu_c = \sum_{i=1}^{N^+} \theta_i^+ \, \mu_{\mathbb{P}_i^+} + \sum_{j=1}^{N^-} \theta_j^- \, \mu_{\mathbb{P}_j^-},$$

with block coefficients $\boldsymbol{\theta}^+ \in \mathbb{R}^{N^+}$, $\boldsymbol{\theta}^- \in \mathbb{R}^{N^-}$ gathered as $\boldsymbol{\theta} = (\boldsymbol{\theta}^+; \boldsymbol{\theta}^-) \in \mathbb{R}^N$. Substituting this expansion into (IEDC-$\mathcal{P}$) and writing the Gram entries directly in terms of the distribution embeddings,

$$K_{ii'}^{++} := \langle \mu_{\mathbb{P}_i^+}, \mu_{\mathbb{P}_{i'}^+} \rangle_{\mathcal{H}}, \qquad K_{jj'}^{--} := \langle \mu_{\mathbb{P}_j^-}, \mu_{\mathbb{P}_{j'}^-} \rangle_{\mathcal{H}}, \qquad K_{ij}^{+-} := \langle \mu_{\mathbb{P}_i^+}, \mu_{\mathbb{P}_j^-} \rangle_{\mathcal{H}},$$

so that $\mathbf{K} := \begin{bmatrix} \mathbf{K}^{++} & \mathbf{K}^{+-} \\ (\mathbf{K}^{+-})^\top & \mathbf{K}^{--} \end{bmatrix}$ and $(\mathbf{K}\boldsymbol{\theta})_i = \langle \mu_{\mathbb{P}_i^+}, \mu_c \rangle_{\mathcal{H}}$, $(\mathbf{K}\boldsymbol{\theta})_{N^+ + j} = \langle \mu_{\mathbb{P}_j^-}, \mu_c \rangle_{\mathcal{H}}$, yields the

kernelized form, which we denote by (IEDC-$\mathcal{K}$) and state in the same split form as (IEDC-$\mathcal{P}$):

$$\begin{aligned}\min_{\boldsymbol{\theta}\in\mathbb{R}^N,\, s\in\mathbb{R},\, t\geq 0,\; \boldsymbol{\omega}^{\pm}\geq 0} \quad & \frac{N\nu}{2}\big(\boldsymbol{\theta}^\top \mathbf{K}\boldsymbol{\theta} - s\big) - \frac{N\eta}{2}t + \sum_{i=1}^{N^+}\omega_i^+ + b\sum_{j=1}^{N^-}\omega_j^- \\ \text{s.t.} \quad & \frac{K_{ii}+s}{2} - (\mathbf{K}\boldsymbol{\theta})_i \leq \omega_i^+, \quad i\in[N^+], \\ & \frac{t}{2} - \frac{K_{N^++j,\,N^++j}+s}{2} + (\mathbf{K}\boldsymbol{\theta})_{N^++j} \leq \omega_j^-, \quad j\in[N^-],\end{aligned} \tag{IEDC-$\mathcal{K}$}$$

where K_{ii} and K_{N^++j,N^++j} denote the diagonal entries of the corresponding blocks. This is the finite-dimensional, kernelized counterpart of (IEDC-$\mathcal{P}$).

Let $(\mu_c^\star, R^\star)$ be an optimal solution of (IEDC-$\mathcal{P}$), with the radius recovered from the change of variables as $(R^\star)^2 = \|\mu_c^\star\|_{\mathcal{H}}^2 - s^\star$. These two objects describe a ball in $\mathcal{H}$, our learned isotropic perturbation region for the unknown target embedding:

$$\mathcal{B}^{\mathrm{IEDC}}_{\nu,\eta,b} \;:=\; \big\{\mu\in\mathcal{H}:\; \|\mu-\mu_c^\star\|_{\mathcal{H}}\leq R^\star\big\}, \tag{2.6}$$

where the subscript records the hyperparameters by which the description is tuned. Because every distribution is identified with its kernel mean embedding, this ball induces an *out-of-distribution ambiguity set* for the unobserved target distribution $\mathbb{P}$,

$$\mathcal{U}^{\mathrm{IEDC}}_{\nu,\eta,b} \;:=\; \Big\{\mathbb{P}\in\mathcal{P}(\Xi):\; \mu_{\mathbb{P}} = \mathbb{E}_{\boldsymbol{\xi}\sim\mathbb{P}}\{\phi(\boldsymbol{\xi})\}\in\mathcal{B}^{\mathrm{IEDC}}_{\nu,\eta,b}\Big\}, \tag{2.7}$$

which gathers precisely the distributions compatible with the inclusive evidence, in the sense that their embeddings lie in the ball, while the geometry keeps the exclusive evidence at a distance. Substituting (2.7) into (1.2) yields the robustified stochastic program for RooDSO, which serves as the worst-case surrogate of Problem 1:

$$\min_{\mathbf{x}\in\mathcal{X}}\; \sup_{\mathbb{P}\in\mathcal{U}^{\mathrm{IEDC}}_{\nu,\eta,b}}\; \mathbb{E}_{\boldsymbol{\xi}\sim\mathbb{P}}\big\{f(\mathbf{x},\boldsymbol{\xi})\big\}. \tag{RooDSO-IEDC}$$

Note that, when $N^- = 0$ and $\eta = 0$, the exclusive term disappears and, up to a rescaling of hyperparameters, (IEDC-$\mathcal{F}$) reduces to the inclusive-only ball of RooDSO [13]. Our formulation therefore nests RooDSO as a special case.

2.2 RooDSO with heterogeneous IEDC (RooDSO-HIEDC)

When the decision maker exploits *both* the inclusive and the exclusive distribution clusters, heterogeneity becomes hard to avoid: the provenance of these clusters is far wider than that of the inclusive clusters used in conventional RooDSO, so the constituent distributions can rarely be trusted to be mutually comparable inside a single pre-specified reproducing kernel Hilbert space. Fixing one kernel in advance may therefore be a poor carrier of the evidence. Accordingly, rather than first embedding the clusters into a given Hilbert space and then learning an SSLM ball there,

as in Section 2.1, we propose to learn the ball and, *at the same time*, to search for a Hilbert space in which the learned ball encloses the inclusive clusters as tightly as possible while keeping the exclusive clusters far away. A natural device to realize such a search is *multiple kernel learning* (MKL), which leads to the heterogeneous model RooDSO-HIEDC developed in this subsection.

Suppose we are given M kernel functions $k_1, \dots, k_M$, each inducing an RKHS $\mathcal{H}_m$ with canonical feature map $\phi_m : \Xi \to \mathcal{H}_m$. Each kernel yields a distribution embedding $\mu_{m,\mathbb{P}} := \mathbb{E}_{\boldsymbol{\xi}\sim\mathbb{P}}\{\phi_m(\boldsymbol{\xi})\} \in \mathcal{H}_m$, and we combine the M spaces through a weight vector $\boldsymbol{\beta} \in \Delta_M := \{\boldsymbol{\beta} \in \mathbb{R}^M_+ : \sum_{m=1}^M \beta_m = 1\}$. Since this composite object is again a kernel mean embedding of $\mathbb{P}$ (now in the block space), we keep the μ-notation reserved for embeddings rather than the ϕ-notation of pointwise features, and write it as a block vector in the direct-sum space

$$\mu_{\boldsymbol{\beta},\mathbb{P}} \;:=\; \big(\sqrt{\beta_1}\,\mu_{1,\mathbb{P}},\;\dots,\;\sqrt{\beta_M}\,\mu_{M,\mathbb{P}}\big) \;\in\; \mathcal{H} := \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_M, \tag{2.8}$$

equipped with the componentwise inner product of the direct sum. In line with Section 2.1, distributional similarity is expressed through inner products of these embeddings rather than through a separate kernel: the composite inner product factors across the M spaces as

$$\langle \mu_{\boldsymbol{\beta},\mathbb{P}_1}, \mu_{\boldsymbol{\beta},\mathbb{P}_2}\rangle_{\mathcal{H}} = \sum_{m=1}^M \beta_m \,\langle \mu_{m,\mathbb{P}_1}, \mu_{m,\mathbb{P}_2}\rangle_{\mathcal{H}_m} = \sum_{m=1}^M \beta_m \int\!\!\int k_m(\boldsymbol{\xi},\boldsymbol{\zeta})\,\mathrm{d}\mathbb{P}_1(\boldsymbol{\xi})\,\mathrm{d}\mathbb{P}_2(\boldsymbol{\zeta}), \tag{2.9}$$

where the last equality expands each marginal inner product by (2.2). The composite feature space, and hence the similarity it induces, is thus itself part of what is learned.

The ball center is written consistently in block form, $\mu_c = \big(\sqrt{\beta_1}\,\mu_{c,1}, \dots, \sqrt{\beta_M}\,\mu_{c,M}\big)$ with blocks $\mu_{c,m} \in \mathcal{H}_m$. Its squared norm and the inner products against the composite embeddings factor across the kernels,

$$\|\mu_c\|^2_{\mathcal{H}} = \sum_{m=1}^M \beta_m \,\|\mu_{c,m}\|^2_{\mathcal{H}_m}, \qquad \langle \mu_{\boldsymbol{\beta},\mathbb{P}}, \mu_c\rangle_{\mathcal{H}} = \sum_{m=1}^M \beta_m \,\langle \mu_{m,\mathbb{P}}, \mu_{c,m}\rangle_{\mathcal{H}_m}, \tag{2.10}$$

and consequently the squared distance from a distribution embedding to the center is the weighted sum of the per-space distances, $\|\mu_{\boldsymbol{\beta},\mathbb{P}} - \mu_c\|^2_{\mathcal{H}} = \sum_{m=1}^M \beta_m \,\|\mu_{m,\mathbb{P}} - \mu_{c,m}\|^2_{\mathcal{H}_m}$.

Substituting (2.8)–(2.10) into the convex program (IEDC-$\mathcal{P}$) promotes the kernel weights to additional decision variables. To prevent the weights from collapsing onto a single kernel, which would silently reduce the model to its single-space special case of Section 2.1, we add the ℓ_2 regularizer $\lambda \sum_{m=1}^M \beta_m^2$ with $\lambda \geq 0$, which over the simplex is minimized at the uniform weights and thereby encourages genuinely multi-kernel representations. The resulting multi-kernel convex

SSLM primal of RooDSO-HIEDC reads

$$\min_{\substack{\boldsymbol{\beta}\in\Delta_M,\ \{\mu_{c,m}\in\mathcal{H}_m\}_{m=1}^M,\ s\in\mathbb{R},\ t\geq 0\\ \boldsymbol{\omega}^+\in\mathbb{R}_+^{N^+},\ \boldsymbol{\omega}^-\in\mathbb{R}_+^{N^-}}} \quad \frac{N\nu}{2}\Big(\sum_{m=1}^M \beta_m\|\mu_{c,m}\|_{\mathcal{H}_m}^2 - s\Big) - \frac{N\eta}{2}t + \lambda\sum_{m=1}^M \beta_m^2 + \sum_{i=1}^{N^+}\omega_i^+ + b\sum_{j=1}^{N^-}\omega_j^-$$

$$\text{s.t.} \quad e_i^+ \leq \omega_i^+, \quad i\in[N^+]; \qquad \frac{t}{2} - e_j^- \leq \omega_j^-, \quad j\in[N^-], \tag{HIEDC-$\mathcal{P}$}$$

where the hinge inputs factor across kernels,

$$e_i^+ := \frac{\|\mu_{\boldsymbol{\beta},\mathbb{P}_i^+}\|_{\mathcal{H}}^2 + s}{2} - \langle\mu_{\boldsymbol{\beta},\mathbb{P}_i^+},\mu_c\rangle_{\mathcal{H}} = \frac{1}{2}\Big(\sum_{m=1}^M \beta_m\|\mu_{m,\mathbb{P}_i^+}\|_{\mathcal{H}_m}^2 + s\Big) - \sum_{m=1}^M \beta_m\langle\mu_{m,\mathbb{P}_i^+},\mu_{c,m}\rangle_{\mathcal{H}_m},$$

and e_j^- is defined analogously with $\mathbb{P}_j^-$ in place of $\mathbb{P}_i^+$.

For a fixed weight vector $\boldsymbol{\beta}$, problem (HIEDC-$\mathcal{P}$) is exactly the convex program (IEDC-$\mathcal{P}$) posed in the composite space $\mathcal{H}$ (whose inner product is the composite one of (2.9)), so the well-posedness and uniqueness results of Section 2.1 transfer verbatim. Jointly over $(\boldsymbol{\beta},\mu_{c,1},\ldots,\mu_{c,M})$, however, the objective is not convex in general. We will discuss its solution in the following section. Let $(\mu_c^\star,R^\star)$ be an optimal solution of (HIEDC-$\mathcal{P}$), with the radius recovered from the change of variables as $(R^\star)^2 = \|\mu_c^\star\|_{\mathcal{H}}^2 - s^\star$, and let $\boldsymbol{\beta}^\star\in\Delta_M$ denote the optimal kernel weights returned by the same problem. These objects describe a ball in the composite space, our learned isotropic perturbation region for the unknown target embedding:

$$\mathcal{B}_{\nu,\eta,b,\lambda}^{\text{HIEDC}} := \big\{\mu\in\mathcal{H}:\ \|\mu-\mu_c^\star\|_{\mathcal{H}}\leq R^\star\big\}, \tag{2.11}$$

where the subscript records the hyperparameters that shape the description, while the learned weights $\boldsymbol{\beta}^\star$ enter through the space itself. Because every distribution is identified with its composite embedding (2.8) evaluated at $\boldsymbol{\beta}^\star$, this ball induces an *out-of-distribution ambiguity set* for the unobserved target distribution $\mathbb{P}$,

$$\mathcal{U}_{\nu,\eta,b,\lambda}^{\text{HIEDC}} := \Big\{\mathbb{P}\in\mathcal{P}(\Xi):\ \mu_{\boldsymbol{\beta}^\star,\mathbb{P}}\in\mathcal{B}_{\nu,\eta,b,\lambda}^{\text{HIEDC}}\Big\}, \tag{2.12}$$

and substituting it into (1.2) gives the heterogeneous robustified program

$$\min_{\mathbf{x}\in\mathcal{X}}\ \sup_{\mathbb{P}\in\mathcal{U}_{\nu,\eta,b,\lambda}^{\text{HIEDC}}}\ \mathbb{E}_{\boldsymbol{\xi}\sim\mathbb{P}}\big\{f(\mathbf{x},\boldsymbol{\xi})\big\}. \tag{RooDSO-HIEDC}$$

Observe that (RooDSO-HIEDC) genuinely contains the single-kernel model (RooDSO-IEDC): when only one kernel is used ($M=1$, $\boldsymbol{\beta}=1$), the composite space $\mathcal{H}$ is just the RKHS $\mathcal{H}$, the composite embedding $\mu_{\boldsymbol{\beta},\mathbb{P}}$ coincides with $\mu_{\mathbb{P}}$, the ℓ_2 penalty $\lambda\sum_m\beta_m^2=\lambda$ becomes a constant, and (HIEDC-$\mathcal{P}$) reduces to the convex program (IEDC-$\mathcal{P}$). Consequently (2.11)–(2.12) and (RooDSO-HIEDC) revert to their counterparts (2.6)–(2.7) and (RooDSO-IEDC). The same reduc-

tion holds whenever the learned weights concentrate on a single kernel (a coordinate of $\boldsymbol{\beta}$ equal to one), which the ℓ_2 term discourages but $\lambda = 0$ permits. RooDSO-IEDC is therefore recovered as the single-kernel special case of RooDSO-HIEDC.

The Gram closed form of (HIEDC-$\mathcal{P}$), a practical scheme for the kernel weights, and the solution of (RooDSO-HIEDC) are developed in Section 3.

2.3 Interpretation of the hyperparameters

We close this section with a closer look at the roles of the hyperparameters ν, η, b, λ. Two complementary readings are useful: a *qualitative* one, which links each parameter to a modeling choice available to the decision maker (Remark 2), and a *quantitative* one, inherited from the ν-property of the underlying convex SSLM [16], which ties the parameters to the fractions of clusters that the learned description keeps, rejects, or leaves exactly at the boundary.

Remark 2 (A qualitative reading of the hyperparameters)**.** *The four hyperparameters of RooDSO-HIEDC have transparent operational meanings, which is one of the attractions of the formulation.*

- $\nu > 0$ *balances the volume* R^2 *of the ball against the requirement that the inclusive clusters be enclosed. A larger* ν *drives a smaller, more compact description, hence a less conservative ambiguity set, at the price of possibly leaving part of the inclusive evidence outside the ball. A smaller* ν *favors covering the inclusive clusters even when the ball grows. It can be read as the decision maker's trust in, and required use of, the* inclusive *(positive) evidence.*

- $b > 0$ *weighs the exclusive-side losses against the inclusive-side ones: increasing* b *forces the description to push the exclusive clusters farther away, encoding stronger reliance on the* exclusive *(negative) evidence.*

- $\eta \geq 0$ *rewards a large margin between the ball and the exclusive clusters. Together with* ν*, it sets how the "plausible but to be avoided" region is carved out, so* (ν, η, b) *jointly encode the decision maker's* risk posture*: how tightly the description hugs the inclusive evidence and how decisively the exclusive evidence is rejected.*

- $\lambda \geq 0$ *penalizes* $\|\boldsymbol{\beta}\|_2^2$ *on the simplex and thereby shapes the* heterogeneity *of the representation.* $\lambda = 0$ *permits the weights to collapse onto a single best kernel, while a larger* λ *drives* $\boldsymbol{\beta}$ *toward the uniform weights and forces the description to exploit several feature spaces jointly. It is the dial that controls how much heterogeneity the model is allowed to use.*

These roles claim that the model's hyperparameters flexibly encode the emphasis on positive versus negative information, the degree of data heterogeneity, and the risk preference of the decision maker. In practice they are selected by validation on held-out cluster data or by expert prior knowledge, subject to the well-posedness constraints (2.4)–(2.5) *of Section 2.1.*

For the quantitative counterpart we transplant to the cluster setting the ν-property proved in [16, Proposition 14]. Fix an optimal solution $(\mu_c^\star, R^\star, t^\star)$ of (IEDC-$\mathcal{P}$) (for fixed weights, likewise of

(HIEDC-$\mathcal{P}$)). Call an *inclusive probability measure* $\mathbb{P}_i^+$ a *support measure* if $\|\mu_{\mathbb{P}_i^+} - \mu_c^\star\|_{\mathcal{H}}^2 \geq (R^\star)^2$, and a *margin-error measure* if the inequality is strict. Likewise, call an *exclusive probability measure* $\mathbb{P}_j^-$ a support measure if $\|\mu_{\mathbb{P}_j^-} - \mu_c^\star\|_{\mathcal{H}}^2 \leq (R^\star)^2 + t^\star$, and a margin-error measure if it lies strictly inside this ball. Let ME^+, SM^+ denote the numbers of inclusive margin errors and support measures, and ME^-, SM^- the numbers of exclusive margin errors and support measures, respectively.

Lemma 1 (ν-property of the IEDC and HIEDC balls; cf. Proposition 14 of [16])**.** *Suppose the admissible regime (2.4)–(2.5) holds. If the optimal margin is positive ($t^\star > 0$), then*

$$\frac{ME^+}{N} \leq \nu + \eta \leq \frac{SM^+}{N}, \qquad \frac{ME^-}{N} \leq \frac{\eta}{b} \leq \frac{SM^-}{N}; \tag{2.13}$$

otherwise ($t^\star = 0$),

$$\frac{ME^+}{N} \leq \nu + b\,\frac{SM^-}{N}, \qquad \nu + b\,\frac{ME^-}{N} \leq \frac{SM^+}{N}, \qquad \frac{\eta}{b} \leq \frac{SM^-}{N}. \tag{2.14}$$

The lemma is Proposition 14 of [16] read at the level of probability distribution embeddings: because the geometry and the optimality conditions of (IEDC-$\mathcal{P}$) coincide, for every fixed kernel weight, with those of the convex SSLM treated there, the proof carries over verbatim. Its content is a family of "fraction certificates" that link the model parameters to the fitted description, and it suggests several practically useful readings.

Remark 3 (A quantitative reading: fraction certificates and practice)**.** *The lemma attaches concrete, data-dependent meaning to the hyperparameters: it says how many of the elicited clusters the fitted description keeps, rejects, or leaves exactly at the boundary. We spell out the consequences that are most useful when calibrating the model and when judging whether the elicited clusters are informative.*

Budget form. *Setting $\kappa^+ := \nu + \eta$ and $\kappa^- := \eta/b$, the positive-margin case (2.13) reads $ME^+/N \leq \kappa^+ \leq SM^+/N$ and $ME^-/N \leq \kappa^- \leq SM^-/N$: κ^+ guarantees that at most a fraction κ^+ of the inclusive distributions are sacrificed while at least a fraction κ^+ are used as supports, and κ^- plays the mirror role on the exclusive side. These two "budgets" are the most direct quantities to calibrate: choosing them is equivalent to choosing how many inclusive distributions may be dropped and how many exclusive distributions may intrude into the margin.*

Coupling and decoupling. *Since η enters both κ^+ and κ^-, adjusting η alone moves the two sides simultaneously, whereas ν and b act on a single side: increasing ν raises only κ^+ (stronger insistence on covering inclusive evidence), and increasing b lowers only κ^- (more decisive rejection of exclusive evidence). This makes precise which knob implements the "emphasis on positive versus negative information".*

Sparsity and interpretability. *Only support measures influence the optimal description: measures strictly inside the ball or strictly beyond the outer sphere carry a zero dual multiplier and can be removed without changing the ambiguity set. Moreover $SM^+ \geq N\kappa^+$, so the description always rests on a controlled number of inclusive supports. The support set is an inspectable* certificate *of*

which elicited distributions actually matter, which is useful for auditing the human input discussed in Section 2.1.

The sign of $t^\star$ as a separability diagnostic. *$t^\star > 0$ means the exclusive evidence is kept out with a strict margin and the strong form (2.13) holds. By contrast, $t^\star = 0$ signals that the geometry saturates on the exclusive evidence (some exclusive cluster pins the outer sphere), so that only the weaker (2.14) applies. In the saturated regime the marginal value of adding yet more exclusive clusters, or of further increasing b, decreases. A persistent $t^\star = 0$ then suggests revisiting the kernels (hence the learned weights $\boldsymbol{\beta}$) or the elicited clusters instead.*

Unification of the classical ν-property. *When there are no exclusive clusters ($N^- = 0$, $\eta = 0$), (2.13) reduces to $ME^+/N^+ \leq \nu \leq SM^+/N^+$: ν is an upper bound on the fraction of inclusive clusters left out and a lower bound on the fraction kept as supports. This is exactly the ν-property used by RooDSO [13], so Lemma 1 interpolates between that classical statement and the two-sided setting introduced here.*

A quantitative check of elicitation quality. *Because $ME^-/N \leq \eta/b$, an unexpectedly large observed fraction of exclusive clusters intruding under a small prescribed η/b indicates that these clusters lie too close to the plausible region to be rejected cleanly, which quantifies the intuition, noted in Remark 1, that one should prefer exclusive clusters that are "very similar yet clearly distinct" from the target. Symmetrically, a large ME^+/N relative to κ^+ signals that inclusive clusters are being sacrificed more than intended, pointing to over-elicitation or to inclusive clusters that were mislabelled as compatible.*

3 Learning and optimization strategies

In this section we discuss how the ambiguity sets of Section 2 are learned from data and how the resulting robustified decision problems are solved. For concreteness we develop the methodology around the multiple-kernel model RooDSO-HIEDC of Section 2.2, the most comprehensive instance considered in this paper. RooDSO-IEDC is its single-kernel special case ($M = 1$ and $\boldsymbol{\beta} = 1$, see the nesting discussion in Section 2.2), and every step below specializes to it immediately by taking a single kernel. We therefore do not repeat the single-kernel derivations separately.

3.1 Learning the ambiguity set

The construction of Section 2.2 is purely geometric: it takes the cluster probability measures as given and defines the ambiguity set through the solution of (HIEDC-$\mathcal{P}$). Turning this into a procedure that runs on data requires two ingredients: a finite-dimensional, tractable form of (HIEDC-$\mathcal{P}$) together with an algorithm for its kernel weights, and a data-driven estimate of the Gram matrices that feed it. For clarity we present the first ingredient in terms of a generic Gram matrix, and only afterwards, at the end of this subsection, describe how that Gram is obtained from samples.

Representer and the finite-dimensional inner problem. For a fixed weight $\boldsymbol{\beta}$, the inner problem is (IEDC-$\mathcal{P}$) posed in the composite space $\mathcal{H}$, whose data are the composite distribution embeddings

$\mu_{\boldsymbol{\beta},\mathbb{P}_l}$ with Gram matrix

$$\mathbf{K}_{\boldsymbol{\beta}} := \sum_{m=1}^{M} \beta_m \mathbf{K}_m \ \in \ \mathbb{R}^{N\times N}.$$

By the representer theorem [10, 21, 16], the center then admits the expansion

$$\mu_c = \sum_{i=1}^{N^+} \theta_i \, \mu_{\boldsymbol{\beta},\mathbb{P}_i^+} + \sum_{j=1}^{N^-} \theta_{N^++j} \, \mu_{\boldsymbol{\beta},\mathbb{P}_j^-} \ = \ \sum_{k=1}^{N} \theta_k \, \mu_{\boldsymbol{\beta},\mathbb{P}_k}, \qquad \boldsymbol{\theta} \in \mathbb{R}^N,$$

and substituting it into (IEDC-$\mathcal{P}$) yields exactly the kernelized program (IEDC-$\mathcal{K}$) of Section 2.1, with its Gram matrix equal to $\mathbf{K}_{\boldsymbol{\beta}}$. We therefore define, rather than restate, the inner value function through that program:

$$\Omega(\boldsymbol{\beta}) \ := \ \big\{ \text{the optimal value of (IEDC-}\mathcal{K}\text{) with } \mathbf{K} = \mathbf{K}_{\boldsymbol{\beta}} \big\},$$

i.e., the best ball attainable in the composite space for a given weight vector. For $M = 1$, $\mathbf{K}_{\boldsymbol{\beta}} = \mathbf{K}_1$, so $\Omega(\boldsymbol{\beta})$ reduces to the optimal value of (IEDC-$\mathcal{P}$) (equivalently of (IEDC-$\mathcal{K}$) on the single kernel), and the outer problem below disappears.

Learning the kernel weights. We first clarify the convexity of the joint problem (HIEDC-$\mathcal{P}$), which is *biconvex* but not *jointly convex*. Every volume term and hinge input multiplies a weight β_m by a quantity built from the corresponding center block and the cluster data. Fixing either block therefore leaves a convex problem in the other, namely a convex quadratic program in the geometric variables $(\mu_{c,1}, \ldots, \mu_{c,M};\ s, t, \boldsymbol{\omega})$ when $\boldsymbol{\beta}$ is held fixed, and a linear, hence convex, function of $\boldsymbol{\beta}$ when the geometric variables are held fixed. Joint convexity nevertheless fails, because these products couple the two blocks. What makes the problem tractable is the reduced view in which the geometric variables have been eliminated, since in that view $\Omega(\boldsymbol{\beta})$ is convex in $\boldsymbol{\beta}$. Indeed, by strong duality it is the maximum of functions that are affine in $\boldsymbol{\beta}$, taken over multipliers whose feasible set does not depend on $\boldsymbol{\beta}$. Hence the two-level program

$$\min_{\boldsymbol{\beta}\in\Delta_M} \ \Omega(\boldsymbol{\beta}) + \lambda \sum_{m=1}^{M} \beta_m^2 \tag{3.1}$$

is convex on the simplex, and we minimize it by alternating an inner solve with a weight update: starting from the uniform weights $\boldsymbol{\beta}^{(0)} = \mathbf{1}/M$, (i) solve the inner QP at the current $\boldsymbol{\beta}$ to obtain the ball and its optimal multipliers; (ii) read the gradient of Ω off these multipliers by the envelope (Danskin) theorem [5, 3], i.e., each coordinate $\partial\Omega/\partial\beta_m$ aggregates the kernel-m Gram terms weighted by the multipliers; and (iii) take a projected-gradient step onto the simplex,

$$\boldsymbol{\beta} \ \leftarrow \ \mathrm{proj}_{\Delta_M}\big(\boldsymbol{\beta} - \gamma\big(\nabla\Omega(\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta}\big)\big), \tag{3.2}$$

where $\Delta_M := \{\boldsymbol{\beta} \in \mathbb{R}_+^M : \ \sum_{m=1}^M \beta_m = 1\}$ is the probability simplex and proj_{Δ_M} is its Euclidean projection, which admits a closed form. Projection is applied because a plain gradient step would

leave the simplex. Newton-type updates are not used because Ω is a maximum of functions affine in $\boldsymbol{\beta}$ and is therefore generally nonsmooth, so that only (sub)gradients are reliably available. Because the outer problem is convex, these projected-subgradient iterations converge to a global minimizer of (3.1) for any step sizes satisfying the Robbins–Monro conditions [20], $\sum_k \gamma_k = \infty$ and $\sum_k \gamma_k^2 < \infty$. A convenient choice is

$$\gamma_k = \gamma_0 \, (k+1)^{-1/2}, \tag{3.3}$$

and each outer step costs a single inner QP solve. The ℓ_2 term prevents the weights from collapsing onto a single kernel.

Inner problem and its dual. The inner problem at fixed $\boldsymbol{\beta}$ is the composite-space instance of (IEDC-$\mathcal{P}$), and it is solved most transparently through its Lagrange dual. Introducing multipliers $\alpha_i^+ \geq 0$, $\alpha_j^- \geq 0$ for the two constraint families of (IEDC-$\mathcal{P}$) and eliminating the primal variables gives a convex QP over the box $[0,1]^{N^+} \times [0,b]^{N^-}$, with two additional linear constraints,

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & -\frac{1}{2N\nu}\,\boldsymbol{\alpha}^\top \mathbf{K}_{\boldsymbol{\beta}}^{\pm}\,\boldsymbol{\alpha} + \frac{1}{2}\,\mathbf{d}_{\boldsymbol{\beta}}^\top \boldsymbol{\alpha} \\ \text{s.t.} \quad & 0 \leq \alpha_i^+ \leq 1, \quad i \in [N^+], \qquad 0 \leq \alpha_j^- \leq b, \quad j \in [N^-], \\ & \sum_{i=1}^{N^+} \alpha_i^+ - \sum_{j=1}^{N^-} \alpha_j^- = N\nu, \qquad \sum_{j=1}^{N^-} \alpha_j^- \geq N\eta, \end{aligned} \tag{3.4}$$

with $\boldsymbol{\alpha} = (\boldsymbol{\alpha}^+; \boldsymbol{\alpha}^-) \in \mathbb{R}^N$. Here $\mathbf{K}_{\boldsymbol{\beta}}^{\pm}$ is the *signed* composite Gram of the distribution embeddings, obtained from the composite Gram $\mathbf{K}_{\boldsymbol{\beta}}$ by negating its cross blocks, i.e.,

$$\mathbf{K}_{\boldsymbol{\beta}}^{\pm} = \begin{bmatrix} \mathbf{K}_{\boldsymbol{\beta}}^{++} & -\mathbf{K}_{\boldsymbol{\beta}}^{+-} \\ -(\mathbf{K}_{\boldsymbol{\beta}}^{+-})^\top & \mathbf{K}_{\boldsymbol{\beta}}^{--} \end{bmatrix},$$

and the linear coefficient collects the diagonal terms, $\mathbf{d}_{\boldsymbol{\beta}} = \big(\operatorname{diag} \mathbf{K}_{\boldsymbol{\beta}}^{++}; -\operatorname{diag} \mathbf{K}_{\boldsymbol{\beta}}^{--}\big)$. This is the composite-kernel analogue of the dual of (IEDC-$\mathcal{K}$). Primal stationarity recovers the center in composite form,

$$\mu_c = \frac{1}{N\nu}\Big(\sum_{i=1}^{N^+} \alpha_i^+ \, \mu_{\boldsymbol{\beta},\mathbb{P}_i^+} - \sum_{j=1}^{N^-} \alpha_j^- \, \mu_{\boldsymbol{\beta},\mathbb{P}_j^-} \Big), \tag{3.5}$$

i.e., the representer emerges from the multipliers. Recalling the kernelized expansion (IEDC-$\mathcal{K}$), $\mu_c = \sum_{k=1}^N \theta_k \, \mu_{\boldsymbol{\beta},\mathbb{P}_k}$ with representer coefficients $\boldsymbol{\theta} \in \mathbb{R}^N$, these multipliers determine the representer coefficients through $\boldsymbol{\theta} = \frac{1}{N\nu}\,\mathbf{D}\,\boldsymbol{\alpha}$, where $\mathbf{D} = \operatorname{diag}(\mathbf{I}_{N^+}, -\mathbf{I}_{N^-}) \in \mathbb{R}^{N \times N}$ is the sign matrix that flips the sign of the exclusive block, $\mathbf{D}\boldsymbol{\alpha} = (\boldsymbol{\alpha}^+; -\boldsymbol{\alpha}^-)$—the same sign pattern that negates the exclusive rows and columns of the signed Gram (3.4). Componentwise, $\theta_i = \alpha_i^+/(N\nu)$ for the inclusive clusters and $\theta_{N^+ + j} = -\alpha_j^-/(N\nu)$ for the exclusive ones.

Strong duality identifies $\Omega(\boldsymbol{\beta})$ with the optimal value of (3.4), whose objective is affine in the weights while its feasible set is not: because $\mathbf{K}_{\boldsymbol{\beta}} = \sum_{m=1}^M \beta_m \mathbf{K}_m$, both the signed Gram and the diagonal vector in (3.4) are β-weighted sums of their single-kernel counterparts, $\mathbf{K}_{\boldsymbol{\beta}}^{\pm} = \sum_{m=1}^M \beta_m \mathbf{K}_m^{\pm}$

and $\mathbf{d}_{\boldsymbol{\beta}} = \sum_{m=1}^{M} \beta_m \mathbf{d}_m$. The envelope (Danskin) theorem therefore differentiates the dual optimum through the multipliers, and at an optimal $\boldsymbol{\alpha}^\star$ it isolates the kernel-m contribution,

$$\frac{\partial \Omega}{\partial \beta_m} = -\frac{1}{2N\nu}\, \boldsymbol{\alpha}^{\star\top} \mathbf{K}_m^{\pm} \boldsymbol{\alpha}^\star + \frac{1}{2}\, \mathbf{d}_m^\top \boldsymbol{\alpha}^\star, \tag{3.6}$$

where $\mathbf{K}_m^{\pm}$ and $\mathbf{d}_m$ are the signed Gram and the diagonal vector built from the single-kernel Gram $\mathbf{K}_m$. This is the gradient used in the update (3.2). The multipliers also carry the support structure underlying Lemma 1: a strictly positive $\alpha_i^{\star+}$ or $\alpha_j^{\star-}$ marks a probability measure that is active in the fitted description, while measures with a zero multiplier can be removed without changing the ball.

The overall procedure is summarized in Algorithm 1.

Algorithm 1 Learning the ambiguity set of RooDSO-HIEDC

Input: the Gram matrices $\mathbf{K}_1, \ldots, \mathbf{K}_M$ of the distribution embeddings; hyperparameters ν, η, b, λ; step sizes $\{\gamma_k\}$. (When only samples are available, these Gram matrices are replaced by their empirical estimates described at the end of this subsection.)

Output: kernel weights $\boldsymbol{\beta}^\star$, ball $(\mu_c^\star, R^\star)$, and ambiguity set $\mathcal{U}^{\mathrm{HIEDC}}_{\nu,\eta,b,\lambda}$.

1: set $\boldsymbol{\beta}^{(0)} \leftarrow (\frac{1}{M}, \ldots, \frac{1}{M})$;
2: **for** $k = 0, 1, 2, \ldots$ until convergence **do**
3: solve the inner dual (3.4) at $\boldsymbol{\beta} = \boldsymbol{\beta}^{(k)}$; obtain $(\mu_c^{(k)}, R^{(k)})$ and the multipliers $\boldsymbol{\alpha}^{\star,(k)}$;
4: compute the gradient $\nabla\Omega(\boldsymbol{\beta}^{(k)})$ by (3.6);
5: update $\boldsymbol{\beta}^{(k+1)} \leftarrow \mathrm{proj}_{\Delta_M}\big(\boldsymbol{\beta}^{(k)} - \gamma_k\big(\nabla\Omega(\boldsymbol{\beta}^{(k)}) + 2\lambda\boldsymbol{\beta}^{(k)}\big)\big)$;
6: **end for**
7: set $\boldsymbol{\beta}^\star \leftarrow \boldsymbol{\beta}^{(K)}$, re-solve the inner problem at $\boldsymbol{\beta}^\star$, and assemble the ball and $\mathcal{U}^{\mathrm{HIEDC}}_{\nu,\eta,b,\lambda}$.

For the single-kernel model RooDSO-IEDC ($M = 1$) Steps 2–5 of Algorithm 1 are skipped: $\boldsymbol{\beta} = 1$ is fixed and only one inner solve of (IEDC-$\mathcal{P}$) (in its dual form (3.4)) is needed. Solving (3.1) (and re-solving the inner QP at the returned $\boldsymbol{\beta}^\star$) yields the ball $\mathcal{B}^{\mathrm{HIEDC}}_{\nu,\eta,b,\lambda}$ and hence the ambiguity set $\mathcal{U}^{\mathrm{HIEDC}}_{\nu,\eta,b,\lambda}$ of Section 2.2. The support/margin-error structure of the solution also yields, through Lemma 1, the diagnostics discussed in Section 2.3. The solution of the robustified program (RooDSO-HIEDC), which reuses the multipliers obtained here, is developed next.

Data-driven approximation of the Gram matrices. So far the inclusive and exclusive cluster probability measures have been treated as known, with all inner products computed from them via (2.2). In practice these measures are never given explicitly: what is observed is, for each *inclusive* cluster $\mathbb{P}_i^+$, a dataset $\mathcal{D}_i^+ = \{\boldsymbol{\xi}_1^{+(i)}, \ldots, \boldsymbol{\xi}_{n_i^+}^{+(i)}\}$ of n_i^+ i.i.d. draws from $\mathbb{P}_i^+$, and, for each *exclusive* cluster $\mathbb{P}_j^-$, a dataset $\mathcal{D}_j^- = \{\boldsymbol{\xi}_1^{-(j)}, \ldots, \boldsymbol{\xi}_{n_j^-}^{-(j)}\}$ of n_j^- i.i.d. draws from $\mathbb{P}_j^-$, all datasets being independent. (The unobserved target distribution $\mathbb{P}$ is never sampled and plays no role in these approximations.) Their empirical measures are $\hat{\mathbb{P}}_i^+ = \frac{1}{n_i^+}\sum_u \delta_{\boldsymbol{\xi}_u^{+(i)}}$ and $\hat{\mathbb{P}}_j^- = \frac{1}{n_j^-}\sum_v \delta_{\boldsymbol{\xi}_v^{-(j)}}$,

and, for each kernel k_m, the corresponding empirical mean embeddings are

$$\hat{\mu}_{m,\mathbb{P}_i^+} = \frac{1}{n_i^+}\sum_{u=1}^{n_i^+}\phi_m(\boldsymbol{\xi}_u^{+(i)}) \in \mathcal{H}_m, \qquad \hat{\mu}_{m,\mathbb{P}_j^-} = \frac{1}{n_j^-}\sum_{v=1}^{n_j^-}\phi_m(\boldsymbol{\xi}_v^{-(j)}) \in \mathcal{H}_m. \tag{3.7}$$

By (2.2), the inner products entering (HIEDC-$\mathcal{P}$) then reduce to empirical double sums over the two families. Assembling them gives the three block Gram matrices of kernel m,

$$\begin{aligned}
\hat{K}^{++}_{m,ii'} &= \langle \hat{\mu}_{m,\mathbb{P}_i^+}, \hat{\mu}_{m,\mathbb{P}_{i'}^+}\rangle_{\mathcal{H}_m} = \frac{1}{n_i^+ n_{i'}^+}\sum_{u=1}^{n_i^+}\sum_{v=1}^{n_{i'}^+} k_m(\boldsymbol{\xi}_u^{+(i)}, \boldsymbol{\xi}_v^{+(i')}),\\
\hat{K}^{--}_{m,jj'} &= \langle \hat{\mu}_{m,\mathbb{P}_j^-}, \hat{\mu}_{m,\mathbb{P}_{j'}^-}\rangle_{\mathcal{H}_m} = \frac{1}{n_j^- n_{j'}^-}\sum_{u=1}^{n_j^-}\sum_{v=1}^{n_{j'}^-} k_m(\boldsymbol{\xi}_u^{-(j)}, \boldsymbol{\xi}_v^{-(j')}),\\
\hat{K}^{+-}_{m,ij} &= \langle \hat{\mu}_{m,\mathbb{P}_i^+}, \hat{\mu}_{m,\mathbb{P}_j^-}\rangle_{\mathcal{H}_m} = \frac{1}{n_i^+ n_j^-}\sum_{u=1}^{n_i^+}\sum_{v=1}^{n_j^-} k_m(\boldsymbol{\xi}_u^{+(i)}, \boldsymbol{\xi}_v^{-(j)}),
\end{aligned} \tag{3.8}$$

together with the diagonal norms $\|\hat{\mu}_{m,\mathbb{P}_i^+}\|^2 = \hat{K}^{++}_{m,ii}$ and $\|\hat{\mu}_{m,\mathbb{P}_j^-}\|^2 = \hat{K}^{--}_{m,jj}$. After this substitution, every quantity in (HIEDC-$\mathcal{P}$) depends on the data only through these per-kernel, block Gram matrices. For a fixed weight they combine into the composite empirical embeddings

$$\hat{\mu}_{\boldsymbol{\beta},\mathbb{P}_i^+} = \big(\sqrt{\beta_m}\,\hat{\mu}_{m,\mathbb{P}_i^+}\big)_{m=1}^M, \qquad \hat{\mu}_{\boldsymbol{\beta},\mathbb{P}_j^-} = \big(\sqrt{\beta_m}\,\hat{\mu}_{m,\mathbb{P}_j^-}\big)_{m=1}^M,$$

so that the geometric formulation of Section 2.2 applies verbatim to the empirical objects. In other words, the problems above are to be read with each Gram matrix $\mathbf{K}_m$ (and hence $\mathbf{K}_{\boldsymbol{\beta}}$, its signed version, and all embeddings) replaced by the corresponding estimators $\hat{\mathbf{K}}_m, \hat{\mathbf{K}}_{\boldsymbol{\beta}}, \hat{\mu}$. Correspondingly, the ambiguity set $\mathcal{U}^{\mathrm{HIEDC}}_{\nu,\eta,b,\lambda}$ of Section 2.2 (and the single-kernel $\mathcal{U}^{\mathrm{IEDC}}_{\nu,\eta,b}$ of Section 2.1) is understood in its hatted, empirical form $\widehat{\mathcal{U}}^{\mathrm{HIEDC}}_{\nu,\eta,b,\lambda}$ (resp. $\widehat{\mathcal{U}}^{\mathrm{IEDC}}_{\nu,\eta,b}$) built on these empirical quantities. We work with these empirical quantities from now on and defer a systematic analysis of the induced finite-sample approximation error to Section 4.

3.2 Solving the robustified optimization problem

We now turn the ambiguity set learned in Section 3.1 into robust decisions by solving the robustified program (RooDSO-HIEDC). Since its inner supremum ranges over distributions whose composite embedding falls in a closed ball of the composite RKHS, the worst-case expectation admits a clean dual description. Throughout this subsection we work with the data-driven ball and ambiguity set: let $\hat{\boldsymbol{\beta}}$, $\hat{\mu}_c$ and $\hat{R}$ denote the kernel weights, center and radius returned by Algorithm 1, and write the empirical ball and its induced ambiguity set as

$$\widehat{\mathcal{B}}^{\mathrm{HIEDC}}_{\nu,\eta,b,\lambda} := \Big\{\mu \in \mathcal{H} :\ \|\mu - \hat{\mu}_c\|_{\mathcal{H}} \le \hat{R}\Big\}, \qquad \widehat{\mathcal{U}}^{\mathrm{HIEDC}}_{\nu,\eta,b,\lambda} := \Big\{\mathbb{P} \in \mathcal{P}(\Xi) :\ \mu_{\hat{\boldsymbol{\beta}},\mathbb{P}} \in \widehat{\mathcal{B}}^{\mathrm{HIEDC}}_{\nu,\eta,b,\lambda}\Big\},$$

where $\mathcal{H}$ is the RKHS of the composite kernel $k_{\hat{\boldsymbol{\beta}}}(\boldsymbol{\xi},\boldsymbol{\zeta}) = \sum_{m=1}^{M} \hat{\beta}_m k_m(\boldsymbol{\xi},\boldsymbol{\zeta})$ with canonical feature map $\varphi_{\hat{\boldsymbol{\beta}}}(\boldsymbol{\xi}) = \big(\sqrt{\hat{\beta}_m}\,\varphi_m(\boldsymbol{\xi})\big)_{m=1}^{M}$, and $\mu_{\hat{\boldsymbol{\beta}},\mathbb{P}} := \mathbb{E}_{\boldsymbol{\xi}\sim\mathbb{P}}\{\varphi_{\hat{\boldsymbol{\beta}}}(\boldsymbol{\xi})\}$ is the composite embedding (2.8) of $\mathbb{P}$. The data-driven robustified program is the instance of (RooDSO-HIEDC) built on this set,

$$\min_{\mathbf{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\widehat{\mathcal{U}}^{\mathrm{HIEDC}}_{\nu,\eta,b,\lambda}} \mathbb{E}_{\boldsymbol{\xi}\sim\mathbb{P}}\big\{f(\mathbf{x},\boldsymbol{\xi})\big\}. \tag{3.9}$$

We open the inner supremum by the generalized duality for RKHS-constrained ambiguity sets developed in [27] (and used, in the single-kernel case, by RooDSO [13]). For completeness we state the tool in a form tailored to our setting.

Lemma 2 (Generalized duality for ball-constrained distributions [27])**.** *Let $\mathcal{H}$ be an RKHS with canonical feature map $\varphi : \Xi \to \mathcal{H}$, and let $\mathcal{M} \subset \mathcal{H}$ be a closed convex set. For the ambiguity set $\mathcal{U} := \{\mathbb{P} \in \mathcal{P}(\Xi) :\ \mathbb{E}_{\boldsymbol{\xi}\sim\mathbb{P}}\{\varphi(\boldsymbol{\xi})\} \in \mathcal{M}\}$, suppose $f(\mathbf{x},\cdot)$ is proper and upper semicontinuous and $\mathcal{U}$ has nonempty relative interior. Then*

$$\min_{\mathbf{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{U}} \mathbb{E}_{\boldsymbol{\xi}\sim\mathbb{P}}\big\{f(\mathbf{x},\boldsymbol{\xi})\big\} = \min_{\substack{\mathbf{x}\in\mathcal{X}\\ g\in\mathcal{H},\,\rho\in\mathbb{R}}} h^*_{\mathcal{M}}(g) + \rho \quad s.t. \quad f(\mathbf{x},\boldsymbol{\xi}) \le g(\boldsymbol{\xi}) + \rho, \quad \forall\,\boldsymbol{\xi}\in\Xi, \tag{3.10}$$

where $g(\boldsymbol{\xi}) = \langle g, \varphi(\boldsymbol{\xi})\rangle_{\mathcal{H}}$ and $h^*_{\mathcal{M}}(g) := \sup_{\mu\in\mathcal{M}}\langle g,\mu\rangle_{\mathcal{H}}$ is the support function of $\mathcal{M}$. Moreover, strong duality holds for the inner supremum at every fixed $\mathbf{x}\in\mathcal{X}$.

The following proposition specializes Lemma 2 to the ball $\widehat{\mathcal{B}}^{\mathrm{HIEDC}}_{\nu,\eta,b,\lambda}$, whose support function is the sum of a center term and a radius term.

Proposition 3 (Dual reformulation of the robustified HIEDC program)**.** *Let $\hat{\boldsymbol{\beta}}$, $\hat{\mu}_c$, $\hat{R}$ be the output of Algorithm 1, and let $\mathcal{S}^+ \subseteq [N^+]$ and $\mathcal{S}^- \subseteq [N^-]$ index the inclusive and exclusive* support *measures, i.e., those whose optimal multipliers are strictly positive. If $f(\mathbf{x},\cdot)$ is proper and upper semicontinuous, then problem (3.9) is equivalent to the convex program*

$$\min_{\substack{\mathbf{x}\in\mathcal{X}\\ g\in\mathcal{H},\,\rho\in\mathbb{R}}} \quad \langle g,\hat{\mu}_c\rangle_{\mathcal{H}} + \hat{R}\,\|g\|_{\mathcal{H}} + \rho \tag{3.11a}$$

$$s.t. \quad f(\mathbf{x},\boldsymbol{\xi}) \le g(\boldsymbol{\xi}) + \rho, \quad \forall\,\boldsymbol{\xi}\in\Xi, \tag{3.11b}$$

where $g(\boldsymbol{\xi}) := \langle g, \varphi_{\hat{\boldsymbol{\beta}}}(\boldsymbol{\xi})\rangle_{\mathcal{H}}$. Expanding the center by its signed representer (3.5), the linear term reads

$$\langle g,\hat{\mu}_c\rangle_{\mathcal{H}} = \frac{1}{N\nu}\Big(\sum_{i\in\mathcal{S}^+} \alpha_i^{\star+}\,\mathbb{E}_{\boldsymbol{\xi}\sim\hat{\mathbb{P}}_i^+}\{g(\boldsymbol{\xi})\} - \sum_{j\in\mathcal{S}^-} \alpha_j^{\star-}\,\mathbb{E}_{\boldsymbol{\xi}\sim\hat{\mathbb{P}}_j^-}\{g(\boldsymbol{\xi})\}\Big), \tag{3.12}$$

i.e., in data-driven form, $\langle g,\hat{\mu}_c\rangle = \frac{1}{N\nu}\big(\sum_{i\in\mathcal{S}^+}\frac{\alpha_i^{\star+}}{n_i^+}\sum_{u=1}^{n_i^+} g(\boldsymbol{\xi}_u^{+(i)}) - \sum_{j\in\mathcal{S}^-}\frac{\alpha_j^{\star-}}{n_j^-}\sum_{v=1}^{n_j^-} g(\boldsymbol{\xi}_v^{-(j)})\big)$.

Two remarks on the reformulation are in order. First, only *support* measures survive in (3.12): measures whose optimal multiplier vanishes make no contribution to the center and hence to the dual objective, mirroring the sparsity structure of Lemma 1. Second, the exclusive measures enter

through the negative sign of (3.12), so that the ball center is pushed away from them and the surrogate g is penalized precisely on the distributions the ambiguity set is designed to exclude.

Problem (3.11) is still infinite-dimensional in the function g and carries a continuum of constraints over Ξ. Fortunately, our model retains a structure closely analogous to that of the single-kernel model studied in [13], so that the scalable and efficient solution strategy developed there can be fully exploited: the constraint set is replaced by a finite support Υ (the pooled samples plus auxiliary discretization points), a robust representer theorem confines g to the span of the finite feature points, yielding a finite-dimensional second-order cone program, and a reduced-complexity representer together with a row-generation scheme make the resulting master problem scalable. Because the composite kernel and the empirical center enter these steps only through block Gram matrices, every heavy operation remains a block Gram multiplication. The introduction of the exclusive distribution clusters and of multiple kernel learning enlarges this computational footprint: the former adds an exclusive block family together with the inclusive–exclusive cross blocks to each Gram matrix, while the latter multiplies the number of such per-kernel Gram blocks by M, so that both the dimension of the kernel matrices and the number of kernel evaluations increase. This growth is, however, largely absorbed by the very structure just described: every block Gram entry is an independent double average over a single pair of clusters, and the M kernel blocks are mutually independent, so the heavy computations parallelize across clusters and kernels and map naturally onto batched GPU kernels. These are precisely the difficulties that modern parallel and GPU computing alleviates effectively.

4 Analysis of the proposed model

This section quantifies the two statistical errors in RooDSO-HIEDC. The first comes from replacing every inclusive and exclusive probability measure by a finite sample from that measure. The second comes from observing only finitely many probability measures from the populations that generate the two families. We first study the stability of the learned ball under the first source of randomness and then establish two-sided out-of-distribution guarantees under both sources. Throughout, $N = N^+ + N^-$, whereas M continues to denote the number of candidate kernels.

4.1 Finite-sample stability of the learned ambiguity set

For each $m \in [M]$, let $\mathbf{K}_m$ and $\hat{\mathbf{K}}_m$ be the population and empirical Gram matrices defined in Sections 2.2 and 3.1, respectively, and let $\mathbf{d}_m$ and $\hat{\mathbf{d}}_m$ be their corresponding signed diagonal vectors. Define the aggregate within-measure sample size and the simultaneous Gram error by

$$\bar{n} := \left(\sum_{i=1}^{N^+} \frac{1}{n_i^+} + \sum_{j=1}^{N^-} \frac{1}{n_j^-} \right)^{-1}, \qquad \varepsilon_n(\delta) := \max_{m \in [M]} \left\{ \|\hat{\mathbf{K}}_m - \mathbf{K}_m\|_2, \|\hat{\mathbf{d}}_m - \mathbf{d}_m\|_2 \right\}. \tag{4.1}$$

For the center itself, which is an element of a Hilbert space rather than a function of Gram entries alone, we also use

$$\zeta_n(\delta) := \left[\sum_{m=1}^{M} \left\{ \sum_{i=1}^{N^+} \|\hat{\mu}_{m,\mathbb{P}_i^+} - \mu_{m,\mathbb{P}_i^+}\|_{\mathcal{H}_m}^2 + \sum_{j=1}^{N^-} \|\hat{\mu}_{m,\mathbb{P}_j^-} - \mu_{m,\mathbb{P}_j^-}\|_{\mathcal{H}_m}^2 \right\} \right]^{1/2}. \tag{4.2}$$

Here and throughout, the subscript n flags a *finite-sample* quantity: both error terms arise from replacing each population object, that is, an embedding $\mu_{m,\mathbb{P}}$ or a Gram entry built from it, by its empirical counterpart obtained from the n_i (resp. n_j) samples within a measure, and their magnitude is governed by the aggregate sample size $\bar{n}$ of (4.1): ε_n aggregates the Gram-level (spectral-norm) discrepancies over all kernels, while ζ_n collects the Hilbert-space embedding errors that enter the recovered center. The argument $\delta \in (0,1)$ is a *confidence parameter*: a bound written $\varepsilon_n(\delta)$, $\zeta_n(\delta)$ is a statement holding with probability at least $1-\delta$ over the samples, and δ enters only through logarithmic factors. The same convention underlies the single-kernel analysis of [13, Propositions 3 and 4], whose moment bound for an empirical mean embedding and spectral-norm bound for the empirical Gram matrix are stated for arbitrary confidence levels. We invoke those results kernel-wise below. We impose the following regularity conditions. They are stated for the multiple-kernel model, and their single-kernel versions are immediate.

Assumption 2 (Kernel dictionary and stability)**.** *The kernels $k_1, \dots, k_M$ are continuous, symmetric, and integrally strictly positive definite, and $\sup_{m\in[M],\,\boldsymbol{\xi}\in\Xi} k_m(\boldsymbol{\xi},\boldsymbol{\xi}) \le C < \infty$. The population Gram matrices satisfy $\underline{\kappa} := \min_{m\in[M]} \lambda_{\min}(\mathbf{K}_m) > 0$, and $\lambda > 0$. Moreover, the population optimizer $\boldsymbol{\beta}^\star$ is interior to the simplex, i.e., $\min_m \beta_m^\star \ge \underline{\beta} > 0$, and the optimal radius is locally identified by a strictly complementary inclusive boundary support measure.*

The moment bound for an empirical mean embedding and the spectral-norm bound for its Gram matrix are already available from [13, Propositions 3 and 4]. Applying those results to each of the M kernels and to both families of probability measures, and taking a union bound over the kernels, taken at confidence level δ/M for each, gives, for a constant $C_K > 0$ depending only on the uniform kernel bound (the derivation is collected in Appendix A.3),

$$\varepsilon_n(\delta) \le C_K \left[\sqrt{\frac{N\{1+\log(NM/\delta)\}}{\bar{n}}} + \frac{1+\log(NM/\delta)}{\bar{n}} \right] \tag{4.3}$$

with probability at least $1-\delta$. We therefore do not repeat the two preliminary concentration results here and concentrate on how this error propagates through multiple-kernel learning. The same mean-embedding result gives

$$\zeta_n(\delta) \le C_\mu \sqrt{\frac{M\{1+\log(NM/\delta)\}}{\bar{n}}} \tag{4.4}$$

on the same event after adjusting constants.

Let $\mathcal{A}$ denote the feasible region of (3.4), set $A := \sup_{\boldsymbol{\alpha}\in\mathcal{A}} \|\boldsymbol{\alpha}\|_2 \le \sqrt{N^+ + b^2 N^-}$, and write the population saddle function as

$$\mathscr{L}(\boldsymbol{\beta},\boldsymbol{\alpha}) := \lambda\|\boldsymbol{\beta}\|_2^2 - \frac{1}{2N\nu}\boldsymbol{\alpha}^\top \mathbf{K}_{\boldsymbol{\beta}}^{\pm}\boldsymbol{\alpha} + \frac{1}{2}\mathbf{d}_{\boldsymbol{\beta}}^\top\boldsymbol{\alpha}. \tag{4.5}$$

Then (3.1) is equivalently $\min_{\boldsymbol{\beta}\in\Delta_M}\max_{\boldsymbol{\alpha}\in\mathcal{A}}\mathscr{L}(\boldsymbol{\beta},\boldsymbol{\alpha})$. Its empirical counterpart, denoted by $\hat{\mathscr{L}}$, replaces every $(\mathbf{K}_m, \mathbf{d}_m)$ by $(\hat{\mathbf{K}}_m, \hat{\mathbf{d}}_m)$.

Proposition 4 (Error bound for the empirical HIEDC optimizer)**.** *Suppose Assumption 2 holds. Let $(\boldsymbol{\beta}^\star, \boldsymbol{\alpha}^\star)$ and $(\hat{\boldsymbol{\beta}}^\star, \hat{\boldsymbol{\alpha}}^\star)$ be the unique saddle points of $\mathscr{L}$ and $\hat{\mathscr{L}}$, respectively. Assume also that $\varepsilon_n(\delta) < \underline{\kappa}$, so that the empirical inner objective remains strictly concave. Define*

$$\mu_{\mathrm{sp}} := \min\left\{2\lambda, \frac{\underline{\kappa}}{N\nu}\right\}, \qquad C_{\mathrm{sp}} := \left[M\left(\frac{A^2}{2N\nu} + \frac{A}{2}\right)^2 + \left(\frac{A}{N\nu} + \frac{1}{2}\right)^2\right]^{1/2}.$$

Then, on the event in (4.3),

$$\left\|\hat{\boldsymbol{\beta}}^\star - \boldsymbol{\beta}^\star\right\|_2 + \|\hat{\boldsymbol{\alpha}}^\star - \boldsymbol{\alpha}^\star\|_2 \le \frac{\sqrt{2}\,C_{\mathrm{sp}}}{\mu_{\mathrm{sp}}}\varepsilon_n(\delta). \tag{4.6}$$

Consequently, for fixed N^+, N^-, and M, both estimators converge at the root-$\bar{n}$ rate up to logarithmic factors.

The proposition is the multiple-kernel, two-sided counterpart of [13, Proposition 5]. The additional component is the stability of the learned kernel weights. It follows from the strong convexity supplied by $\lambda\|\boldsymbol{\beta}\|_2^2$ and the strong concavity supplied by the signed Gram quadratic. Thus, $\lambda > 0$ is not merely a device for avoiding a one-kernel solution: it also identifies the population kernel mixture and makes its finite-sample perturbation controllable. This observation is in line with the role of regularized kernel-weight learning in [7, 4].

To compare centers learned under different kernel weights, we regard all composite embeddings as elements of the fixed ambient direct sum $\mathcal{H} := \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_M$, with the m-th coordinate scaled by $\sqrt{\beta_m}$. Let $(\mu_c^\star, R^\star)$ and $(\hat{\mu}_c^\star, \hat{R}^\star)$ denote the population and empirical HIEDC balls obtained from these embeddings.

Theorem 5 (Error bound for the HIEDC center and radius)**.** *Suppose Assumption 2 holds. There exists a constant $C_{\mathcal{B}} > 0$, independent of the within-measure sample sizes, such that, with probability at least $1-\delta$,*

$$\|\hat{\mu}_c^\star - \mu_c^\star\|_{\mathcal{H}} + \left|(\hat{R}^\star)^2 - (R^\star)^2\right| \le C_{\mathcal{B}}\{\varepsilon_n(\delta) + \zeta_n(\delta)\}. \tag{4.7}$$

In particular, for fixed N^+, N^-, and M, the center and squared radius converge at rate $\mathcal{O}(\bar{n}^{-1/2})$, up to logarithmic factors.

Theorem 5 extends [13, Theorem 6] in two ways: the center is a signed combination of inclusive and exclusive distribution embeddings, and the metric itself is estimated through $\boldsymbol{\beta}$. The interior-

weight condition makes $\boldsymbol{\beta} \mapsto \sqrt{\boldsymbol{\beta}}$ locally Lipschitz. If it is removed, $|\sqrt{x} - \sqrt{y}| \le \sqrt{|x-y|}$ still yields consistency, but the norm error of the center can in general deteriorate to $\mathcal{O}(\varepsilon_n^{1/2})$. Quantities depending linearly on the composite Gram remain root-$\bar{n}$ stable. When $M = 1$, there is no kernel-weight error, and the results above reduce to the corresponding finite-sample statements for RooDSO-IEDC.

4.2 Two-sided out-of-distribution generalization guarantees

We now account for the upper layer of randomness. RooDSO [13] models all source distributions as independent draws from a *single* meta-distribution. Because exclusive evidence speaks about where the target does not come from, it cannot be regarded as further draws from the same meta-distribution that generates the inclusive evidence. The two-sided setting therefore calls for two meta-distributions, which we posit as follows.

Assumption 3 (Two-layer randomness and the two meta-distributions)**.** [1] *Let Ψ^+ be the meta-distribution generating probability measures judged compatible with the target, and let Ψ^- generate probability measures judged incompatible with it. Then*

$$\mathbb{P}_1^+, \dots, \mathbb{P}_{N^+}^+ \overset{\text{i.i.d.}}{\sim} \Psi^+, \qquad \mathbb{P}_1^-, \dots, \mathbb{P}_{N^-}^- \overset{\text{i.i.d.}}{\sim} \Psi^-, \tag{4.8}$$

with the two families independent, while the unobserved target distribution is an independent draw $\mathbb{P} \sim \Psi^+$.

For a learned population ball, define its signed squared-distance score

$$h(\mathbb{Q}) := \|\mu_{\boldsymbol{\beta}^\star, \mathbb{Q}} - \mu_c^\star\|_{\mathcal{H}_{\boldsymbol{\beta}^\star}}^2 - (R^\star)^2. \tag{4.9}$$

Thus $h(\mathbb{Q}) \le 0$ means that $\mathbb{Q}$ is admitted by the HIEDC ambiguity set. Let $\mathcal{G}$ be the class of scores of the form (4.9) generated by feasible kernel weights and bounded ball centers and radii. Denote its Rademacher complexity under q draws from $\Psi^\pm$ by $\mathfrak{R}_q^\pm(\mathcal{G})$. Uniformly bounded kernels and centers ensure $\mathfrak{R}_q^\pm(\mathcal{G}) \le C_{\mathcal{G}} \sqrt{\log(2M)/q}$ for some constant $C_{\mathcal{G}} > 0$, where the logarithmic dependence is the usual price of learning from a finite kernel dictionary [4].

The fraction certificates in Lemma 1 use N as their denominator. For generalization within each family, it is therefore convenient to define

$$\tau^+ := \min\left\{\frac{N(\nu+\eta)}{N^+},\ 1\right\}, \qquad \tau^- := \min\left\{\frac{N\eta}{bN^-},\ 1\right\}. \tag{4.10}$$

[1] The assumption is stated with a single meta-distribution Ψ^- for notational simplicity, but it is compatible with heterogeneous exclusive evidence drawn from several distinct sources. One may read Ψ^- as the *mixture* of the meta-distributions that generate the various exclusive sources—or, equivalently, assign each exclusive source its own meta-distribution Ψ_k^- and apply the bound source-wise—in which case the exclusive guarantee of Theorem 6 holds for a new incompatible distribution drawn from the same mixture. A genuinely new, untrained exclusive source, by contrast, falls outside its scope, since no observation of it enters the training set.

Theorem 6 (Two-sided out-of-distribution generalization)**.** *Suppose Assumption 3 holds and the population HIEDC solution has $t^\star > 0$. For any $\delta \in (0,1)$ and $\gamma > 0$, with probability at least $1-\delta$ over the probability measures in (4.8), the following inequalities hold simultaneously:*

$$\Pr_{\mathbb{P}' \sim \Psi^+} \{h(\mathbb{P}') > \gamma\} \le \tau^+ + \frac{2}{\gamma}\mathfrak{R}^+_{N^+}(\mathcal{G}) + \sqrt{\frac{\log(4/\delta)}{2N^+}}, \tag{4.11a}$$

$$\Pr_{\mathbb{P}' \sim \Psi^-} \{h(\mathbb{P}') \le 0\} \le \tau^- + \frac{2}{t^\star}\mathfrak{R}^-_{N^-}(\mathcal{G}) + \sqrt{\frac{\log(4/\delta)}{2N^-}}. \tag{4.11b}$$

The first inequality says that a new compatible distribution lies in the slightly enlarged ambiguity set $\mathcal{U}_\gamma^{\mathrm{HIEDC}} := \{\mathbb{Q} : h(\mathbb{Q}) \le \gamma\}$ with high probability. The second says that a new incompatible distribution is unlikely to enter the original ambiguity set. The margin $t^\star$ is therefore not only geometric: it is the statistical buffer that supports out-of-distribution rejection. If $t^\star = 0$, the exclusive rejection bound becomes vacuous and the inclusive term τ^+ must be replaced using the data-dependent positive-error certificate in (2.14). This is precisely the separability warning discussed in Remark 3.

We next return to the empirically learned ball. Define its score $\hat{h}$ by replacing $(\boldsymbol{\beta}^\star, \mu_c^\star, R^\star)$ in (4.9) with $(\hat{\boldsymbol{\beta}}^\star, \hat{\mu}_c^\star, \hat{R}^\star)$, and let $\Delta_n(\delta)$ be any bound satisfying

$$\sup_{\mathbb{Q} \in \mathcal{P}(\Xi)} |\hat{h}(\mathbb{Q}) - h(\mathbb{Q})| \le \Delta_n(\delta) \tag{4.12}$$

with probability at least $1-\delta$ over the within-measure samples. Under Assumption 2, Theorem 5 and boundedness of the kernels give $\Delta_n(\delta) \le C_h\{\varepsilon_n(\delta) + \zeta_n(\delta)\}$ for a constant $C_h > 0$.

Corollary 7 (Two-layer guarantee for the empirical HIEDC set)**.** *Under the conditions of Theorem 6 and Assumption 2, the inclusive bound (4.11a) remains valid with the event replaced by*

$$\left\{\mathbb{P}' \notin \widehat{\mathcal{U}}^{\mathrm{HIEDC}}_{\gamma+\Delta_n}\right\}, \qquad \widehat{\mathcal{U}}^{\mathrm{HIEDC}}_{\gamma+\Delta_n} := \{\mathbb{Q} : \hat{h}(\mathbb{Q}) \le \gamma + \Delta_n\}.$$

If $\Delta_n(\delta) < t^\star$, the exclusive bound (4.11b) remains valid after replacing its complexity term by $2\mathfrak{R}^-_{N^-}(\mathcal{G})/(t^\star - \Delta_n)$ and its event by $\{\mathbb{P}' \in \widehat{\mathcal{U}}^{\mathrm{HIEDC}}\}$. The joint confidence level is at least $1-2\delta$ over both layers of randomness.

Unlike a direct reuse of the single-kernel result, the corollary displays the within-measure estimation error explicitly. This is necessary because the training inputs are empirical embeddings whereas a new target is a probability measure with its population embedding. The correction vanishes at the rate in (4.3), and does not alter the $(N^+)^{-1/2}$ and $(N^-)^{-1/2}$ upper-layer rates.

Finally, coverage of a new draw from Ψ^+ translates directly into a performance certificate for the robust decision.

Proposition 8 (Out-of-distribution guarantee for RooDSO-HIEDC)**.** *Let*

$$\widehat{V}_{\gamma+\Delta_n}(\mathbf{x}) := \sup_{\mathbb{Q}\in\widehat{\mathcal{U}}^{\mathrm{HIEDC}}_{\gamma+\Delta_n}} \mathbb{E}_{\mathbb{Q}}\{f(\mathbf{x},\boldsymbol{\xi})\}.$$

On the event of Corollary 7,

$$\Pr_{\mathbb{P}'\sim\Psi^+}\left\{\forall\,\mathbf{x}\in\mathcal{X}:\ \mathbb{E}_{\mathbb{P}'}\{f(\mathbf{x},\boldsymbol{\xi})\}\le\widehat{V}_{\gamma+\Delta_n}(\mathbf{x})\right\}\ge 1-\tau^+-\frac{2}{\gamma}\mathfrak{R}^+_{N^+}(\mathcal{G})-\sqrt{\frac{\log(4/\delta)}{2N^+}}. \tag{4.13}$$

In particular, the worst-case value at an optimizer of the enlarged empirical RooDSO-HIEDC problem is an upper confidence bound on its expected cost under a new target distribution.

Sending $N^+\to\infty$, $N^-\to\infty$, and $\bar{n}\to\infty$ while M is fixed drives the complexity terms $\mathfrak{R}^{\pm}_{N^{\pm}}(\mathcal{G})$ and the estimation correction Δ_n to zero. Letting these limits run through the two bounds of Theorem 6, the inclusive bound (4.11a) yields asymptotic coverage of the enlarged set $\mathcal{U}^{\mathrm{HIEDC}}_{\gamma}$ of level at least $1-\tau^+$, i.e., $\Pr_{\mathbb{P}'\sim\Psi^+}\{h(\mathbb{P}')>\gamma\}\le\tau^+$ in the limit, while the exclusive bound (4.11b) yields an asymptotic false-acceptance probability at most τ^-, i.e., $\Pr_{\mathbb{P}'\sim\Psi^-}\{h(\mathbb{P}')\le 0\}\le\tau^-$ in the limit. These are the two-sided analogues of the $1-\nu$ property in RooDSO [13]. For $M=1$, all statements specialize to RooDSO-IEDC without further argument.

5 Concluding remarks

In this paper we developed robust out-of-distribution stochastic optimization under a target distribution that is completely unobserved, so that neither the distribution itself nor any sample from it is available at decision time, and showed how human judgment, rather than being set aside in an age of abundant automated data, can be systematically mined to inform decisions in precisely such cold-start settings. Our point of departure was the observation that knowledge about the unknown has *two* sides: not only can experts state which distributions the target resembles (inclusive clusters), they are often even better at stating which distributions it does *not* resemble (exclusive clusters), and the latter may be easier to elicit. We turned this into a complete mathematical model, RooDSO-IEDC, in which a convex small-sphere–large-margin problem on kernel mean embeddings encircles the inclusive evidence while keeping the exclusive evidence out, thereby producing an ambiguity set that is simultaneously data-driven, decision-oriented, and *well posed*: by replacing the classical SVDD-type ball of RooDSO [13] with the convex SSLM formulation [16], the optimal description (center and radius) is provably unique under finite samples, so the conservatism of the downstream robust decision is no longer ambiguous. Because the relevant source distributions are typically heterogeneous and resist any single kernel, we then extended the model to RooDSO-HIEDC, composing several reproducing kernel Hilbert spaces and learning their combination jointly with the ambiguity set through multiple kernel learning, with RooDSO-IEDC recovered as its single-kernel special case. We also showed that the hyperparameters of the model transparently encode the decision maker's emphasis on positive versus negative information, the

degree of data heterogeneity, and the desired risk posture, providing a principled calibration handle. On the computational side, we derived a dual formulation in which every heavy operation reduces to block Gram-matrix computations, and presented scalable learning and optimization strategies compatible to modern parallel and GPU hardware. Finally, we provided a statistical analysis of the proposed model, establishing finite-sample stability of the learned ball and two-sided out-of-distribution generalization guarantees. Together these ingredients offer, we believe, a coherent answer to the challenge posed in the introduction: robust decisions under an unobserved distribution, grounded in both what the unknown is and what it is not.

Several directions deserve future investigation. First, the present model treats all clusters as unconditional probability measures. In practice, relevant distributions are often accompanied by contextual information, and augmenting each cluster with covariates would enable conditional estimation and out-of-distribution decision-making under side information. Our framework can be extended along two dimensions, either by learning a covariate-to-distribution mapping from sources endowed with covariates, or by letting each cluster be itself conditional on context. Second, the meta-distribution underlying the two families is summarized here only through the first-order kernel mean embeddings that the ambiguity set retains. Exploiting higher-order or richer statistical profiles of the meta-distribution (e.g., covariance information or dependence structure across clusters) could sharpen the learned set when the number of available clusters is large. Third, the elicitation process itself is treated as exogenous. Formalizing how inclusive and exclusive clusters should be queried, how many are needed, and how mislabelled clusters can be detected from the learned certificates would connect the model to active learning and to the statistical diagnostics developed in Section 2.3. Finally, an empirical evaluation on synthetic and real-world cold-start tasks, comparing the proposed methodology against RooDSO and other distributionally robust baselines, is an important next step toward validating the framework in the application domains that motivate it.

References

[1] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[2] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[3] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 2nd edition, 1999.

[4] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Generalization bounds for learning kernels. In *Proceedings of the 27th International Conference on Machine Learning*, pages 247–254, 2010.

[5] John M. Danskin. *The Theory of Max-Min and its Application to Weapons Allocation Problems*. Springer, 1967.

[6] Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153, 1989.

[7] Biao Han, Chao Shang, and Dexian Huang. Multiple kernel learning-aided robust optimization: Learning algorithm, computational tractability, and usage in multi-stage decision-making. *European Journal of Operational Research*, 292(3):1004–1018, 2021.

[8] Lars Peter Hansen and Thomas J. Sargent. Robust control and model uncertainty. *American Economic Review*, 91(2):60–66, 2001.

[9] Janellen Huttenlocher. Some effects of negative instances on the formation of simple concepts. *Psychological Reports*, 11(1):35–42, 1962.

[10] George Kimeldorf and Grace Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

[11] Peter Klibanoff, Massimo Marinacci, and Sujoy Mukerji. A smooth model of decision making under ambiguity. *Econometrica*, 73(6):1849–1892, 2005.

[12] Daniel Kuhn, Soroosh Shafiee, and Wolfram Wiesemann. Distributionally robust optimization. *Acta Numerica*, 34:579–804, 2025.

[13] Xianyu Li, Huan Xu, Xiaolin Huang, and Chao Shang. Robust out-of-distribution stochastic optimization. *arXiv preprint arXiv:2604.20147*, April 2026.

[14] Baoding Liu. *Uncertainty Theory*. Springer, Berlin, 2nd edition, 2007.

[15] Feng Liu, Zhi Chen, Ruodu Wang, and Shuming Wang. Newsvendor under ambiguity and misspecification. *Manufacturing & Service Operations Management*, page msom.2025.197, May 2026.

[16] Hongying Liu, Hao Wang, Haoran Chu, and Yibo Wu. Towards convexity in anomaly detection: A new formulation of SSLM with unique optimal solutions. *Journal of Machine Learning Research*, 27(89):1–36, May 2026.

[17] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

[18] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.

[19] Krikamol Muandet and Bernhard Schölkopf. One-class support measure machines for group anomaly detection. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 449–458, 2013.

[20] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[21] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[22] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2021.

[23] Kenneth L. Smoke. Negative instances in concept learning. *Journal of Experimental Psychology*, 16(4):583–588, 1933.

[24] David M.J. Tax and Robert P.W. Duin. Support vector data description. *Machine Learning*, 54(1):45–66, January 2004.

[25] Xiaoming Wang, Fulai Chung, and Shitong Wang. Theoretical analysis for solution of support vector data description. *Neural Networks*, 24(4):360–369, 2011.

[26] Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.

[27] Jia-Jie Zhu, Wittawat Jitkrittum, Moritz Diehl, and Bernhard Schölkopf. Kernel distributionally robust optimization: Generalized duality theorem and stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 280–288, 2021.

Appendix A Proofs

A.1 Additional notation

Before proceeding, we introduce notations that are not used in the main text but will be required in the sequel. The notation $\|Z\|_\infty$ denotes the essential supremum norm of a random variable Z, defined as $\|Z\|_\infty := \inf\{C \geq 0 \,|\, \Pr(\|Z\| \leq C) = 1\}$, namely the smallest constant C such that $\|Z\| \leq C$ holds almost surely under the underlying distribution. We use the notation $X \perp\!\!\!\perp Y$ to denote that the random variables X and Y are independent. We use $\chi_{\mathcal{S}}$ for the $\{0,1\}$-valued indicator of a set $\mathcal{S}$, and $\iota_{\mathcal{S}}$ for the convex-analysis indicator, which is 0 on $\mathcal{S}$ and $+\infty$ otherwise. We denote by $\mathrm{Proj}_{\mathcal{H}}$ the orthogonal projector onto $\mathcal{H}$. We use $A_n \downarrow A$ to denote a decreasing sequence of sets or events, i.e., $A_{n+1} \subseteq A_n$ for all n, and $\bigcap_{n\geq 1} A_n = A$. For any bounded operator $\Delta : \mathcal{H}_1 \to \mathcal{H}_2$, where $\mathcal{H}_1$ and $\mathcal{H}_2$ are Hilbert spaces, we define its induced operator norm $\|\Delta\|_{\mathrm{op}} := \sup_{\|g\|_{\mathcal{H}_1}=1} \|\Delta(g)\|_{\mathcal{H}_2}$, and use Δ^* to denote its adjoint operator. We write $a \lesssim b$ to mean $a \leq Db$ for some universal constant $D > 0$ that may change from line to line.

A.2 Proof of Proposition 3

Apply Lemma 2 with $\mathcal{H} = \mathcal{H}_{\hat{\boldsymbol{\beta}}}$ (in the unified notation of Section 2.2, the ambient space $\mathcal{H}$ equipped with the composite inner product), $\varphi = \varphi_{\hat{\boldsymbol{\beta}}}$, and $\mathcal{M} = \widehat{\mathcal{B}}^{\mathrm{HIEDC}}_{\nu,\eta,b,\lambda}$. The support function of the ball is, by Cauchy–Schwarz and its attainability at the boundary point $\hat{\mu}_c + \hat{R}\, g/\|g\|$ (whenever $g \neq 0$),

$$h^*_{\widehat{\mathcal{B}}}(g) = \sup_{\|\mu - \hat{\mu}_c\| \leq \hat{R}} \langle g, \mu \rangle = \langle g, \hat{\mu}_c \rangle + \hat{R}\, \|g\|,$$

which inserted in the dual of Lemma 2 gives (3.11). The expansion (3.12) follows by substituting the center $\hat{\mu}_c = \frac{1}{N\nu}(\sum_{i\in\mathcal{S}^+} \alpha_i^{\star+} \hat{\mu}_{\hat{\boldsymbol{\beta}},\mathbb{P}_i^+} - \sum_{j\in\mathcal{S}^-} \alpha_j^{\star-} \hat{\mu}_{\hat{\boldsymbol{\beta}},\mathbb{P}_j^-})$ and using the reproducing property $\langle g, \hat{\mu}_{\hat{\boldsymbol{\beta}},\mathbb{P}} \rangle = \mathbb{E}_{\boldsymbol{\xi}\sim\hat{\mathbb{P}}}\{g(\boldsymbol{\xi})\}$.

A.3 Derivation of the simultaneous concentration bounds

We spell out how (4.3) and (4.4) follow from the single-kernel estimates of [13, Propositions 3 and 4]. Throughout, fix a kernel m and write $\bar{n}$ for the aggregate sample size (4.1). For one fixed kernel, the Gram matrix $\mathbf{K}_m$ is an $N \times N$ matrix over the $N = N^+ + N^-$ probability measures. Reading the number of sources in [13, Proposition 4] as our N and its confidence level as δ_m, that result gives, for any $\delta_m \in (0,1)$,

$$\|\hat{\mathbf{K}}_m - \mathbf{K}_m\|_2 \;\leq\; \sqrt{\frac{8N\,C^2\big(1 + 2\log(N/\delta_m)\big)}{\bar{n}}} + \frac{2C\big(1 + 2\log(N/\delta_m)\big)}{\bar{n}}$$

with probability at least $1 - \delta_m$, where C is the uniform kernel bound of Assumption 2. The diagonal vector $\hat{\mathbf{d}}_m - \mathbf{d}_m$ is no larger in magnitude: its entries are diagonal deviations of the same empirical Gram, each of which concentrates at the rate $n_i^{-1/2}$ up to logarithms, so after taking the

ℓ_2 norm over the N entries and absorbing constants it obeys the same bound on the same event.

To make the bound hold *simultaneously* for all M kernels, we set the per-kernel confidence to $\delta_m = \delta/M$. By the union bound,

$$\Pr\Big\{\forall\, m \in [M]:\ \|\hat{\mathbf{K}}_m - \mathbf{K}_m\|_2 \text{ and } \|\hat{\mathbf{d}}_m - \mathbf{d}_m\|_2 \text{ obey the bound above}\Big\} \geq 1 - \sum_{m=1}^{M} \delta_m = 1 - \delta.$$

Substituting δ/M for δ_m turns $\log(N/\delta_m)$ into $\log(NM/\delta)$. Since $\varepsilon_n(\delta)$ in (4.1) is the maximum over the M kernels of these deviations, every term above is bounded by the same expression with $\log(NM/\delta)$ in place of $\log(N/\delta_m)$. To recover the condensed form (4.3), note that $1 + 2\log x \leq 2(1 + \log x)$ for $x \geq 1$, so the absolute factors $\sqrt{8}$, 2, and the kernel bound C fold into a single constant $C_K > 0$ that depends only on the uniform kernel bound. Hence

$$\varepsilon_n(\delta) \ \leq\ C_K\Big[\sqrt{\frac{N\big(1 + \log(NM/\delta)\big)}{\bar{n}}} + \frac{1 + \log(NM/\delta)}{\bar{n}}\Big]$$

with probability at least $1 - \delta$.

The embedding bound (4.4) is obtained analogously from the moment inequality of [13, Proposition 3]: applied to one embedding $\hat{\mu}_{m,\mathbb{P}_i^+}$ with n_i^+ samples (resp. n_j^- for the exclusive family) and converted to a high-probability statement by Markov's inequality, it gives $\|\hat{\mu}_{m,\mathbb{P}_i^+} - \mu_{m,\mathbb{P}_i^+}\|_{\mathcal{H}_m} \lesssim (n_i^+)^{-1/2}$ up to logarithms. Squaring each such deviation, summing over the $N^+ + N^-$ measures and the M kernels, and using the definition of $\bar{n}$ in (4.1), $\sum_i (n_i^+)^{-1} + \sum_j (n_j^-)^{-1} = \bar{n}^{-1}$, yields $\zeta_n^2(\delta) \lesssim M\,\bar{n}^{-1}$ times logarithmic factors. Taking the square root and applying the same union bound over the M kernels and the two families gives (4.4).

A.4 Proof of Proposition 4

Write $z = (\boldsymbol{\beta}, \boldsymbol{\alpha})$ and introduce the saddle operator

$$T(z) := \big(\nabla_{\boldsymbol{\beta}} \mathscr{L}(z), -\nabla_{\boldsymbol{\alpha}} \mathscr{L}(z)\big),$$

with $\hat{T}$ defined analogously from $\hat{\mathscr{L}}$. Since $\mathbf{K}_m^{\pm} = \mathbf{D}\mathbf{K}_m\mathbf{D}$, congruence by the orthogonal sign matrix preserves eigenvalues. Hence, for every $\boldsymbol{\beta} \in \Delta_M$,

$$\lambda_{\min}(\mathbf{K}_{\boldsymbol{\beta}}^{\pm}) \geq \sum_{m=1}^{M} \beta_m \lambda_{\min}(\mathbf{K}_m) \geq \underline{\kappa}.$$

The function $\mathscr{L}$ is therefore 2λ-strongly convex in $\boldsymbol{\beta}$ and $\underline{\kappa}/(N\nu)$-strongly concave in $\boldsymbol{\alpha}$. Its saddle operator is consequently strongly monotone with modulus μ_{sp}.

Both saddle points solve their respective variational inequalities over the common set $\Delta_M \times \mathcal{A}$.

Combining these two inequalities and using strong monotonicity gives

$$\mu_{\mathrm{sp}}\|\hat{z}^\star - z^\star\|_2^2 \le \langle (T-\hat{T})(\hat{z}^\star), \hat{z}^\star - z^\star\rangle \le \|(T-\hat{T})(\hat{z}^\star)\|_2 \|\hat{z}^\star - z^\star\|_2. \tag{A.1}$$

For any feasible $\boldsymbol{\alpha}$, the perturbation of the m-th weight derivative satisfies

$$\left| -\frac{1}{2N\nu}\boldsymbol{\alpha}^\top(\hat{\mathbf{K}}_m^{\pm} - \mathbf{K}_m^{\pm})\boldsymbol{\alpha} + \frac{1}{2}(\hat{\mathbf{d}}_m - \mathbf{d}_m)^\top \boldsymbol{\alpha} \right| \le \left(\frac{A^2}{2N\nu} + \frac{A}{2}\right)\varepsilon_n(\delta),$$

whereas the perturbation of the multiplier derivative is at most

$$\left(\frac{A}{N\nu} + \frac{1}{2}\right)\varepsilon_n(\delta).$$

It follows that $\|(T-\hat{T})(\hat{z}^\star)\|_2 \le C_{\mathrm{sp}}\varepsilon_n(\delta)$. Substitution into (A.1), followed by $\|a\|_2 + \|b\|_2 \le \sqrt{2}\|(a,b)\|_2$, proves (4.6).

A.5 Proof of Theorem 5

The m-th block of the population center recovered from (3.5) is

$$\mu_{c,m}^\star = \frac{\sqrt{\beta_m^\star}}{N\nu}\left(\sum_{i=1}^{N^+}\alpha_i^{\star+}\mu_{m,\mathbb{P}_i^+} - \sum_{j=1}^{N^-}\alpha_j^{\star-}\mu_{m,\mathbb{P}_j^-}\right), \tag{A.2}$$

and the empirical block has the same expression with hats. Add and subtract the expressions obtained by changing, one at a time, the multiplier, the embedding, and the square root of the kernel weight. The triangle inequality, the box bound $\|\boldsymbol{\alpha}\|_2 \le A$, and $\|\mu_{m,\mathbb{Q}}\|_{\mathcal{H}_m} \le \sqrt{C}$ then yield

$$\|\hat{\mu}_c^\star - \mu_c^\star\|_{\mathcal{H}} \le C_1\|\hat{\boldsymbol{\alpha}}^\star - \boldsymbol{\alpha}^\star\|_2 + C_2\zeta_n(\delta) + C_3\|\sqrt{\hat{\boldsymbol{\beta}}^\star} - \sqrt{\boldsymbol{\beta}^\star}\|_2. \tag{A.3}$$

The interior-weight condition and consistency from Proposition 4 imply that, for sufficiently small $\varepsilon_n(\delta)$, both weight vectors have coordinates at least $\underline{\beta}/2$. The mean value theorem therefore gives

$$\|\sqrt{\hat{\boldsymbol{\beta}}^\star} - \sqrt{\boldsymbol{\beta}^\star}\|_2 \le \frac{1}{\sqrt{2\underline{\beta}}}\|\hat{\boldsymbol{\beta}}^\star - \boldsymbol{\beta}^\star\|_2.$$

Combining this inequality with (4.6) proves the claimed center bound.

By local identification, there is an inclusive index i_0 that remains a strictly complementary boundary support measure under all sufficiently small perturbations. Hence

$$(R^\star)^2 = \|\mu_{\boldsymbol{\beta}^\star,\mathbb{P}_{i_0}^+} - \mu_c^\star\|_{\mathcal{H}}^2, \qquad (\hat{R}^\star)^2 = \|\hat{\mu}_{\hat{\boldsymbol{\beta}}^\star,\mathbb{P}_{i_0}^+} - \hat{\mu}_c^\star\|_{\mathcal{H}}^2.$$

Using $|\|a\|^2 - \|b\|^2| \le (\|a\| + \|b\|)\|a-b\|$, boundedness of all embeddings, and (A.3) proves the radius bound after enlarging the constant. Equations (4.3) and (4.4) give the stated rate.

A.6 Proof of Theorem 6

For inclusive coverage, introduce the ramp loss

$$\ell_\gamma^+(u) := \begin{cases} 0, & u \le 0, \\ u/\gamma, & 0 < u < \gamma, \\ 1, & u \ge \gamma. \end{cases}$$

It is $1/\gamma$-Lipschitz and $\chi_{(\gamma,\infty)}(u) \le \ell_\gamma^+(u)$. Moreover, its empirical average over the inclusive measures is at most ME^+/N^+, since only measures outside the learned ball can make a positive contribution. The standard symmetrization and contraction argument used in [13, Theorem 7] thus gives, with probability at least $1 - \delta/2$,

$$\Pr_{\Psi^+}\{h(\mathbb{P}') > \gamma\} \le \frac{ME^+}{N^+} + \frac{2}{\gamma}\mathfrak{R}_{N^+}^+(\mathcal{G}) + \sqrt{\frac{\log(4/\delta)}{2N^+}}.$$

Lemma 1 and $t^\star > 0$ imply $ME^+/N^+ \le \tau^+$, proving (4.11a).

For exclusion, use the reverse ramp

$$\ell_{t^\star}^-(u) := \begin{cases} 1, & u \le 0, \\ 1 - u/t^\star, & 0 < u < t^\star, \\ 0, & u \ge t^\star. \end{cases}$$

This loss is $1/t^\star$-Lipschitz and majorizes $\chi_{(-\infty,0]}(u)$. Its empirical average over exclusive measures is at most ME^-/N^-, because it vanishes whenever an exclusive embedding lies beyond the outer sphere. The same symmetrization argument and $ME^-/N^- \le \tau^-$ give (4.11b) with probability at least $1 - \delta/2$. A union bound completes the proof.

A.7 Proof of Corollary 7

On the event (4.12),

$$\hat{h}(\mathbb{Q}) > \gamma + \Delta_n \quad \Longrightarrow \quad h(\mathbb{Q}) > \gamma.$$

The inclusive assertion therefore follows directly from (4.11a). On the exclusive side, $\hat{h}(\mathbb{Q}) \le 0$ implies $h(\mathbb{Q}) \le \Delta_n$. Replace the reverse ramp in the preceding proof by the function that equals one on $(-\infty, \Delta_n]$, decreases linearly to zero on $(\Delta_n, t^\star)$, and vanishes on $[t^\star, \infty)$. Its Lipschitz constant is $1/(t^\star - \Delta_n)$, while its empirical average is still bounded by ME^-/N^-. This proves the stated exclusive bound. Taking a union bound over the upper-layer and lower-layer events gives confidence at least $1 - 2\delta$.

A.8 Proof of Proposition 8

If $\mathbb{P}' \in \widehat{\mathcal{U}}^{\mathrm{HIEDC}}_{\gamma+\Delta_n}$, then, for every $\mathbf{x} \in \mathcal{X}$,

$$\mathbb{E}_{\mathbb{P}'}\{f(\mathbf{x}, \boldsymbol{\xi})\} \leq \sup_{\mathbb{Q} \in \widehat{\mathcal{U}}^{\mathrm{HIEDC}}_{\gamma+\Delta_n}} \mathbb{E}_{\mathbb{Q}}\{f(\mathbf{x}, \boldsymbol{\xi})\} = \widehat{V}_{\gamma+\Delta_n}(\mathbf{x}).$$

Thus the coverage event is contained in the simultaneous performance event in (4.13). The result follows from the inclusive part of Corollary 7.