Learning Decision-Sufficient Representations for Linear Optimization

Yuhan Ye YYH03@MIT.EDU
Saurabh Amin AMINS@MIT.EDU
Asuman Özdağlar ASUMAN@MIT.EDU
Massachusetts Institute of Technology

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study how to construct compressed datasets that suffice to recover optimal decisions in linear programs with unknown cost vector c lying in a prior set $\mathcal{C}$. Recent work by Bennouna et al. (2025a) provides an exact geometric characterization of sufficient decision datasets (SDDs) via an intrinsic decision-relevant dimension $d^\star$. However, their algorithm for constructing minimum-size SDDs requires solving mixed-integer programs. In this paper, we establish hardness results: computing $d^\star$ is NP-hard and deciding whether a dataset is globally sufficient is coNP-hard, thereby resolving the open problem posed in Bennouna et al. (2026). To circumvent worst-case intractability, we introduce pointwise sufficiency, a relaxation that requires sufficiency for an individual cost vector. We provide a polynomial-time cutting-plane algorithm to construct pointwise-sufficient decision datasets under nondegeneracy. In a data-driven regime with i.i.d. costs, we propose a cumulative algorithm that aggregates decision-relevant directions across samples, yielding a stable compression scheme of size at most $d^\star$. This leads to a distribution-free PAC guarantee: w.h.p. over the training sample, the pointwise sufficiency failure probability on a fresh draw is at most $\tilde{O}(d^\star/n)$, and this rate is tight up to logarithmic factors. Finally, we apply decision-sufficient representations to contextual linear optimization, obtaining compressed predictors with generalization bounds scaling as $\tilde{O}(\sqrt{d^\star/n})$ rather than $\tilde{O}(\sqrt{d/n})$, where d is the ambient cost dimension.

Keywords: Linear programming, Sample compression, PAC learning, Computational complexity, Decision-focused learning, Contextual optimization, Polyhedral geometry

1. Introduction

In many real-world decision problems, the optimization objective depends on parameters that are not directly observable and must be inferred from data. With the surge in available data, it has become common to use empirical evidence alongside contextual knowledge to support decision-making. The main challenge is to characterize the smallest set of information needed to identify an optimal decision and to recover it efficiently from finite samples via computationally tractable algorithms. Motivated by this, this paper studies the following fundamental question:

Which and how many objective measurements are sufficient to identify an optimal decision, and can we learn such measurements with provable guarantees in polynomial time?

We formulate the decision-making problem as a linear optimization with an unknown cost vector and a known feasible region. The decision-maker solves

$$\min_{x \in \mathcal{X}} \; c^\top x, \tag{1}$$

where $\mathcal{X} := \{x \in \mathbb{R}^d : Ax = b,\ x \geq 0\}$ is a nonempty bounded polytope for some $A \in \mathbb{R}^{m \times d}$ with full row rank and $b \in \mathbb{R}^m$. The objective vector $c \in \mathbb{R}^d$ is unknown, but it is known a

priori to lie in an uncertainty set $\mathcal{C} \subset \mathbb{R}^d$. Rather than observing c directly, the decision-maker deploys a fixed dataset (measurement set) $\mathcal{D} = \{q_1, \ldots, q_l\} \subset \mathbb{R}^d$ chosen in advance and observes the corresponding inner products $s(c; \mathcal{D}) := (q_1^\top c, \ldots, q_l^\top c) \in \mathbb{R}^l$, as linear measurements of the objective. Given observations $s \in \mathbb{R}^l$, we restrict plausible cost vectors to the *fiber* $\mathcal{C}(\mathcal{D}, s) := \{c' \in \mathcal{C} : q_i^\top c' = s_i,\ i = 1, \ldots, l\}$.

Acquiring objective information is often the bottleneck: observing or predicting the full d-dimensional vector c can be unnecessary when optimal decisions only depend on a few *decision-relevant* directions. A central geometric message is that decision-making depends on c only through a low-dimensional set of decision-relevant directions. In particular, for open convex $\mathcal{C}$, Bennouna et al. (2025a) provide an exact geometric characterization: a dataset is globally sufficient on $\mathcal{C}$ if and only if its span contains the decision-relevant subspace $W_\star$ (formally introduced in Equation (3)), whose intrinsic dimension $d^\star$ is often much smaller than the ambient dimension d. However, turning this characterization into an efficient procedure appears difficult; their Algorithm 2 constructs a minimum-size global sufficient dataset by solving a mixed-integer program at each iteration. In a recent work (Bennouna et al., 2026), whether a basis of $W_\star$ can be constructed by a polynomial-time algorithm is stated as an open problem.

In Section 3, we show that computing $d^\star$ is NP-hard (Theorem 5) and that even the weakest global sufficiency test—deciding whether the empty dataset $\mathcal{D} = \varnothing$ is globally sufficient—is coNP-hard already when $\mathcal{C}$ is an open polyhedron specified in H-representation[1] (Theorem 10). Under P $\neq$ NP, this gives a negative answer to the open problem in Bennouna et al. (2026): it is computationally hard in general to construct a basis of $W_\star$ and hence a minimum-size global sufficient dataset. The hardness results establish a computational barrier for worst-case global sufficiency but leave open two important questions: (i) Can we efficiently construct sufficient datasets for individual cost instances? (ii) In a data-driven regime with typical costs, can we learn datasets with good average-case guarantees?

We answer both affirmatively, demonstrating a computational-statistical gap between worst-case complexity and average-case learnability. Following the data-driven algorithm design paradigm (Gupta and Roughgarden, 2017, 2020; Balcan, 2021), we relax from *global (uniform)* sufficiency over all $c \in \mathcal{C}$ to a *distributional* notion: we assume costs are drawn i.i.d. from an unknown distribution P_c supported on $\mathcal{C}$, and aim to learn a dataset $\hat{\mathcal{D}}$ that is sufficient for a fresh draw $c \sim P_c$ with high probability. We address this via two intermediate steps: (i) a tractable *pointwise relaxation* for individual instances (answering question (i) above), and (ii) a cumulative learning algorithm with PAC guarantees (answering question (ii)). This leads to a sample-complexity question:

How many training instances are needed to learn such a $\hat{\mathcal{D}}$ in polynomial time?

To obtain a polynomial-time learning algorithm from realized cost samples, we introduce *pointwise sufficiency*: every cost vector still compatible with the observed measurements of a realized c must support a common optimal decision. Equivalently, the remaining fiber lies inside a single optimality cone. This relaxation is tractable because it only asks whether the fiber lies in one optimality cone. Under a standard nondegeneracy assumption on $\mathcal{X}$, checking pointwise sufficiency reduces to a geometric containment test that solves at most $d - m$ instances of the face-intersection (FI) subproblem for a fixed optimal basis. When $\mathcal{C}$ is a polytope given in explicit H-representation or an ellipsoid, each FI call can be solved in polynomial time; see Property 19.

1. An *H-polyhedron* is specified by finitely many linear inequalities, e.g., $P = \{x \in \mathbb{R}^d : Hx \leq h\}$. A *polytope* is a bounded polyhedron; we use *H-polytope* when emphasizing an inequality representation of a polytope. An *open polyhedron* is obtained by replacing non-strict inequalities by strict ones.

In Section 4, we develop a sequential offline cutting-plane algorithm (Algorithm 1) that discovers pointwise-sufficient datasets. The key algorithmic innovation is our facet-hit cutting-plane rule: when the containment test fails, we identify the first facet of the optimality cone encountered along the segment joining an interior point to the witness of failure. This ensures the queried facet normal is genuinely decision-relevant. A naive approach querying arbitrary violated constraints can fail, as we demonstrate via a counterexample in Appendix C.6. Each query produces a direction linearly independent of previous ones, and the procedure terminates after at most $d^\star$ queries.

In Section 5, in a data-driven regime with i.i.d. costs, we run the pointwise routine cumulatively with warm starts (Algorithm 2) and expand the measurement set only on a small subset of "hard" instances. This yields a realizable stable compression scheme (Hanneke and Kontorovich, 2021) with compression size at most $d^\star$. Consequently, with probability at least $1-\delta$ over n training samples, the pointwise-sufficiency failure probability on a fresh draw is at most $\widetilde{O}(d^\star/n)$ (Theorem 23). This fast rate exploits realizability: our algorithm achieves zero empirical sufficiency loss via the compression framework, contrasting with data-driven projection approaches for LPs (Sakaue and Oki, 2024; Iwata and Sakaue, 2025) and broader data-driven algorithm design analyses (Balcan et al., 2024a) that control objective value gaps and obtain characteristic $\widetilde{O}(1/\sqrt{n})$ rates via uniform convergence. In practice, many decision systems repeatedly solve linear programs over a fixed feasible polytope $\mathcal{X}$ while the cost vector $c \in \mathcal{C}$ varies. Examples include (i) *repeated LP* with time-varying costs (routing, resource allocation), (ii) *contextual linear optimization*, where one learns a predictor for c and plugs it into the LP, and (iii) preference-based decision making, where feedback is limited to comparisons or other aggregate signals. This motivates the two-stage pipeline below:

- **Stage I (Representation Discovery).** Learn $W_\star$ from i.i.d. training instances via sequential offline linear queries with a distribution-free certificate on the probability of sufficiency failure.
- **Stage II (Task-Specific Deployment).** Use the learned subspace for dimension reduction in repeated LPs, contextual LPs, and preference-based decision making.

In Section 6, as an illustration, we integrate this framework into SPO+ training for contextual linear optimization by restricting to the discovered decision-relevant subspace. This reduces the dimension parameter appearing in the generalization bound (see Theorem 25).

We summarize our main contributions as follows, and defer the literature review to Appendix A.

- **Hardness results.** Building on the SDD geometry of Bennouna et al. (2025a), we show that computing the intrinsic decision-relevant dimension $d^\star$ is NP-hard (Theorem 5). We also prove that deciding whether the empty dataset $\mathcal{D} = \varnothing$ is globally sufficient (in the decision sense) is coNP-hard (Theorem 10); for open convex $\mathcal{C}$, this yields that constructing a minimum-size global SDD is NP- and coNP-hard (Corollary 11). Finally, verifying pointwise sufficiency is coNP-hard in general (Theorem 9).
- **Per-instance level: finding pointwise SDDs in polynomial time.** Under nondegeneracy, we design a facet-hit cutting-plane algorithm (Algorithm 1) that sequentially queries decision-relevant directions. The facet-hit rule ensures each query is linearly independent and lies in $W_\star$, returning a pointwise-sufficient dataset (Theorem 18) in polynomial time.
- **Distributional level: learning decision-sufficient representations with fast rates.** Warm-starting the pointwise routine over i.i.d. costs (Algorithm 2) yields a realizable stable compression scheme (Hanneke and Kontorovich, 2021) with compression size at most $d^\star$. This leads to a distribution-free PAC certificate scaling as $\widetilde{O}(d^\star/n)$ (Theorem 23), matching our lower bound

(Theorem 24). The fast rate exploits zero empirical loss and linear independence of decision-relevant queries.

- **Application to contextual linear optimization.** We integrate decision-sufficient representation into predictor training for contextual linear optimization, yielding a compressed prediction model with improved generalization guarantees. In particular, the dimension term in the bound is reduced from d to $d^\star$ (Theorem 25).

2. Preliminaries: Characterizing sufficient datasets via LP geometry

Recall that the fundamental question we address is: *which fixed measurement sets $\mathcal{D}$ are sufficient to solve the LP*? In other words, when do the measurements produced by one dataset $\mathcal{D}$, together with the prior restriction $c \in \mathcal{C}$, already determine an optimal decision? Following Bennouna et al. (2025a), we adopt a *global* notion of informativeness in which a single fixed dataset $\mathcal{D}$ must work uniformly for all $c \in \mathcal{C}$. We now formalize this notion.

Definition 1 (Global sufficient decision dataset) *A dataset $\mathcal{D} := \{q_1, \ldots, q_l\} \subseteq \mathbb{R}^d$ is a **sufficient decision dataset** (SDD) for the uncertainty set $\mathcal{C} \subseteq \mathbb{R}^d$ and decision set $\mathcal{X}$ if there exists a decision rule $\widehat{X} : \mathbb{R}^l \to \mathcal{P}(\mathcal{X})$, where $\mathcal{P}(\mathcal{X})$ denotes the collection of subsets of $\mathcal{X}$, such that*

$$\forall\, c \in \mathcal{C}, \qquad \widehat{X}\big(c^\top q_1, \ldots, c^\top q_l\big) \;=\; \arg\min_{x \in \mathcal{X}} c^\top x.$$

To characterize sufficient datasets, we recall a few notions from LP geometry. Let $\mathcal{X}^\angle$ denote the set of *extreme points* of $\mathcal{X}$. For $x^\star \in \mathcal{X}^\angle$, define the feasible direction cone $FD(x^\star) := \{\delta \in \mathbb{R}^d : \exists \varepsilon > 0,\ x^\star + \varepsilon\delta \in \mathcal{X}\}$ and define the *optimality cone* $\Lambda(x^\star) := \{c \in \mathbb{R}^d :\ x^\star \in \arg\min_{x \in \mathcal{X}} c^\top x\}$.

Let $D(x^\star)$ be the set of *extreme directions* of the polyhedral cone $FD(x^\star)$. For $\delta \in D(x^\star)$, define the corresponding boundary face $F(x^\star, \delta) := \Lambda(x^\star) \cap \{\delta\}^\perp = \{c \in \Lambda(x^\star) :\ c^\top \delta = 0\}$, and the set of *relevant extreme directions*

$$\Delta(\mathcal{X}, \mathcal{C}) := \Big\{\delta \in \mathbb{R}^d :\ \exists x^\star \in \mathcal{X}^\angle,\ \delta \in D(x^\star),\ \text{and } F(x^\star, \delta) \cap \mathcal{C} \neq \emptyset\Big\}. \tag{2}$$

Equivalently, δ is relevant if the corresponding boundary face between optimality regions is attainable by some cost in $\mathcal{C}$. For any cost vector c, let $\mathcal{X}^\star(c) := \arg\min_{x \in \mathcal{X}} c^\top x$ denote the (possibly set-valued) set of optimal solutions. For a set $\mathcal{C} \subseteq \mathbb{R}^d$, define

$$\mathcal{X}^\star(\mathcal{C}) := \bigcup_{c \in \mathcal{C}} \big(\mathcal{X}^\star(c) \cap \mathcal{X}^\angle\big), \quad \dim(\mathcal{X}^\star(\mathcal{C})) := \operatorname{span}\{x - x' :\ x, x' \in \mathcal{X}^\star(\mathcal{C})\}. \tag{3}$$

We refer to $W_\star := \dim(\mathcal{X}^\star(\mathcal{C}))$ as the *decision-relevant subspace*, and set $d^\star := \dim(W_\star)$. The following results are due to Bennouna et al. (2025a). The first theorem states that global sufficiency is equivalent to spanning all directions along which optimality can change.

Theorem 2 (Bennouna et al. (2025a), Theorem 1) *Let $\mathcal{C}$ be open and convex. A dataset $\mathcal{D}$ is an SDD for $(\mathcal{X}, \mathcal{C})$ if and only if*

$$\Delta(\mathcal{X}, \mathcal{C}) \subseteq \operatorname{span}(\mathcal{D}).$$

While $\Delta(\mathcal{X}, \mathcal{C})$ is defined via faces of optimality cones, the next theorem gives an equivalent characterization directly in decision space: the span of these relevant directions matches the span of *differences of reachable optima*.

Theorem 3 (Bennouna et al. (2025a), Theorem 2) *For any convex set $\mathcal{C} \subset \mathbb{R}^d$,*

$$\operatorname{span} \Delta(\mathcal{X}, \mathcal{C}) = \operatorname{dir}(\mathcal{X}^\star(\mathcal{C})).$$

Combining Theorems 2 and 3 yields the following subspace characterization.

Corollary 4 (Subspace characterization Bennouna et al. (2025a), Corollary 1) *Let $\mathcal{C}$ be open and convex. A dataset $\mathcal{D}$ is an SDD for $(\mathcal{X}, \mathcal{C})$ if and only if*

$$\operatorname{dir}(\mathcal{X}^\star(\mathcal{C})) \subseteq \operatorname{span}(\mathcal{D}),$$

hence, the minimal size of global SDD is $d^\star$.

In particular, the necessity direction of Theorem 2 uses that $\mathcal{C}$ is open. Accordingly, any argument or algorithm that invokes this direction requires an openness assumption on $\mathcal{C}$. In contrast, Theorem 3 requires only convexity and applies to both open and closed convex sets.[2]

3. Hardness and Relaxation

3.1. Computational hardness.

The geometric characterizations in Section 2 are information-theoretic in nature: they identify the subspace that any global sufficient dataset must capture. Algorithmically, a natural goal is to compute the minimum size of a global SDD and to construct one. By Corollary 4, when $\mathcal{C}$ is open and convex, this minimum size equals the intrinsic decision-relevant dimension $d^\star = \dim \operatorname{dir}\big(\mathcal{X}^\star(\mathcal{C})\big)$. We state informal versions of our hardness results below; full statements and proofs are deferred to Appendix B. Our reductions construct highly structured instances: shortest-path flow polytopes with budgeted arc-length perturbations.

Theorem 5 (Informal) *It is* NP*-hard to compute the intrinsic decision-relevant dimension $d^\star$, given as input a polytope $\mathcal{X} \subseteq \mathbb{R}^d$ and a polyhedral uncertainty set $\mathcal{C}$ specified in H-representation. This hardness persists whether $\mathcal{C}$ is given as a closed polyhedron or as an open polyhedron.*

When $\mathcal{C}$ is open and convex (in particular, for the open polyhedral instances in Theorem 5), combining Corollary 4 and Theorem 5 implies that both computing the size of a minimum global SDD and constructing a minimum global SDD are NP-hard.

3.2. A relaxation: pointwise sufficient datasets

The computational hardness of finding minimum global SDDs motivates us to relax the concept to an instance-wise notion:

Definition 6 (Pointwise sufficient decision dataset) *Fix a (possibly unknown) $c \in \mathcal{C}$. A finite query dataset $\mathcal{D}$ is **pointwise sufficient at** c if there exists a decision $x^\star \in \mathcal{X}$ such that*

$$x^\star \in \mathcal{X}^\star(c') \qquad \forall\, c' \in \mathcal{C}\big(\mathcal{D}, s(c; \mathcal{D})\big).$$

Equivalently, the data-consistent fiber $\mathcal{C}(\mathcal{D}, s(c; \mathcal{D}))$ is contained in a single optimality region of $\mathcal{X}$.

2. This distinction is important: our pointwise-sufficiency algorithms (Section 4) leverage Theorem 3 under a closed convex prior, while the global SDD size characterization requires $\mathcal{C}$ to be open and convex.

Remark 7 *Since $c \in \mathcal{C}(\mathcal{D}, s(c;\mathcal{D}))$, Definition 6 can equivalently be stated as: $\mathcal{D}$ is pointwise sufficient at c if there exists $x^\star \in \mathcal{X}^\star(c)$ such that $x^\star$ remains optimal for every $c' \in \mathcal{C}$ that produces the same measurements, i.e., $s(c';\mathcal{D}) = s(c;\mathcal{D})$. This parallels the structure of Definition 1 but applies to a single cost vector rather than all $c \in \mathcal{C}$.*

The following basic properties of pointwise sufficiency follow immediately from the definition.

Property 8 ***(i) Monotonicity.*** *If $\mathcal{D}$ is pointwise sufficient at c and $\mathcal{D}' \supseteq \mathcal{D}$, then $\mathcal{D}'$ is also pointwise sufficient at c, since $\mathcal{C}(\mathcal{D}', s(c;\mathcal{D}')) \subseteq \mathcal{C}(\mathcal{D}, s(c;\mathcal{D}))$.* ***(ii) Global $\Rightarrow$ pointwise.*** *If $\mathcal{D}$ is an SDD for $(\mathcal{X}, \mathcal{C})$, then $\mathcal{D}$ is pointwise sufficient at every $c \in \mathcal{C}$.*

A natural verification problem is: given $(\mathcal{X}, \mathcal{C})$, a dataset $\mathcal{D}$, and a cost vector $c \in \mathcal{C}$, decide whether $s(c;\mathcal{D})$ already suffices to determine an optimal decision. The next theorem shows that even the verification problem is intractable in full generality.

Theorem 9 (Informal) *It is* coNP*-hard to decide, given a bounded polytope $\mathcal{X} \subseteq \mathbb{R}^d$, a polyhedral uncertainty set $\mathcal{C} \subseteq \mathbb{R}^d$ specified in H-representation, a dataset $\mathcal{D}$, and a cost vector $c \in \mathcal{C}$, whether $\mathcal{D}$ is pointwise sufficient at c.*

The proof of Theorem 9 proceeds via a reduction from the classical *H-in-V polytope containment* problem.[3] In our hard instance, the V-polytope is induced by the optimality cone of an extreme point in a highly degenerate LP instance, and we take $\mathcal{C}$ to be the specific H-polytope. The same containment-based reduction also shows that even the weakest global problem—deciding whether *no data* ($\mathcal{D} = \varnothing$) already suffices—is coNP-hard.

Theorem 10 (Informal) *It is* coNP*-hard to decide whether the empty dataset $\mathcal{D} = \varnothing$ is globally sufficient for $(\mathcal{X}, \mathcal{C})$, when $\mathcal{X} \subseteq \mathbb{R}^d$ is a bounded polytope and $\mathcal{C}$ is an open polyhedron specified in H-representation. Moreover, computing the intrinsic decision-relevant dimension $d^\star$ is* coNP*-hard.*

Consequently, computing the minimum size of global SDDs and outputting such a dataset are coNP-hard. Combining Theorems 5 and 10 with Corollary 4 yields the following final statement.

Corollary 11 *For convex and open $\mathcal{C}$, constructing a minimum-size global SDD and computing its minimum size are both* NP*-hard and* coNP*-hard.*

The coNP-hardness in Theorem 9 arises from degenerate LP geometry, where a single optimal extreme point may correspond to many bases, and its optimality region can have a complicated representation. Going forward, we adopt a standard nondegeneracy assumption to recover tractability.

Assumption 12 *The polytope $\mathcal{X} = \{x \in \mathbb{R}^d : Ax = b,\ x \geq 0\}$ is nondegenerate: every extreme point $x^\star \in \mathcal{X}^\angle$ has exactly m strictly positive components.*

Under Assumption 12, each extreme point has a single associated optimality cone that admits a simple linear inequality description. This makes it possible to test whether a fiber $\mathcal{C}(\mathcal{D}, s)$ is contained in a candidate cone by solving only a polynomial number of LPs (as we do in Algorithm 1). In contrast, without Assumption 12, even deciding pointwise sufficiency is coNP-hard by Theorem 9.

3. A *V-polytope* is a polytope specified by its vertices (a vertex representation), e.g., $Q = \mathrm{conv}\{v_1, \ldots, v_M\}$. The H-in-V containment problem asks whether $P \subseteq Q$ given (H, h) and $\{v_j\}_{j=1}^M$.

4. A Tractable Algorithm that Finds Pointwise SDD

In this section, we present a sequential offline cutting-plane routine that constructs a non-adaptive pointwise-sufficient dataset for a fixed (and possibly unknown) cost vector $c \in \mathcal{C}$. The routine maintains the data-consistent fiber $\mathcal{C}_k = \mathcal{C}(\mathcal{D}, s(c;\mathcal{D}))$, checks whether $\mathcal{C}_k$ lies inside one optimality cone, and if not, adds the normal of a reachable violated facet. Under Assumption 12, each iteration solves one LP over $\mathcal{X}$ and at most $d-m$ convex minimization subproblems over the current fiber. Every new queried direction is decision-relevant and linearly independent of the previous ones, so the procedure makes at most $d^\star$ queries. Unlike the global SDD characterization in Section 2, pointwise sufficiency is a containment statement about a single fiber and does not require $\mathcal{C}$ to be open. Accordingly, the only structural assumption we use on the prior set is convexity; we take it closed only for algorithmic convenience (Remark 14). Remaining proofs appear in Appendix C.

Assumption 13 *The uncertainty set $\mathcal{C} \subseteq \mathbb{R}^d$ is convex.*

4.1. Pointwise sufficiency as optimality-cone containment

To make the logic of the algorithm transparent, we first translate pointwise sufficiency into a cone-containment test around one optimal basis. Fix a basis $B \subseteq \{1,\dots,d\}$ with $|B| = m$ and let N denote its complement. Write $A = [A_B\ A_N]$, and let A_j be the jth column of A. The corresponding basic feasible solution is $x(B)$ with $x_N(B) = 0$ and $x_B(B) = A_B^{-1}b$. For each nonbasic index $j \in N$, let $\delta(B,j) \in \mathbb{R}^d$ denote the standard edge direction obtained by increasing x_j from 0, i.e., $\delta_N(B,j) = e_j$ and $\delta_B(B,j) = -A_B^{-1}A_j$. Under Assumption 12, feasible bases are in one-to-one correspondence with vertices of $\mathcal{X}$. Moreover, for any feasible basis B, the optimality region of its corresponding $x(B)$ is the polyhedral cone

$$\Lambda(B) := \{c \in \mathbb{R}^d :\ c^\top \delta(B,j) \geq 0\ \ \forall j \in N\}. \tag{4}$$

After querying directions $q_1,\dots,q_k$, let $Q_k = [q_1\ \cdots\ q_k] \in \mathbb{R}^{d\times k}$ and $s_k = Q_k^\top c$. The current fiber is $\mathcal{C}_k := \{c' \in \mathcal{C} :\ Q_k^\top c' = s_k\}$. By Definition 6, the dataset is pointwise sufficient at c once there exists a basis B such that $\mathcal{C}_k \subseteq \Lambda(B)$. Using (4), this containment reduces to checking that $\min_{c'\in\mathcal{C}_k}(c')^\top\delta(B,j) \geq 0$ for all $j \in N$, which motivates the face-intersection subproblem below.

The face-intersection (FI) subproblem. For a direction $\delta \in \mathbb{R}^d$ and a fiber $\mathcal{C}_k$, define

$$m_k(\delta)\ :=\ \inf_{c'\in\mathcal{C}_k} (c')^\top\delta. \tag{5}$$

When the infimum is attained, write $c_k^{\text{out}}(\delta) \in \arg\min_{c'\in\mathcal{C}_k}(c')^\top\delta$; otherwise, any $c_k^{\text{out}}(\delta) \in \mathcal{C}_k$ with $(c_k^{\text{out}}(\delta))^\top\delta < 0$ is a valid witness when $m_k(\delta) < 0$. Thus $m_k(\delta) < 0$ iff $\mathcal{C}_k$ intersects the open halfspace $\{c : c^\top\delta < 0\}$. Under Assumption 13, the value problem defining $m_k(\delta)$ is convex. In the algorithmic settings of interest, when $\mathcal{C}$ is polyhedral it is an LP; when $\mathcal{C}$ is an ellipsoid, $\mathrm{FI}(\delta;\mathcal{C}_k)$ admits a closed form (Proposition 33).

4.2. A facet-hit cutting-plane algorithm

Algorithm 1 gives the full procedure. At each iteration it anchors at the realized cost vector c (i.e., we take $c^{\text{in}} := c \in \mathcal{C}_k$), solves the LP under c^{in} to obtain an optimal vertex solution $x(B)$; by Assumption 12, this vertex uniquely determines the corresponding feasible basis B, and tests

Algorithm 1 A cutting-plane algorithm that finds pointwise SDD

Input: LP data (A, b), prior set $\mathcal{C}$, a cost vector $c \in \mathcal{C}$, and an initial dataset $\mathcal{D}_{\text{init}} \subset \mathbb{R}^d$.

1: Initialize $\mathcal{D} \leftarrow \mathcal{D}_{\text{init}}$.
2: Let $k \leftarrow |\mathcal{D}|$; form $Q_k = [q_1 \ \cdots \ q_k]$ from $\mathcal{D}$ (any fixed order) and set $s_k \leftarrow Q_k^\top c$.
3: **while** true **do**
4: Set $\mathcal{C}_k := \{c' \in \mathcal{C} : Q_k^\top c' = s_k\}$ and set $c^{\text{in}} \leftarrow c$.
5: Solve $\min\{(c^{\text{in}})^\top x : Ax = b,\ x \geq 0\}$ and obtain an optimal basis B with nonbasis N.
6: **(Containment test via face-intersection subproblem)** For each $j \in N$, form $\delta_j := \delta(B, j)$ and compute

$$m_j := \min_{c' \in \mathcal{C}_k} (c')^\top \delta_j, \quad c_j^{\text{out}} \in \arg\min_{c' \in \mathcal{C}_k} (c')^\top \delta_j. \qquad (\text{FI}(\delta_j; \mathcal{C}_k))$$

7: Let $j_0 \in \arg\min_{j \in N} m_j$, set $m_{\min} := m_{j_0}$, $c^{\text{out}} := c_{j_0}^{\text{out}}$.
8: **if** $m_{\min} \geq 0$ **then**
9: **Return** dataset $\mathcal{D}$ and certificate basis B (decision $x(B)$); **break**.
10: **else**
11: **(Facet-hit rule)** For each $j \in N$ with $(c^{\text{out}})^\top \delta_j < 0$, set $\alpha_j := \frac{(c^{\text{in}})^\top \delta_j}{(c^{\text{in}})^\top \delta_j - (c^{\text{out}})^\top \delta_j} \in [0, 1)$.
12: Let $\alpha^\star := \min \alpha_j$ and pick any $j^\star \in \arg\min \alpha_j$.
13: Set $q_{k+1} := \delta_{j^\star}$ and set $\sigma_{k+1} \leftarrow q_{k+1}^\top c$.
14: Update: $\mathcal{D} \leftarrow \mathcal{D} \cup \{q_{k+1}\}$, $Q_{k+1} \leftarrow [Q_k \ \ q_{k+1}]$, $s_{k+1} \leftarrow (s_k^\top, \ \sigma_{k+1})^\top$, $k \leftarrow k + 1$.
15: **end if**
16: **end while**

whether *every* cost vector in the fiber is contained in the corresponding cone $\Lambda(B)$. This is done by evaluating each facet inequality via the face-intersection subproblem. If the minimum violation is nonnegative, the fiber is fully inside $\Lambda(B)$ and pointwise sufficiency holds. Otherwise, a witness point $c^{\text{out}} \in \mathcal{C}_k$ may violate several facet inequalities of $\Lambda(B)$.

Remark 14 (Closedness is not necessary) *Pointwise sufficiency and Algorithm 1 interact with the prior only through containment tests of the form $\mathcal{C}_k \subseteq \Lambda(B)$. Since $\Lambda(B)$ is closed, $\mathcal{C}_k \subseteq \Lambda(B)$ holds iff $\text{cl}(\mathcal{C}_k) \subseteq \Lambda(B)$. Thus, for these containment tests, one may replace a (possibly nonclosed) fiber by its closure without changing any certification outcome. Moreover, any violating witness $c_j^{\text{out}} \in \mathcal{C}_k \setminus \Lambda(B)$ (whenever it exists) remains valid after this replacement.*

The facet-hit rule identifies the first facet of $\Lambda(B)$ encountered along the segment $[c^{\text{in}}, c^{\text{out}}]$, and thus guarantees the existence of a boundary point $c^{\text{hit}} \in \mathcal{C}_k \cap \Lambda(B)$ with $(c^{\text{hit}})^\top \delta(B, j^\star) = 0$ (Lemma 16), which is what makes the new query direction decision-relevant. A concrete counterexample showing that an arbitrary violated facet can fail is given in Appendix C.6.

Remark 15 (Picking arbitrary $c^{\text{in}} \in \mathcal{C}_k$) *Algorithm 1 is stated in a* fixed-anchor *form (we set $c^{\text{in}} := c$ at each iteration). The same facet-hit cutting-plane idea and all results in this section also apply when the cost vector c is* unknown *and one only has oracle access to inner products $q^\top c$: in that case, line 4 may pick* any *anchor $c^{\text{in}} \in \mathcal{C}_k$, and then proceed identically.*

4.3. Correctness and basic properties

We now record a few basic facts about Algorithm 1. The next two lemmas are technical but important for our later analysis in Section 5: Lemma 16 shows that every new queried direction is

genuinely decision-relevant. Lemma 17 guarantees that the dataset grows with linearly independent directions. Theorem 18 shows the algorithm terminates and indeed certifies pointwise sufficiency.

Lemma 16 *In Algorithm 1, any newly added direction q_{k+1} lies in $\Delta(\mathcal{X}, \mathcal{C}) \subseteq \mathrm{dir}(\mathcal{X}^\star(\mathcal{C}))$.*

Lemma 17 *In Algorithm 1, the queried directions are linearly independent. In particular, the algorithm makes at most $d^\star$ queries.*

Theorem 18 (Correctness) *Algorithm 1 terminates after at most $d^\star + 1$ iterations and returns a dataset $\mathcal{D}$ that is pointwise sufficient at the (possibly unknown) c.*

Property 19 *Assume that $\mathcal{C}$ is either (i) a polytope given in H-representation, or (ii) an ellipsoid. Then Algorithm 1 can be implemented to run in time **polynomial** in the input size. In particular, it makes at most $d^\star$ oracle queries of the form $q^\top c$ and solves at most $(d^\star + 1)$ LPs over $\mathcal{X}$ and $(d^\star + 1)(d - m)$ face-intersection subproblems (LPs in case (i), and closed form in case (ii)).*

5. Learning from Distributional Data

Section 4 constructs a pointwise-sufficient query set for a single cost vector c. We now move to a data-driven regime in which the LP (1) is solved repeatedly with random $c \sim P_c$ supported on $\mathcal{C}$, and we observe i.i.d. samples $c_1, \ldots, c_n$. Our goal is to learn a *compressed, decision-sufficient representation*: a small set of query directions that are pointwise sufficient for a fresh draw $c \sim P_c$ with high probability, together with a *distribution-free certificate* on its failure probability. Proofs for this section appear in Appendix D.

For a given dataset $\mathcal{D}$ (query directions), define the *0–1 sufficiency loss*

$$\ell(\mathcal{D}, c) \;:=\; \mathbf{1}\{\mathcal{D} \text{ is not pointwise sufficient at } c\}. \tag{6}$$

We aim to output a dataset $\mathcal{D}$ with small *risk* $R(\mathcal{D}) := \mathbb{P}_{c\sim P_c}[\ell(\mathcal{D}, c) = 1]$.

5.1. A cumulative algorithm

We now run the pointwise cutting-plane routine sequentially on each training sample c_i and *accumulate* the queried directions. Algorithm 2 initializes with an empty dataset and, for $i = 1, \ldots, n$, invokes Algorithm 1 on c_i using the current dataset as a warm start. If c_i is already certified by the current dataset, nothing changes; otherwise, new directions are added until c_i becomes pointwise sufficient. We call an index i *hard* if processing c_i adds at least one new direction, i.e., $\mathcal{D}_i \neq \mathcal{D}_{i-1}$.

The next three lemmas formalize the learning-theoretic structure underlying this cumulative procedure. Together, they show that Algorithm 2 induces a *stable, realizable* sample compression scheme (Hanneke and Kontorovich, 2021, Definitions 7–8) of compression size at most $d^\star$. This is the key mechanism we use in the next subsection to obtain a distribution-free fast-rate certificate on the true failure probability $R(\mathcal{D}_n) = \Pr_{c\sim P_c}[\ell(\mathcal{D}_n, c) = 1]$.

Lemma 20 (Realizability) *Algorithm 2 returns $\mathcal{D}_n$ with $\ell(\mathcal{D}_n, c_i) = 0$ for all $i = 1, \ldots, n$.*

Lemma 21 (Stability) *The final dataset $\mathcal{D}_n$ returned by Algorithm 2 is fully determined by the compressed subsequence $(c_i)_{i\in T}$, independent of the remaining samples.*

Lemma 22 (Compression size bound) *Under Assumption 12, $|T| \le |\mathcal{D}_n| \le d^\star$.*

Algorithm 2 Learning sufficient decision datasets over samples

Input: Prior $\mathcal{C}$, LP data (A, b), i.i.d. samples $c_1, \ldots, c_n$ (via oracle access to $q^\top c_i$).
1: Initialize dataset $\mathcal{D}_0 \leftarrow \emptyset$, hard index set $T \leftarrow \emptyset$.
2: **for** $i = 1$ to n **do**
3: Run Algorithm 1 on c_i with initialization $\mathcal{D}_{\text{init}} = \mathcal{D}_{i-1}$. Let the returned dataset be $\mathcal{D}_i$.
4: **if** $\mathcal{D}_i \neq \mathcal{D}_{i-1}$ **then**
5: Mark i as hard: $T \leftarrow T \cup \{i\}$.
6: **end if**
7: **end for**
8: **Return** final dataset $\mathcal{D}_n$ and hard set T.

5.2. A distribution-free certificate via stable compression

We can now invoke the clean stable-compression generalization bound of Hanneke and Kontorovich (2021, Corollary 11) to certify the true failure probability $R(\mathcal{D}_n) = \mathbb{P}_{c\sim P_c}[\ell(\mathcal{D}_n, c) = 1]$. For an alternative viewpoint, see Campi and Garatti, 2023.

Theorem 23 (Certificate via stable compression) *For any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the draw of $c_1, \ldots, c_n$, the output $(\mathcal{D}_n, T)$ of Algorithm 2 satisfies*

$$R(\mathcal{D}_n) \;\le\; \frac{4}{n}\Big(6|T| + \ln\frac{e}{\delta}\Big) \;\le\; \frac{4}{n}\Big(6d^\star + \ln\frac{e}{\delta}\Big), \tag{7}$$

where the second inequality uses $|T| \le d^\star$ from Lemma 22.

Equation (7) gives the fast-rate certificate $R(\mathcal{D}_n) \le O\big((d^\star + \ln(1/\delta))/n\big)$. Recent data-driven projection approaches for LPs learn a low-dimensional embedding that preserves feasibility and objective values approximately and then control a downstream error via uniform-convergence-style analyses; this leads to the characteristic $1/\sqrt{n}$ dependence on sample size and complexity terms tied to the pseudo-dimension of the learned projection family (Balcan et al., 2024a; Sakaue and Oki, 2024). Our guarantee is complementary. We target decision sufficiency (a binary property) and exploit polyhedral geometry to ensure only decision-relevant directions can ever be queried. This guarantees a stable compression scheme, yielding $\widetilde{O}(d^\star/n)$ certificates that depend on the intrinsic dimension $d^\star$. This fast rate is also characteristic of realizability; in agnostic settings where zero empirical loss is unattainable, compression-based methods are subject to $\Omega\big(\sqrt{k\log(n/k)/n}\big)$ lower bounds (Hanneke and Kontorovich, 2019), where k is the compression size.

Finally, it is natural to ask whether the dependence on $d^\star$ and n can be improved. The next theorem shows that, at least for the concrete cumulative procedure in Algorithm 2, the fast-rate certificate is tight up to constants: one needs $n = \Omega(d^\star/\varepsilon)$ samples to drive the pointwise-sufficiency failure probability below ε with constant confidence. At a high level, the proof follows the classical "rare-types" construction used to obtain lower bounds for realizable sample-compression schemes (Littlestone and Warmuth, 1986). However, we cannot simply import a generic compression lower bound as a black box, because here the compression map is not arbitrary: it must arise from LP optimality geometry through Algorithm 1. Accordingly, we explicitly construct a linear program, a convex prior set $\mathcal{C}$ with $\dim(\mathrm{dir}(\mathcal{X}^\star(\mathcal{C}))) = d^\star$, and a distribution P_c over c such that each rare event forces the discovery of a distinct decision-relevant direction.

Theorem 24 *Fix any integer $d^\star \geq 2$ and any $\varepsilon \in (0, 1/4)$. There exist an ambient dimension $d \geq d^\star$, a nondegenerate LP polytope $\mathcal{X} \subseteq \mathbb{R}^d$, a convex uncertainty set $\mathcal{C} \subseteq \mathbb{R}^d$ with $\dim(\mathrm{dir}(\mathcal{X}^\star(\mathcal{C}))) = d^\star$, and a distribution P_c supported on $\mathcal{C}$ such that the following holds. If $\mathcal{D}_n$ is the output of Algorithm 2 on n i.i.d. samples from P_c and $n \leq \frac{d^\star - 1}{8\varepsilon}$, then $\mathbb{P}(R(\mathcal{D}_n) > \varepsilon) \geq \frac{1}{2}$.*

6. Application: Model Compression for Contextual Linear Optimization

In this section, we illustrate how decision-sufficient representations yield a principled model-compression guarantee for contextual linear optimization (CLO). Let $(\xi, c) \sim P$, where $\xi \in \Xi \subseteq \mathbb{R}^p$ is a context and $c \in \mathcal{C} \subseteq \mathbb{R}^d$ (a.s.) is the cost vector of a downstream linear program over a known bounded polytope $\mathcal{X} \subseteq \mathbb{R}^d$. We assume that $\mathcal{C}$ is convex. Let P_c denote the marginal distribution of c. The goal in CLO is to train a predictor $f : \Xi \to \mathbb{R}^d$ from contextual data so that the induced plug-in decisions $x^\star(f(\xi))$ achieve low out-of-sample loss. Throughout, we fix a deterministic oracle $x^\star : \mathbb{R}^d \to \mathcal{X}^\angle$ such that $x^\star(v) \in \arg\min_{x \in \mathcal{X}} v^\top x$ for all v (e.g., with lexicographic tie-breaking).

Ellipsoidal prior with a canonical lifting map. Throughout this section, we specialize the prior set $\mathcal{C}$ to an ellipsoid $\mathcal{C} := \{c \in \mathbb{R}^d : \ (c - c_0)^\top \Sigma^{-1} (c - c_0) \leq 1\}$, for some $\Sigma \succ 0$, $c_0 \in \mathbb{R}^d$. Let $W \subseteq \mathbb{R}^d$ be any t-dimensional subspace with orthonormal basis $U \in \mathbb{R}^{d \times t}$. We define the *lifting matrix* $\mathcal{L}_U := \Sigma U (U^\top \Sigma U)^{-1}$, and the corresponding *canonical lifting map* (Appendix E.1)

$$\mathrm{lift}_U(s) := c_0 + \mathcal{L}_U s, \qquad s \in \mathbb{R}^t. \tag{8}$$

Two-stage pipeline. Our deployment follows a two-stage pipeline that separates representation discovery from predictor training. In *Stage I*, we recover a decision-sufficient subspace $\hat{W} \subseteq \mathbb{R}^d$ with orthonormal basis $\hat{U}$. In *Stage II*, we draw independent contextual samples $S := \{(\xi_i, c_i)\}_{i=1}^n \sim P$ and train a predictor that first predicts a coordinate in the learned subspace $\hat{W}$.

Throughout this section, we assume that Stage I recovers the decision-relevant subspace

$$W_\star := \mathrm{dir}(\mathcal{X}^\star(\mathcal{C})),$$

where $\mathcal{C}$ is the closed ellipsoidal prior above. Let $U_\star \in \mathbb{R}^{d \times d^\star}$ be an orthonormal basis for $W_\star$. In Appendix E.6, we give a data-driven Stage I construction from contextual samples (ξ, c) using Algorithm 2, which recovers an almost sufficient subspace under a margin condition and incurs an additional misspecification term.

6.1. SPO training in a decision-sufficient subspace

Given a predictor $\hat{c} = f(\xi)$, the plug-in decision is $x^\star(\hat{c})$ and the SPO loss is $\ell_{\mathrm{SPO}}(\hat{c}, c) := c^\top x^\star(\hat{c}) - c^\top x^\star(c) \geq 0$. The corresponding *SPO risk* of a predictor $f : \Xi \to \mathbb{R}^d$ is $R_{\mathrm{SPO}}(f) := \mathbb{E}_{(\xi,c)\sim P}\big[\ell_{\mathrm{SPO}}(f(\xi), c)\big] = \mathbb{E}_{(\xi,c)\sim P}\big[c^\top x^\star(f(\xi)) - c^\top x^\star(c)\big]$. In practice, given i.i.d. samples $S = \{(\xi_i, c_i)\}_{i=1}^n$ in Stage II, we train θ by minimizing either the empirical SPO risk $\hat{R}_{\mathrm{SPO}}(f) := \frac{1}{n}\sum_{i=1}^n \ell_{\mathrm{SPO}}(f(\xi_i), c_i)$ or its convex surrogate $\hat{R}_{\mathrm{SPO+}}(f) := \frac{1}{n}\sum_{i=1}^n \ell_{\mathrm{SPO+}}(f(\xi_i), c_i)$. Following Elmachtoub and Grigas (2022), a stochastic subgradient of $\ell_{\mathrm{SPO+}}(\hat{c}, c)$ can be computed with two oracle calls (see Appendix E.2).

Focusing on the decision-sufficient subspace, one can parametrize a cost predictor by first predicting a $d^\star$-dimensional centered coordinate $g_\theta : \Xi \to \mathbb{R}^{d^\star}$ and then lifting it back to $\mathbb{R}^d$ via

$$\hat{c}_\theta(\xi) = \mathrm{lift}_{U_\star}(g_\theta(\xi)) = c_0 + \mathcal{L}_{U_\star} g_\theta(\xi), \qquad \mathcal{L}_{U_\star} := \Sigma U_\star (U_\star^\top \Sigma U_\star)^{-1}. \tag{9}$$

For linear coordinate models of the form $g_B(\xi) = B\xi$ with $B \in \mathbb{R}^{d^\star \times p}$, this reduces the number of trainable parameters from dp to $d^\star p$. An explicit stochastic subgradient step is shown in Appendix E.3. Pseudocode for Stage II training is given in Appendix E.4.

6.2. Improved generalization bound

Consider the compressed affine-linear hypothesis class $\mathcal{H}_{U_\star,d^\star} := \{f_B(\xi) = c_0 + \mathcal{L}_{U_\star} B\xi : \ B \in \mathbb{R}^{d^\star \times p}\}$. We define the SPO range constant $\omega_{\mathcal{X}}(\mathcal{C}) := \sup_{c\in\mathcal{C}} \left(\max_{x\in\mathcal{X}} c^\top x - \min_{x\in\mathcal{X}} c^\top x\right)$. With these notations in place, we can state the following guarantee for Stage II training.

Theorem 25 *Let $\mathcal{H}_{U_\star,d^\star}$ be as above and define the $\mathcal{C}$-valued affine-linear classes*

$$\mathcal{H}(\mathcal{C}) := \{f_A(\xi) = c_0 + A\xi : \ A \in \mathbb{R}^{d\times p},\ f_A(\xi) \in \mathcal{C} \text{ a.s.}\}, \quad \mathcal{H}_{U_\star,d^\star}(\mathcal{C}) := \{f \in \mathcal{H}_{U_\star,d^\star} : \ f(\xi) \in \mathcal{C} \text{ a.s.}\}.$$

(1) ***No misspecification loss.*** *Let $f_\star \in \arg\min_{f\in\mathcal{H}(\mathcal{C})} R_{\mathrm{SPO}}(f)$ and define its compressed version $\hat{f}_\star(\xi) := \mathrm{lift}_{U_\star}\big(U_\star^\top (f_\star(\xi) - c_0)\big) = c_0 + \mathcal{L}_{U_\star} U_\star^\top (f_\star(\xi) - c_0) \in \mathcal{H}_{U_\star,d^\star}(\mathcal{C})$.*
Then $\hat{f}_\star \in \arg\min_{f\in\mathcal{H}_{U_\star,d^\star}(\mathcal{C})} R_{\mathrm{SPO}}(f)$ and $R_{\mathrm{SPO}}(f_\star) = R_{\mathrm{SPO}}(\hat{f}_\star)$.

(2) ***Improved generalization bound in the compressed class.*** *For any $\delta \in (0,1)$, with probability at least $1-\delta$ over S, the following bound holds* simultaneously for all $f \in \mathcal{H}_{U_\star,d^\star}$*:*

$$R_{\mathrm{SPO}}(f) \le \hat{R}_{\mathrm{SPO}}(f) + 2\,\omega_{\mathcal{X}}(\mathcal{C}) \sqrt{\frac{2(d^\star p + 1)\log(n|\mathcal{X}^{\angle}|^2)}{n}} + \omega_{\mathcal{X}}(\mathcal{C}) \sqrt{\frac{\log(1/\delta)}{2n}}. \quad (10)$$

The proof is in Appendix E.5. Compared to El Balghiti et al. (2023), the bound in (10) replaces the ambient dimension d in the dominant $\widetilde{O}(1/\sqrt{n})$ generalization error term by the decision-relevant intrinsic dimension $d^\star$. In Appendix E.6, we show how to estimate $\hat{W}$ in Stage I from n_I contextual samples (ξ, c). This introduces an additional representation-estimation error term of order $\widetilde{O}(n_I^{-\min\{1,\alpha/2\}})$ under the margin condition in Assumption 42 (see Theorem 44). A numerical experiment illustrating the resulting gains in sample efficiency is presented in Appendix E.8.

7. Concluding Remarks and Open Problems

This paper develops a computational and statistical theory of learning decision-sufficient representations for linear optimization. We show that computing the intrinsic decision-relevant dimension $d^\star$ and constructing minimum-size global SDDs are NP-hard and coNP-hard. To go beyond these worst-case barriers, we introduce a pointwise relaxation that admits a polynomial-time cutting-plane algorithm under nondegeneracy, and prove that the resulting cumulative learner yields a stable compression scheme with a distribution-free fast-rate certificate: with high probability over the training sample, the pointwise sufficiency failure probability on a fresh draw is at most $\tilde{O}(d^\star/n)$. We also illustrate how the learned representation can reduce the effective dimension in contextual linear optimization. We further propose the following open problems.

Hardness beyond degeneracy and approximability. Our current hardness results for computing the intrinsic decision-relevant dimension $d^\star$ (and hence constructing minimum-size global SDDs) rely on highly degenerate hard instances. It therefore remains open whether our hardness results persist under the nondegeneracy assumption on $\mathcal{X}$. Moreover, it is open whether one can compute a global SDD whose cardinality is within a constant factor of the minimum.

Adaptive versus non-adaptive decision information. Another open direction is to separate non-adaptive and adaptive notions of information. The parameter $d^\star$ concerns fixed batch datasets, whereas Algorithm 1 adaptively constructs a fixed pointwise-sufficient dataset and Algorithm 2 outputs a fixed dataset used non-adaptively at test time. It remains open to characterize the worst-case adaptive query complexity of pointwise sufficiency and to determine whether test-time branching-program strategies can provably use fewer than $d^\star$ queries per instance. Likewise, Theorem 24 is a lower bound for Algorithm 2. Characterizing the best computational-statistical guarantee achievable by arbitrary polynomial-time learning algorithms remains open.

Extension to noisy and robust settings. Our framework focuses on the noiseless setting, where each linear measurement $q_i^\top c$ is observed exactly. Extending our algorithms to noisy observations and characterizing the resulting sample complexity are open directions. Intuitively, such an extension would likely require an additional margin condition on P_c that controls how often c lies close to the boundary of an optimality cone.

Motivated by recent advances in robust learning and robust statistical estimation, this setting also raises a natural learning-theoretic question: can one develop robust sample-compression schemes for decision-sufficient representations? For instance, suppose that the learner observes corrupted costs $\widehat{c}$ whose distribution $\widehat{P}$ satisfies $d_{\mathrm{TV}}(\widehat{P}, P_c) \le \varepsilon$, which captures settings where an ε-fraction of the data may be arbitrary or even adversarial outliers. Can one design a robust learner with certificates? A possible route is to allow a controlled redundancy in the hard-sample budget, expanding the effective size from $d^\star$ to $d^\star + r(\varepsilon, n)$, so that the extra budget can absorb outlier-induced violations while preserving a stable-compression-style guarantee for the clean distribution.

SDD Geometry beyond LPs. Finally, we restrict attention to linear optimization with a fixed feasible region in this paper. Extending the decision-sufficient representation framework to broader problem classes, such as mixed-integer programs, quadratic programs, or LPs with varying constraints, is an important direction for future work.

Acknowledgments

We thank Omar Bennouna, Jiawei Zhang, and Amine Bennouna for useful discussions.

References

Maria-Florina Balcan. Data-driven algorithm design. In Tim Roughgarden, editor, *Beyond the Worst-Case Analysis of Algorithms*, pages 626–645. Cambridge University Press, 2021.

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 65–72, 2006.

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009.

Maria-Florina Balcan, Vaishnavh Nagarajan, Ellen Vitercik, and Colin White. Learning-theoretic foundations of algorithm configuration for combinatorial partitioning problems. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, volume 65 of *Proceedings of Machine Learning Research*, pages 213–274. PMLR, 2017.

Maria-Florina Balcan, Travis Dick, and Ellen Vitercik. Dispersion for data-driven algorithm design, online learning, and private optimization. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 603–614. IEEE Computer Society, 2018.

Maria-Florina Balcan, Tuomas Sandholm, and Ellen Vitercik. Refined bounds for algorithm configuration: The knife-edge of dual class approximability. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 580–590. PMLR, 2020.

Maria-Florina Balcan, Dan DeBlasio, Travis Dick, Carl Kingsford, Tuomas Sandholm, and Ellen Vitercik. How much data is sufficient to learn high-performing algorithms? Generalization guarantees for data-driven algorithm design. *Journal of the ACM*, 71(5):32:1–32:58, 2024a.

Maria-Florina Balcan, Travis Dick, Tuomas Sandholm, and Ellen Vitercik. Learning to branch: Generalization guarantees and limits of data-independent discretization. *Journal of the ACM*, 71 (2):13:1–13:73, 2024b.

Maria-Florina F. Balcan, Siddharth Prasad, Tuomas Sandholm, and Ellen Vitercik. Structural analysis of branch-and-cut and the learnability of Gomory mixed integer cuts. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Peter Bartlett, Piotr Indyk, and Tal Wagner. Generalization bounds for data-driven numerical linear algebra. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2013–2040. PMLR, 2022.

Omar Bennouna, Amine Bennouna, Saurabh Amin, and Asuman Ozdaglar. What data enables optimal decisions? An exact characterization for linear optimization. In *Advances in Neural Information Processing Systems*, 2025a. NeurIPS 2025. arXiv:2505.21692.

Omar Bennouna, Jiawei Zhang, Saurabh Amin, and Asuman E. Ozdaglar. Contextual optimization under model misspecification: A tractable and generalizable approach. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 3749–3775. PMLR, 2025b.

Omar Bennouna, Amine Bennouna, Saurabh Amin, and Asuman Ozdaglar. Data informativeness in linear optimization under uncertainty. *arXiv preprint arXiv:2602.15365*, 2026. URL https://arxiv.org/abs/2602.15365.

Yonatan Bilu and Nathan Linial. Are stable instances easy? *Combinatorics, Probability and Computing*, 21(5):643–660, 2012.

Avrim Blum and Joel Spencer. Coloring random and semi-random k-colorable graphs. *Journal of Algorithms*, 19(2):204–234, 1995.

Olivier Bousquet, Steve Hanneke, Shay Moran, and Nikita Zhivotovskiy. Proper learning, helly number, and an optimal SVM bound. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of the 33rd Conference on Learning Theory (COLT)*, volume 125 of *Proceedings of Machine Learning Research*, pages 582–609. PMLR, 2020.

Marco C. Campi and Simone Garatti. Compression, generalization and learning. *Journal of Machine Learning Research*, 24(339):1–74, 2023.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

Sanjoy Dasgupta. Two faces of active learning. *Theoretical Computer Science*, 412(19):1767–1781, 2011.

Othman El Balghiti, Adam N. Elmachtoub, Paul Grigas, and Ambuj Tewari. Generalization bounds in the predict-then-optimize framework. *Mathematics of Operations Research*, 48(4):2043–2065, 2023.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68 (1):9–26, 2022.

Sally Floyd and Manfred K. Warmuth. Sample compression, learnability, and the vapnik-chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995.

Robert M. Freund and James B. Orlin. On the complexity of four polyhedral set containment problems. *Mathematical Programming*, 33(2):139–145, 1985.

Aditya Gangrade, Tianrui Chen, and Venkatesh Saligrama. Safe linear bandits over unknown polytopes. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of the 37th Conference on Learning Theory (COLT)*, volume 247 of *Proceedings of Machine Learning Research*, pages 1755–1795. PMLR, 2024.

Thore Graepel, Ralf Herbrich, and John Shawe-Taylor. Pac-bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59(1-2):55–76, 2005.

Peter Gritzmann and Victor Klee. Computational complexity of inner and outer j-radii of polytopes in finite-dimensional normed spaces. *Mathematical Programming*, 59(2):163–213, 1993.

Andrew Guillory and Jeff A. Bilmes. Average-case active learning with costs. In *Algorithmic Learning Theory, 20th International Conference, ALT 2009. Proceedings*, volume 5809 of *Lecture Notes in Computer Science*, pages 141–155. Springer, 2009.

Andrew Guillory and Jeff A. Bilmes. Interactive submodular set cover. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 415–422, 2010.

Rishi Gupta and Tim Roughgarden. A PAC approach to application-specific algorithm selection. *SIAM Journal on Computing*, 46(3):992–1017, 2017.

Rishi Gupta and Tim Roughgarden. Data-driven algorithm design. *Communications of the ACM*, 63(6):87–94, 2020.

Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, 7(2-3):131–309, 2014.

Steve Hanneke and Aryeh Kontorovich. A sharp lower bound for agnostic learning with sample compression schemes. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 489–505. PMLR, 2019.

Steve Hanneke and Aryeh Kontorovich. Stable sample compression schemes: New applications and an optimal SVM margin bound. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 697–721. PMLR, 2021.

Yichun Hu, Nathan Kallus, Xiaojie Mao, and Yanchen Wu. Contextual linear optimization with bandit feedback. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Tomoharu Iwata and Shinsaku Sakaue. Learning to generate projections for reducing dimensionality of heterogeneous linear programming problems. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 26627–26641. PMLR, 2025.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in Modern Analysis and Probability*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society, 1984.

Kai Kellner and Thorsten Theobald. Sum of squares certificates for containment of H-polytopes in V-polytopes. *SIAM Journal on Discrete Mathematics*, 30(2):763–776, 2016.

S. R. Kulkarni, S. K. Mitter, and J. N. Tsitsiklis. Active learning using arbitrary binary valued queries. *Machine Learning*, 11(1):23–35, 1993.

Tung Quoc Le, Anh Tuan Nguyen, and Viet Anh Nguyen. Provably data-driven lagrangian relaxation for mixed integer linear programming. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. Accepted at ICML 2026. arXiv:2605.19052.

Eva Ley and Maximilian Merkert. Solution methods for partial inverse combinatorial optimization problems in which weights can only be increased. *Journal of Global Optimization*, 93(1):263–298, 2025.

Nick Littlestone and Manfred K. Warmuth. Relating data compression and learnability. Technical report, University of California, Santa Cruz, 1986. Technical report.

Heyuan Liu and Paul Grigas. Online contextual decision-making with a smart predict-then-optimize method, 2022.

Mo Liu, Paul Grigas, Heyuan Liu, and Zuo-Jun Max Shen. Active learning for contextual linear optimization: A margin-based approach, 2023. arXiv:2305.06584v2 (Jan 2025).

Zakaria Mhammedi. Online convex optimization with a separation oracle. In Nika Haghtalab and Ankur Moitra, editors, *Proceedings of the 38th Conference on Learning Theory (COLT)*, volume 291 of *Proceedings of Machine Learning Research*, pages 4033–4077. PMLR, 2025.

Shay Moran and Amir Yehudayoff. Sample compression schemes for VC classes. *Journal of the ACM*, 63(3):21:1–21:10, 2016.

Anh Tuan Nguyen and Viet Anh Nguyen. Provably data-driven projection method for quadratic programming. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(29):24541–24548, 2026.

Pierre-Louis Poirion, Bruno F. Lourenço, and Akiko Takeda. Random projections of linear and semidefinite problems with linear inequalities. *Linear Algebra and its Applications*, 664:24–60, 2023.

Tim Roughgarden. Beyond worst-case analysis. *Communications of the ACM*, 62(3):88–96, 2019.

Tim Roughgarden, editor. *Beyond the Worst-Case Analysis of Algorithms*. Cambridge University Press, 2021. ISBN 9781108637435.

Utsav Sadana, Abhilash Chenreddy, Erick Delage, Alexandre Forel, Emma Frejinger, and Thibaut Vidal. A survey of contextual optimization methods for decision-making under uncertainty. *European Journal of Operational Research*, 320(2):271–289, 2025.

Shinsaku Sakaue and Taihei Oki. Generalization bound and learning methods for data-driven projections in linear programming. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Noah Schutte, Grigorii Veviurko, Krzysztof Postek, and Neil Yorke-Smith. Sufficient decision proxies for decision-focused learning, 2025.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. ISBN 9781107057135.

Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM*, 51(3):385–463, 2004.

Yunhao Tang, Shipra Agrawal, and Yuri Faenza. Reinforcement learning for integer programming: Learning to cut. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9367–9376. PMLR, 2020.

Matus Telgarsky. Stochastic linear optimization never overfits with quadratically-bounded losses on general data. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of the 35th Conference on Learning Theory (COLT)*, volume 178 of *Proceedings of Machine Learning Research*, pages 5453–5488. PMLR, 2022.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Ky Khac Vu, Pierre-Louis Poirion, and Leo Liberti. Random projections for linear programming. *Mathematics of Operations Research*, 43(4):1051–1071, 2018.

David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014.

Appendix A. Related literature

Dimension reduction and model compression for LP. Dimension reduction is a classical tool for coping with high-dimensional optimization and learning. Random projections and sketching preserve geometry with high probability (e.g., Johnson–Lindenstrauss (Johnson and Lindenstrauss, 1984); see (Woodruff, 2014)) and are ubiquitous in numerical linear algebra as well as learning-theoretic analyses (Bartlett et al., 2022). For linear programs, random projections can reduce problem size while approximately preserving feasibility and objective values (Vu et al., 2018; Poirion et al., 2023), and Sakaue and Oki (2024) and Iwata and Sakaue (2025) propose data-driven projections with generalization guarantees. In contrast, we seek exact optimizer recovery over a known polytope $\mathcal{X}$ by identifying the directions that can change the optimal solution; we quantify the intrinsic decision dimension by $d^\star$. Our work is most closely related to Bennouna et al. (2025a), which characterizes global sufficient decision datasets for LP under convex open priors and gives an iterative construction with a mixed-integer program at each step. More recently, Nguyen and Nguyen (2026) further extend this data-driven projection viewpoint from LPs to convex quadratic programs, proving generalization guarantees for learned projection matrices.

Data-driven algorithm design. Beyond worst-case analysis argues that worst-case complexity can be overly pessimistic and instead advocates structured or distributional models of "relevant" instances (Roughgarden, 2019, 2021). Representative examples include perturbation-resilient instances (Bilu and Linial, 2012), smoothed analysis (Spielman and Teng, 2004), and planted or semi-random models (Blum and Spencer, 1995). Data-driven algorithm design is a principled ML instantiation of this viewpoint, learning an algorithmic object (e.g., an algorithm family or configuration) from i.i.d. samples while retaining provable performance guarantees (Gupta and Roughgarden, 2020; Balcan, 2021; Gupta and Roughgarden, 2017; Balcan et al., 2017, 2024a,b). Under bounded loss, typical results yield uniform-convergence generalization gaps on the order of $\tilde{O}(\sqrt{\mathrm{Pdim}(\mathcal{A})/n})$, and sharper bounds can be obtained via refined complexity notions such as dispersion and "knife-edge" structure (Balcan et al., 2018, 2020). Beyond LP and QP dimension reduction, this viewpoint has also been applied to optimization problems such as integer programming, including learning cutting-plane policies for branch-and-cut (Tang et al., 2020; Balcan et al., 2022) and learning Lagrangian relaxations for MILP (Le et al., 2026). Our setting fits this paradigm by treating the queried directions as the learned object. For our specific problem, we exploit LP geometry to derive a stable compression scheme, yielding fast-rate certificates scaling as $\widetilde{O}(d^\star/n)$. At a technical level, our geometry-aware cutting-plane viewpoint is also reminiscent of oracle-based frameworks that access constraints through separation (Mhammedi, 2025).

Compression-based generalization. Compression-based analyses provide distribution-free generalization and scenario-type certificates, often yielding fast k/n-type rates when the learned object admits a small compression of size k (Valiant, 1984; Littlestone and Warmuth, 1986; Floyd and Warmuth, 1995; Moran and Yehudayoff, 2016; Graepel et al., 2005). In the realizable regime, stable compression can further sharpen logarithmic factors (Bousquet et al., 2020; Hanneke and Kontorovich, 2021). In fully agnostic settings, however, $1/n$ rates are impossible in general: sharp lower bounds show worst-case rates of order $\Theta\big(\sqrt{k\log(n/k)/n}\big)$ (Hanneke and Kontorovich, 2019). Our cumulative algorithm in Section 5 fits naturally into this view: each newly queried direction is triggered by a "hard" sample revealing a genuinely new decision-relevant facet, producing a compression set of size at most $d^\star$ and a fast-rate certificate.

Polyhedral containment. Our hardness results rely on classical complexity phenomena in polyhedral computation. In particular, deciding containment of an H-polytope in a V-polytope is coNP-complete (Freund and Orlin, 1985; Gritzmann and Klee, 1993). Sum-of-squares certificates provide powerful relaxations for containment questions and yield refined results for structured instances (Kellner and Theobald, 2016). Since pointwise sufficiency can be phrased as a containment statement that a data-consistent slice lies within an optimality region, these results naturally connect to the verification problem studied in our paper.

Contextual optimization. Our paper is also motivated by decision-focused (predict-then-optimize) learning, where one learns predictive models to support downstream optimization (Elmachtoub and Grigas, 2022; El Balghiti et al., 2023); see also the survey (Sadana et al., 2025). Beyond the batch setting, recent work studies online and active learning variants of contextual linear optimization, including margin-based active learning (Liu et al., 2023) and online contextual decision-making with SPO-type surrogates (Liu and Grigas, 2022); for background on active learning, see Dasgupta (2011); Hanneke (2014). Bandit/partial-feedback formulations are studied in Hu et al. (2024), and related safe-exploration objectives appear in safe linear bandits over unknown polytopes (Gangrade et al., 2024). On the statistical side, El Balghiti et al. (2023) derive uniform generalization bounds for the SPO loss via the Natarajan dimension of the induced decision class; in the polyhedral case, their bounds depend only logarithmically on the number of extreme points, yielding rates on the order of $\tilde{O}\left(\sqrt{pd\,\log(n|\mathcal{X}^{\angle}|)/n}\right)$ for linear predictors. Recent work addresses model misspecification in contextual optimization (Bennouna et al., 2025b) and benign generalization behavior in stochastic linear optimization under quadratically bounded losses (Telgarsky, 2022); see also Schutte et al. (2025) on sufficient proxy representations.

Active learning and adaptive measurement. Active learning studies how to adaptively acquire information—labels, features, or more general tests—to reduce query cost relative to passive sampling. A central idea is to query only where candidate hypotheses disagree, as formalized by disagreement-based active learning (Hanneke, 2014); see also Dasgupta (2011). Classical work distinguishes sample complexity from label (query) complexity in agnostic active learning (Balcan et al., 2006, 2009). Our per-instance model is closer in spirit to arbitrary-query and experimental-design views of active learning (Kulkarni et al., 1993; Cohn et al., 1996), since we design linear measurements $q^\top c$ of a possibly unknown c. Frameworks accounting for heterogeneous query costs are also related to our measurement budget (Guillory and Bilmes, 2009, 2010). At the intersection with decision-focused learning, Liu et al. (2023) studies margin-based active learning for contextual linear optimization. In contrast, we leverage polyhedral LP geometry: our facet-hit rule queries a violated optimality-cone facet normal, guaranteeing exact decision identification after at most $d^\star$ measurements and enabling a stable-compression view across i.i.d. instances.

Appendix B. Formal statement for hardness and other proofs for Section 3

Roadmap and proof sketches. This appendix collects proofs for the complexity results in Section 3. We provide formal statements of the hardness results, and then give self-contained reductions.

- **NP-hardness.** Theorem 5 and its formal counterpart Theorem 26 show that computing the decision-relevant dimension $d^\star := \dim \mathrm{dir}(X^\star(\mathcal{C}))$ is NP-hard, even when X is a shortest-path flow polytope and even for both the closed and open budget sets $\mathcal{C}_{\mathrm{cl}}$ and $\mathcal{C}_{\mathrm{op}}$. We start

from 3-SAT and use the PISPP-W+ construction of (Ley and Merkert, 2025, Theorem 3.1). Given a formula φ, the reduction outputs a directed acyclic graph (DAG) together with baseline arc-lengths d, a budget vector κ, a budget B, and a required arc r, defining a budgeted uncertainty set of admissible length vectors $c = d + w$ (with $w \in \mathbb{Q}^A_{\geq 0}$ and $\kappa^\top w \leq B$). The formula φ is satisfiable if and only if there exists such a modification w for which some shortest s–t path with respect to $d + w$ contains the required arc r.

We then show that deciding the feasibility of the resulting PISPP-W+ instance reduces to checking whether $\dim \operatorname{dir}(X^\star(\mathcal{C})) > \dim \operatorname{dir}(X^\star_r(\mathcal{C}))$, where $X_r := \{x \in X : x_r = 0\}$; see (11). Geometrically, restricting to the face $x_r = 0$ removes a decision-relevant extreme direction if and only if there exists a feasible modification that makes some shortest s–t path use r. For the open set $\mathcal{C}_{\text{op}}$, Lemma 28 "opens" the budget via a small topological perturbation without changing the shortest-path structure and thus maintaining the answer. Under an open convex $\mathcal{C}$, Corollary 4 then translates this into NP-hardness of computing the minimum global-SDD size.

- coNP**-hardness.** Theorem 9 reduces the *H-in-V polytope containment problem* to checking whether the empty dataset is pointwise sufficient at a fixed cost, yielding coNP-hardness of the verification problem. Concretely, starting from the restricted hard family $P = [-1, 1]^d \subseteq Q$ with $0 \in \operatorname{int}(Q)$ (Gritzmann and Klee, 1993), we construct a bounded *standard-form* polytope $X = \{z : Az = b,\ z \geq 0\}$ with a distinguished vertex x_0 such that the relevant slice of its optimality cone satisfies

$$((y, 1), 0, 0) \in \Lambda(x_0) \quad \Longleftrightarrow \quad (y, 1) \in \operatorname{cone}\{(v_i, 1)\}_{i=1}^M \quad \Longleftrightarrow \quad y \in Q.$$

We then set $\mathcal{C} = \{((y, 1), 0, 0) : y \in P\}$ and take $\mathcal{D} = \emptyset$, so the data-consistent fiber equals all of $\mathcal{C}$. Since x_0 is the unique minimizer for the reference cost $c_0 = ((0, 1), 0, 0)$ (using $0 \in \operatorname{int}(Q)$), pointwise sufficiency reduces to the cone containment $\mathcal{C} \subseteq \Lambda(x_0)$, which is equivalent to $P \subseteq Q$.

Theorem 10 strengthens this to zero-query global sufficiency by replacing $\mathcal{C}$ with an open, full-dimensional "thickening" $\mathcal{C}_{\text{op}}$ defined via a linear *effective-cost* map T (so that optimizing over X depends on c only through $T(c)$): we choose an open polyhedron B_{op} of effective costs with strictly positive last coordinate and set $\mathcal{C}_{\text{op}} := \{c : T(c) \in B_{\text{op}}\}$. In the YES case $P \subseteq Q$, we have $B_{\text{op}} \subseteq \operatorname{cone}\{(v_i, 1)\}_{i=1}^M$, and since B_{op} is open this implies $B_{\text{op}} \subseteq \operatorname{int} \operatorname{cone}\{(v_i, 1)\}_{i=1}^M$; consequently x_0 is the unique optimizer for every $c \in \mathcal{C}_{\text{op}}$ and the optimal solution set is constant, so a decoder with $|D| = 0$ exists. In the NO case, B_{op} contains an effective cost outside $\operatorname{cone}\{(v_i, 1)\}_{i=1}^M$, yielding two costs in $\mathcal{C}_{\text{op}}$ with different optimizers and ruling out any decoder with $|D| = 0$.

B.1. Proof of Theorem 5

Theorem 26 (Formal version of Theorem 5) *Fix a coordinate index $r \in [n]$. The following decision problem is* NP*-hard: given a bounded polytope $X \subseteq \mathbb{R}^n$ and a polyhedral uncertainty set $\mathcal{C} \subseteq \mathbb{R}^n$ specified in H-representation, decide whether*

$$\dim \operatorname{dir}(X^\star(\mathcal{C})) \ > \ \dim \operatorname{dir}(X^\star_r(\mathcal{C})), \tag{11}$$

where $X_r := \{x \in X : \ x_r = 0\}$ *and* $X_r^\star(\mathcal{C})$ *is defined as in Equation* (3) *with* X *replaced by* X_r. *Consequently, computing* $\dim \operatorname{dir}(X^\star(\mathcal{C}))$ *is* NP*-hard under polynomial-time Turing reductions.*

The hardness persists even when X *is the* s*–*t *unit-flow polytope of a directed acyclic graph and* $\mathcal{C}$ *is a budgeted set of arc-length increases of the form*

$$\mathcal{C}_{\mathrm{cl}} = \{d + w : \ w \geq 0, \ \kappa^\top w \leq B\} \qquad \textit{or} \qquad \mathcal{C}_{\mathrm{op}} = \{d + w : \ w > 0, \ \kappa^\top w < B + \eta\},$$

for any fixed rational $\eta \in (0, 1)$.

Proof We prove NP-hardness via the 3-SAT⇔PISPP-W+ construction of Ley and Merkert (2025, Theorem 3.1). Given a 3-SAT instance φ, their reduction produces a partial inverse shortest path instance with only weight increases (PISPP-W+), consisting of a directed acyclic graph $G = (V, A)$ with source s and sink t, initial arc-lengths $d \in \mathbb{Q}^A_{\geq 0}$, modification costs $\kappa \in \mathbb{Q}^A_{>0}$, a budget $B \in \mathbb{Q}_{>0}$, and a single required arc $r \in A$. The equivalent PISPP-W+ decision question is whether there exists a modification vector $w \in \mathbb{Q}^A_{\geq 0}$ with $\kappa^\top w \leq B$ such that some shortest s–t path with respect to the modified arc-lengths $d + w$ contains r. We proceed with the proof in three steps.

(1) **Restating the hard instance and key properties:** recall the explicit 3-SAT→PISPP-W+ construction and the structural properties we will use.

(2) **Openifying the uncertainty set:** show that passing from the closed budgeted set $\mathcal{C}_{\mathrm{cl}}$ to the slightly relaxed open set $\mathcal{C}_{\mathrm{op}}$ does not change the answer to the PISPP-W+ decision question on these instances.

(3) **Dimension comparison:** encode s–t paths as extreme points of a flow polytope and reduce the PISPP-W+ decision question to the strict inequality (11).

Step 1: Hard instance and structural properties. We first restate the explicit 3-SAT→PISPP-W+ construction in the notation needed here. Let φ have variables $x_1, \ldots, x_n$ and clauses $B_1, \ldots, B_m$. Write each clause as $B_j = \{\ell_{j1}, \ell_{j2}, \ell_{j3}\}$ where each literal $\ell_{jk} \in \{x_i, \bar{x}_i : \ i \in [n]\}$ (if a clause has fewer than three literals, duplicate one). For each ℓ_{jk} we create a *clause–literal vertex* b_{jk} labeled by ℓ_{jk}.

Vertices. Let V consist of

$$\{s_0, \ldots, s_n\} \ \cup \ \{t_0, \ldots, t_m\} \ \cup \ \{x_i, \bar{x}_i : \ i \in [n]\} \ \cup \ \{b_{jk} : \ j \in [m], \ k \in [3]\}.$$

Set the source $s := s_0$ and the sink $t := t_m$.

Arcs. We build a DAG $G = (V, A)$ with a variable layer (from s_0 to s_n) and a clause layer (from t_0 to t_m), connected by one *required* arc and additional *shortcut* arcs.

- *Variable gadgets:* for each $i \in [n]$, add the four arcs

$$(s_{i-1}, x_i), \ (x_i, s_i), \ (s_{i-1}, \bar{x}_i), \ (\bar{x}_i, s_i).$$

  Thus, any s_0–s_n path chooses exactly one of $\{x_i, \bar{x}_i\}$ for each i.

- *Clause gadgets:* for each $j \in [m]$ and $k \in [3]$, add

$$(t_{j-1}, b_{jk}),\ (b_{jk}, t_j).$$

  Thus, any t_0–t_m path chooses exactly one literal vertex b_{jk} per clause j.

- *Required arc:* add $r := (s_n, t_0)$, which is the unique required arc ($R = \{r\}$).

- *Shortcut arcs:* for each clause–literal vertex b_{jk} labeled by a literal on variable x_i, add exactly one arc from the *opposite* variable vertex into b_{jk}:

$$\ell_{jk} = x_i \ \Rightarrow\ (\bar{x}_i, b_{jk}) \in A, \qquad \ell_{jk} = \bar{x}_i \ \Rightarrow\ (x_i, b_{jk}) \in A.$$

  Equivalently, the tail of the shortcut arc is the variable vertex that makes ℓ_{jk} *false*.

Initial arc-lengths and modification costs. Set the initial lengths d by

$$d(a) = 1 \text{ for every non-shortcut arc } a, \qquad d(a) = 2(n-i+j) \text{ for a shortcut arc } a = (\cdot, b_{jk}) \text{ with } \ell_{jk} \in \{x_i, \bar{x}_i\}.$$

Let the budget be $B := n + 2m$ and set modification costs κ by

$$\kappa_a = 1 \text{ for every non-shortcut arc } a, \qquad \kappa_a = B + 1 \text{ for every shortcut arc } a.$$

Lemma 27 (Structural properties) *The above instance satisfies the following properties (cf. Observations 1–4 and the claim in the proof of (Ley and Merkert, 2025, Theorem 3.1)):*

(i) ***Degree constraints.*** *Each x_i and $\bar{x}_i$ has exactly one ingoing arc (from s_{i-1}), and each b_{jk} has exactly one outgoing arc (to t_j).*

(ii) ***Path structure ("r or one shortcut").*** *Every s–t path contains either the required arc r or exactly one shortcut arc (but not both).*

(iii) ***Unit gap under d.*** *Every s–t path using r has length $2n + 2m + 1$ under d, while every s–t path using a shortcut arc has length $2n + 2m$ under d. In particular (with $w = 0$), all shortest s–t paths avoid r.*

(iv) ***Encoding of assignments + "false literal $\Rightarrow$ available shortcut".*** *If an s–t path P uses r, then it visits exactly one of $\{x_i, \bar{x}_i\}$ for every $i \in [n]$ and exactly one b_{jk} in every clause gadget $j \in [m]$. Define the induced truth assignment by $x_i^P = 1$ iff $x_i \in P$. If P visits a literal vertex b_{jk} that is false under x^P, then the unique shortcut arc entering b_{jk} has its tail on P.*

(v) ***3-SAT $\Leftrightarrow$ PISPP-W+.*** *The mapping $\varphi \mapsto (G, d, \kappa, B, r)$ is computable in polynomial time and satisfies the following equivalence: the formula φ is satisfiable if and only if there exists a modification vector $w \in \mathbb{Q}^A_{\geq 0}$ with $\kappa^\top w \leq B$ such that some shortest s–t path with respect to the modified arc-lengths $d + w$ contains the required arc r. Equivalently, if φ is unsatisfiable then for every $w \in \mathbb{Q}^A_{\geq 0}$ with $\kappa^\top w \leq B$, every shortest s–t path under $d + w$ avoids r.*

Proof (i) is immediate from the arc construction. (ii) The only arcs that can enter the clause layer are $r = (s_n, t_0)$ and the shortcut arcs into some b_{jk}. Once the path enters the clause layer, it cannot return to the variable layer (there are no such arcs), and each b_{jk} has only the outgoing arc (b_{jk}, t_j); hence no s–t path can use two shortcuts, and no path can use both r and a shortcut. (iii) Any s–t path using r traverses $2n$ unit-length arcs in the variable gadgets, then r (length 1), then $2m$ unit-length arcs in the clause gadgets, totaling $2n + 2m + 1$. A path using a shortcut from variable i into clause j has prefix length $2(i-1)+1$ up to the tail variable vertex, then shortcut length $2(n-i+j)$, then suffix length $1 + 2(m-j)$ in the clause layer, totaling

$$(2(i-1)+1) + 2(n-i+j) + (1+2(m-j)) = 2n + 2m.$$

(iv) The first statement follows since each variable gadget offers exactly two disjoint choices and each clause gadget offers exactly three disjoint choices. For the last statement, if b_{jk} is false under x^P, then P must have visited the opposite variable vertex ($\bar{x}_i$ if $\ell_{jk} = x_i$, and x_i if $\ell_{jk} = \bar{x}_i$), which is exactly the tail of the unique shortcut arc into b_{jk}.

(v) is exactly the claim of Theorem 3.1 in Ley and Merkert (2025), which shows NP-completeness of PISPP-W+. We provide a proof sketch here:

($\Rightarrow$) Let x^* satisfy φ. W.l.o.g. reorder literals in each clause so that b_{j1} is true. Build the s–t path P that uses r by following x^* in the variable gadgets and b_{j1} in each clause gadget. Set $w = 1$ on the unique entry arc into the unchosen variable vertex (one per variable) and set $w = 1$ on (b_{j2}, t_j) and (b_{j3}, t_j) (two per clause); set $w = 0$ on all remaining arcs (in particular on shortcut arcs). Then $\kappa^\top w = n + 2m = B$, and any shortcut path must traverse at least one penalized arc, so its length increases by ≥ 1, closing the initial gap $d(Q) = d(P) - 1$ and implying P is (one of) the shortest paths.

($\Leftarrow$) Suppose $\kappa^\top w \leq B$ and some shortest path P under $d + w$ uses r; let x^P be the assignment induced by P. If some clause is false under x^P, then the visited literal vertex b_{jk} is false, and by (iv) the entering shortcut arc e has its tail on P. Replacing the corresponding subpath R of P by e gives a shortcut path Q with $d(Q) = d(P) - 1$ (by (iii)). Thus $(d+w)(Q) - (d+w)(P) = -1 + w(e) - w(R) \leq -1 + w(e)$, so shortestness of P forces $w(e) \geq 1$. Since e is a shortcut arc with $\kappa_e = B + 1$, this implies $\kappa^\top w \geq \kappa_e w(e) > B$, a contradiction. ■

Step 2: Openifying the uncertainty set does not change the answer. This step is only needed to ensure hardness persists for an open uncertainty set. We consider two uncertainty sets of admissible arc-length vectors:

$$\mathcal{C}_{\text{cl}} := \{\, c = d+w :\ w \in \mathbb{R}^A_{\geq 0},\ \kappa^\top w \leq B \,\} \quad \text{and} \quad \mathcal{C}_{\text{op}} := \{\, c = d+w :\ w \in \mathbb{R}^A_{>0},\ \kappa^\top w < B+\eta \,\},$$

where we fix any rational constant $\eta \in (0,1)$ (e.g. $\eta = \frac{1}{2}$). The lemma below shows that for the hard instances above, replacing $\mathcal{C}_{\text{cl}}$ by $\mathcal{C}_{\text{op}}$ does not change the answer to the underlying question.

Lemma 28 *Fix any $\eta \in (0,1)$. For the PISPP-W+ instance produced by Ley and Merkert (2025) from φ, the formula φ is satisfiable if and only if there exists $c \in \mathcal{C}_{\text{op}}$ such that some shortest s–t path with respect to c contains r.*

Proof ($\Rightarrow$) Suppose φ is satisfiable. By Lemma 27(v), there exists $w_0 \in \mathbb{Q}^A_{\geq 0}$ with $\kappa^\top w_0 \leq B$ such that for $c_0 := d + w_0$ there is a shortest s–t path using r.

Since G is a DAG, fix any topological ordering $\rho : V \to \{0, 1, \ldots, |V| - 1\}$. For (a rational) $\varepsilon > 0$, define a perturbation $\delta \in \mathbb{Q}^A$ by

$$\delta(u, v) := \varepsilon(\rho(v) - \rho(u)) \qquad \forall (u, v) \in A.$$

Then $\delta > 0$ componentwise. Moreover, for any s–t path $P = (v_0 = s, v_1, \ldots, v_k = t)$, the perturbation telescopes:

$$\sum_{i=0}^{k-1} \delta(v_i, v_{i+1}) = \varepsilon \sum_{i=0}^{k-1} (\rho(v_{i+1}) - \rho(v_i)) = \varepsilon(\rho(t) - \rho(s)),$$

which is a constant independent of P. Hence, adding δ shifts the length of every s–t path by the same constant, so the set of shortest s–t paths is unchanged when we replace c_0 by $c_0 + \delta$.

Finally, choose ε small enough so that $\kappa^\top \delta < \eta$. For instance, since $\rho(v) - \rho(u) \le |V| - 1$ for every arc,

$$\kappa^\top \delta = \sum_{a \in A} \kappa_a \delta_a \le \varepsilon(|V| - 1) \sum_{a \in A} \kappa_a,$$

so it suffices to take $\varepsilon := \eta / \big(2(|V| - 1) \sum_{a \in A} \kappa_a\big)$. With this choice, $w := w_0 + \delta$ satisfies $w > 0$ and $\kappa^\top w < B + \eta$, i.e. $c := d + w \in \mathcal{C}_{\text{op}}$, and there still exists a shortest path using r.

($\Leftarrow$) We prove the contrapositive. Suppose φ is not satisfiable, assume that there exists a shortest s–t path P with respect to $c := d + w \in \mathcal{C}_{\text{op}}$ that contains r.

By Lemma 27(ii), P contains no shortcut arc. Define the induced assignment x^P as in Lemma 27(iv). Since φ is unsatisfiable, there exists a clause index j that is not satisfied by x^P. By Lemma 27(iv), P visits exactly one literal vertex b_{jk} in clause gadget j; because clause j is unsatisfied, this visited literal vertex b_{jk} is false under x^P. Therefore Lemma 27(iv) yields that the unique shortcut arc e entering b_{jk} has its tail on P.

Let R denote the (nonempty) subpath of P from the tail of e to the vertex b_{jk}, and define a new s–t path Q by following P up to the tail of e, then traversing e, and then following P from b_{jk} to t. Then Q uses a shortcut arc, so by Lemma 27(iii) we have $d(Q) = d(P) - 1$.

Since P and Q coincide outside of R and e, we obtain

$$(d + w)(Q) - (d + w)(P) = (d(Q) - d(P)) + (w(Q) - w(P)) = -1 + w(e) - w(R).$$

Because R contains at least one arc and $w > 0$ componentwise, we have $w(R) > 0$. Thus if P is shortest we must have $0 \le (d + w)(Q) - (d + w)(P)$, implying $w(e) > 1$.

As e is a shortcut arc, $\kappa_e = B + 1$, hence

$$\kappa^\top w \ge \kappa_e w(e) > (B + 1) \cdot 1 = B + 1.$$

In particular, for any $\eta \in (0, 1)$ we have $\kappa^\top w > B + \eta$, contradicting the assumption $\kappa^\top w < B + \eta$ required for $d + w \in \mathcal{C}_{\text{op}}$. Therefore, no such $w > 0$ can make a shortest s–t path use r. ■

Step 3: Reduction from PISPP-W+ to a comparison of decision-relevant dimensions. We now translate the PISPP-W+ instance into a linear optimization problem over the s–t unit-flow polytope. The coordinate indexed by the required arc r will play the role of the distinguished coordinate in (11), and we will compare (the dimensions of) reachable optimal directions with and without imposing $x_r = 0$.

Given the instance (G, d, κ, B, r), let $p := |A|$ and index coordinates of $\mathbb{R}^p$ by arcs. Define the s–t unit-flow polytope

$$X := \Big\{ x \in \mathbb{R}^p_{\geq 0} : \sum_{a \in \delta^+(v)} x_a - \sum_{a \in \delta^-(v)} x_a = \begin{cases} 1 & v = s, \\ -1 & v = t, \\ 0 & \text{otherwise,} \end{cases} \Big\}.$$

Because G is acyclic, the extreme points of X are exactly the incidence vectors of s–t paths. Let $X_r := \{x \in X : \ x_r = 0\}$ be the face of flows that avoid the required arc. Note that X_r is a face of X, hence

$$X_r^{\angle} = X^{\angle} \cap X_r = \{x \in X^{\angle} : \ x_r = 0\}. \tag{12}$$

For $\mathcal{C} \in \{\mathcal{C}_{\text{cl}}, \mathcal{C}_{\text{op}}\}$, define the reachable optimal sets $X^\star(\mathcal{C})$ and $X_r^\star(\mathcal{C})$ as in Equation (3).

The next lemma is the key structural property for the dimension comparison: it identifies the reachable optimal extreme points of the restricted face X_r and shows that all of them are also reachable for X.

Lemma 29 *For $\mathcal{C} \in \{\mathcal{C}_{\text{cl}}, \mathcal{C}_{\text{op}}\}$ constructed above, we have*

$$X_r^\star(\mathcal{C}) = X_r^{\angle} \qquad \textit{and} \qquad X_r^{\angle} \subseteq X^\star(\mathcal{C}).$$

Proof For $\mathcal{C} = \mathcal{C}_{\text{cl}}$, let $c^0 := d \in \mathcal{C}$ (take $w = 0$). For $\mathcal{C} = \mathcal{C}_{\text{op}}$, let δ be the topological perturbation from the proof of Lemma 28 and set $c^0 := d + \delta \in \mathcal{C}_{\text{op}}$. In either case, by the unit-gap property under d and by telescoping of δ, every s–t path that avoids r is shortest under c^0, and every path that uses r is strictly longer. Combine with (12), we have

$$X^\star(c^0) = \{x \in X^{\angle} : x_r = 0\} = X_r^{\angle},$$

and in particular $X_r^{\angle} \subseteq X^\star(\mathcal{C})$.

Moreover, for this same c^0, every extreme point of X_r (i.e., every s–t path avoiding r) is optimal for $\min\{(c^0)^\top x : x \in X_r\}$, so $X_r^{\angle} \subseteq X_r^\star(\mathcal{C})$. The reverse inclusion $X_r^\star(\mathcal{C}) \subseteq X_r^{\angle}$ holds by definition, hence $X_r^\star(\mathcal{C}) = X_r^{\angle}$. ■

We are now ready to complete the reduction. To prove Theorem 26, it suffices to establish the following claim.

Claim 30 *The formula φ is satisfiable if and only if* $\dim \operatorname{dir}(X^\star(\mathcal{C})) > \dim \operatorname{dir}(X_r^\star(\mathcal{C}))$.

Proof By Lemma 29, we always have $X_r^\star(\mathcal{C}) \subseteq X^\star(\mathcal{C})$, hence

$$\operatorname{dir}(X_r^\star(\mathcal{C})) \subseteq \operatorname{dir}(X^\star(\mathcal{C})). \tag{13}$$

($\Rightarrow$) Assume φ is satisfiable. We claim that there exists $c^+ \in \mathcal{C}$ such that some shortest s–t path w.r.t. c^+ contains r. If $\mathcal{C} = \mathcal{C}_{\text{cl}}$, this follows from Lemma 27(v) (equivalently, Ley and Merkert (2025, Theorem 3.1)); if $\mathcal{C} = \mathcal{C}_{\text{op}}$, it follows from Lemma 28.

Let $x^+ \in X^\star(c^+) \cap X^{\angle} \subseteq X^\star(\mathcal{C})$ be the incidence vector of such a shortest path, so $x_r^+ = 1$. By Lemma 29, we also have $X_r^{\angle} \subseteq X^\star(\mathcal{C})$; pick any $x^0 \in X_r^{\angle}$, so $x_r^0 = 0$. Then $v := x^+ - x^0 \in \operatorname{dir}\big(X^\star(\mathcal{C})\big)$ and satisfies $v_r = 1$.

On the other hand, every $u \in \operatorname{dir}\big(X_r^\star(\mathcal{C})\big)$ satisfies $u_r = 0$, and therefore $v \notin \operatorname{dir}\big(X_r^\star(\mathcal{C})\big)$. Combining this with (13) yields

$$\operatorname{dir}\big(X_r^\star(\mathcal{C})\big) \subsetneq \operatorname{dir}\big(X^\star(\mathcal{C})\big),$$

and thus $\dim \operatorname{dir}(X^\star(\mathcal{C})) > \dim \operatorname{dir}(X_r^\star(\mathcal{C}))$.

($\Leftarrow$) We prove the contrapositive. Assume φ is not satisfiable. Then, for $\mathcal{C} = \mathcal{C}_{\text{cl}}$ the corresponding PISPP-W+ instance is infeasible by Ley and Merkert (2025), and for $\mathcal{C} = \mathcal{C}_{\text{op}}$ it is infeasible by Lemma 28. Equivalently, for every $c \in \mathcal{C}$, every extreme optimal solution of $\min\{c^\top x : x \in X\}$ avoids the arc r, i.e., satisfies $x_r = 0$. Since X_r is a face of X, we have $X_r^\lessdot = \{x \in X^\lessdot : x_r = 0\}$, and therefore

$$X^\star(c) \cap X^\lessdot \subseteq X_r^\lessdot \qquad \forall\, c \in \mathcal{C}.$$

Taking the union over $c \in \mathcal{C}$ gives $X^\star(\mathcal{C}) \subseteq X_r^\lessdot$. On the other hand, Lemma 29 yields $X_r^\lessdot \subseteq X^\star(\mathcal{C})$ and $X_r^\star(\mathcal{C}) = X_r^\lessdot$. Hence $X^\star(\mathcal{C}) = X_r^\star(\mathcal{C})$, and thus $\dim \operatorname{dir}(X^\star(\mathcal{C})) = \dim \operatorname{dir}(X_r^\star(\mathcal{C}))$. ■

This completes the reduction. Therefore deciding whether (11) holds is NP-hard.

Finally, to deduce the hardness of computing $\dim \operatorname{dir}(X^\star(\mathcal{C}))$ as a function problem, note that if we could compute $\dim \operatorname{dir}(X^\star(\mathcal{C}))$ in polynomial time, we could compute both sides of (11) (one call on $(X, \mathcal{C})$ and one call on $(X_r, \mathcal{C})$) and decide the inequality, implying NP-hardness under polynomial-time Turing reductions. ■

B.2. Proof of Theorem 9

Theorem 31 (Formal version of Theorem 9) *The following decision problem is* coNP*-hard: given a bounded polytope $X \subseteq \mathbb{R}^n$, a polyhedral uncertainty set $\mathcal{C} \subseteq \mathbb{R}^n$ specified in H-representation, a dataset $\mathcal{D}$, and a cost vector $c \in \mathcal{C}$, decide whether $\mathcal{D}$ is pointwise sufficient at c in the sense of Definition 6. Consequently, computing the size of a minimum pointwise SDD (and the corresponding search problem of finding one) is* coNP*-hard.*

The H-in-V polytope containment problem. An instance consists of two polytopes $P, Q \subseteq \mathbb{R}^d$ where

$$P = \{z \in \mathbb{R}^d : \ Hz \le h\} \qquad \text{and} \qquad Q = \operatorname{conv}\{v_1, \dots, v_M\},$$

i.e., P is given in *H-representation* and Q is given in *V-representation*. The decision problem asks whether $P \subseteq Q$. This problem is coNP-complete (Freund and Orlin, 1985). Moreover, the coNP-hardness persists under strong structural restrictions; in particular, it is already coNP-complete to decide whether the standard cube is contained in an affine image of a cross polytope (Gritzmann and Klee, 1993). A convenient modern reference is Kellner and Theobald (2016, Proposition 2.1). In our reduction, we work with the following convenient hard family: given $Q = \operatorname{conv}\{v_1, \dots, v_M\} \subseteq \mathbb{R}^d$ that is full-dimensional and satisfies $0 \in \operatorname{int}(Q)$, decide whether the hypercube $P = [-1, 1]^d$ is contained in Q.

Construction of a standard-form LP instance. Let $P = [-1,1]^d$ and $Q = \mathrm{conv}\{v_1, \dots, v_M\}$ be such a hard instance. Set $n_0 := d+1$ and define the homogenized vectors $\bar{v}_i := (v_i, 1) \in \mathbb{R}^{n_0}$. Let $\bar{V} \in \mathbb{R}^{M \times n_0}$ be the matrix whose ith row is $\bar{v}_i^\top$, and set $\beta := \bar{V}\mathbf{1} \in \mathbb{R}^M$, where $\mathbf{1}$ denotes the all-ones vector. We introduce variables $w, r \in \mathbb{R}^{n_0}$ and $s \in \mathbb{R}^M$, and write $z := (w, r, s) \in \mathbb{R}^n$ with $n := 2n_0 + M$. Define the standard-form polytope

$$X := \Big\{ z \in \mathbb{R}^n : \underbrace{\begin{bmatrix} I_{n_0} & I_{n_0} & 0 \\ \bar{V} & 0 & -I_M \end{bmatrix}}_{=:A} z = \underbrace{\begin{bmatrix} 2\cdot\mathbf{1} \\ \beta \end{bmatrix}}_{=:b}, \ z \geq 0 \Big\}. \tag{14}$$

The matrix A has full row rank, and X is nonempty and bounded (indeed, $w + r = 2 \cdot \mathbf{1}$ implies $0 \leq w \leq 2 \cdot \mathbf{1}$ and $0 \leq r \leq 2 \cdot \mathbf{1}$, while $s = \bar{V}w - \beta$ is then bounded as well). Let

$$z_0 := (\mathbf{1}, \mathbf{1}, 0) \in X.$$

We consider the empty dataset $\mathcal{D} = \varnothing$ and the cost vector

$$c_0 := (e_{n_0}, 0, 0) \in \mathbb{R}^n,$$

where $e_{n_0} \in \mathbb{R}^{n_0}$ is the last standard basis vector. Finally, define the polyhedral uncertainty set

$$\mathcal{C} := \big\{ ((y,1), 0, 0) \in \mathbb{R}^n : \ y \in P \big\}.$$

Note that $c_0 \in \mathcal{C}$ since $0 \in [-1,1]^d$.

We claim that $\mathcal{D} = \varnothing$ is pointwise sufficient at c_0 if and only if $P \subseteq Q$.

Step 1: the optimality cone at z_0. For a standard-form polytope $X = \{z : Az = b, \ z \geq 0\}$ and an extreme point z_0, the optimality cone can be written via KKT as

$$\Lambda(z_0) = \Big\{ c \in \mathbb{R}^n : \ \exists\, y, \rho \text{ s.t. } c = A^\top y + \rho, \ \rho \geq 0, \ \rho_j = 0 \text{ whenever } (z_0)_j > 0 \Big\}. \tag{15}$$

Here $z_0 = (\mathbf{1}, \mathbf{1}, 0)$ has strictly positive components on the w- and r-coordinates and zeros on the s-coordinates. Writing $y = (\alpha, \mu) \in \mathbb{R}^{n_0} \times \mathbb{R}^M$, we have

$$A^\top y = \big(\alpha + \bar{V}^\top \mu, \ \alpha, \ -\mu\big).$$

Therefore, for costs of the form $((\tilde{y}, \tilde{t}), 0, 0)$ we obtain the characterization

$$((\tilde{y}, \tilde{t}), 0, 0) \in \Lambda(z_0) \quad \Longleftrightarrow \quad (\tilde{y}, \tilde{t}) \in \mathrm{cone}\{\bar{v}_1, \dots, \bar{v}_M\}. \tag{16}$$

Step 2: z_0 is the unique minimizer for c_0. Let $u := (x,t) := w - \mathbf{1} \in \mathbb{R}^d \times \mathbb{R}$. From (14), feasibility implies $w \in [0,2]^{n_0}$ and

$$s_i = \bar{v}_i^\top w - \beta_i = \bar{v}_i^\top (w - \mathbf{1}) = v_i^\top x + t \ \geq 0 \qquad \forall i \in [M].$$

Thus the projection $z \mapsto u = w - \mathbf{1}$ identifies X with the bounded polytope

$$\widetilde{X} := \{(x,t) \in \mathbb{R}^d \times \mathbb{R} : \ v_i^\top x + t \geq 0 \ \forall i \in [M], \ -1 \leq (x,t) \leq 1\}.$$

Moreover, minimizing $c_0^\top z$ over X is equivalent (up to an additive constant) to minimizing t over $\widetilde{X}$, since $c_0^\top z = e_{n_0}^\top w = t + 1$.

We now show that $(0, 0)$ is the unique minimizer of $\min\{t : (x, t) \in \widetilde{X}\}$. First, we claim that every feasible point satisfies $t \geq 0$. Indeed, if $t < 0$ then $v_i^\top x \geq -t > 0$ for all i, hence $z^\top x > 0$ for all $z \in Q = \mathrm{conv}\{v_1, \ldots, v_M\}$. But $0 \in \mathrm{int}(Q)$ implies that for any $x \neq 0$ there exists $\varepsilon > 0$ such that $-\varepsilon x/\|x\| \in Q$, which yields $(-\varepsilon x/\|x\|)^\top x < 0$, a contradiction. Thus $t \geq 0$.

Therefore the minimum value of t equals 0 (since $(0, 0) \in \widetilde{X}$). When $t = 0$, feasibility requires $v_i^\top x \geq 0$ for all i, which again forces $x = 0$ by the same argument using $0 \in \mathrm{int}(Q)$. Hence $(x, t) = (0, 0)$ is the unique minimizer in $\widetilde{X}$, and consequently $z_0 = (\mathbf{1}, \mathbf{1}, 0)$ is the unique minimizer in X:

$$X^\star(c_0) = \{z_0\}.$$

Step 3: pointwise sufficiency reduces to cone containment. Since $\mathcal{D} = \varnothing$, the data-consistent fiber equals $\mathcal{C}$. Pointwise sufficiency at c_0 requires a decision $z^\star \in X$ that is optimal for all costs in $\mathcal{C}$ (and in particular for c_0). By Step 2, this forces $z^\star = z_0$. Therefore,

$$\mathcal{D} = \varnothing \text{ is pointwise sufficient at } c_0 \iff z_0 \in X^\star(c) \ \ \forall c \in \mathcal{C} \iff \mathcal{C} \subseteq \Lambda(z_0).$$

Step 4: $\mathcal{C} \subseteq \Lambda(z_0)$ iff $P \subseteq Q$. By (16), for any $y \in \mathbb{R}^d$ we have

$$((y, 1), 0, 0) \in \Lambda(z_0) \iff (y, 1) \in \mathrm{cone}\{\bar{v}_1, \ldots, \bar{v}_M\} \iff y \in Q,$$

where the last equivalence uses that $(y, 1) = \sum_{i=1}^M \alpha_i (v_i, 1)$ with $\alpha_i \geq 0$ holds if and only if $\sum_{i=1}^M \alpha_i = 1$ and $y = \sum_{i=1}^M \alpha_i v_i$. Applying this pointwise over $y \in P = [-1, 1]^d$ yields $\mathcal{C} \subseteq \Lambda(z_0) \iff P \subseteq Q$.

This completes a polynomial-time reduction from H-in-V containment to checking pointwise sufficiency, proving coNP-hardness.

Finally, since our reduction uses $\mathcal{D} = \varnothing$, deciding whether the minimum pointwise SDD size equals 0 is already coNP-hard; hence, computing the minimum size (and producing a minimum pointwise SDD) is coNP-hard.

B.3. Proof of Theorem 10

Theorem 32 (Formal version of Theorem 10) *The following decision problem is* coNP*-hard: given a bounded polytope $X \subseteq \mathbb{R}^n$ and a polyhedral uncertainty set $\mathcal{C} \subseteq \mathbb{R}^n$ specified in H-representation, decide whether the empty dataset $\mathcal{D} = \varnothing$ is a global SDD for $(X, \mathcal{C})$ in the sense of Definition 1. The hardness persists even when $\mathcal{C}$ is an open, full-dimensional polyhedron. Consequently, computing the size of a minimum global SDD (and the corresponding search problem of finding one) is* coNP*-hard.*

Proof We reduce from the same restricted H-in-V polytope containment problem used in the proof of Theorem 9: given a full-dimensional polytope $Q = \mathrm{conv}\{v_1, \ldots, v_M\} \subseteq \mathbb{R}^d$ with $0 \in \mathrm{int}(Q)$, decide whether $P = [-1, 1]^d \subseteq Q$, which is coNP-complete. We reuse the standard-form polytope X from (14). In particular, $X = \{z = (w, r, s) \in \mathbb{R}^n : Az = b,\ z \geq 0\}$ with $n = 2(d + 1) + M$, and it contains the distinguished vertex $z_0 = (\mathbf{1}, \mathbf{1}, 0)$.

A full-dimensional open uncertainty set. Write costs as $c = (c_w, c_r, c_s) \in \mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \times \mathbb{R}^M$ with $n_0 = d + 1$. Define the *effective cost on* w by the linear map

$$T(c) \;:=\; c_w - c_r + \bar{V}^\top c_s \;\in\; \mathbb{R}^{n_0}. \tag{17}$$

Indeed, for any feasible $z = (w, r, s) \in X$ we have $r = 2 \cdot \mathbf{1} - w$ and $s = \bar{V}w - \beta$, so

$$c^\top z = c_w^\top w + c_r^\top (2 \cdot \mathbf{1} - w) + c_s^\top (\bar{V}w - \beta) = T(c)^\top w \;+\; \underbrace{2c_r^\top \mathbf{1} - c_s^\top \beta}_{\text{constant over } X}. \tag{18}$$

Therefore $\arg\min_{z \in X} c^\top z$ depends on c only through $T(c)$.

Let $B_{\mathrm{op}} \subseteq \mathbb{R}^{n_0}$ denote the open polyhedron

$$B_{\mathrm{op}} := \Big\{ (\tilde{y}, \tilde{t}) \in \mathbb{R}^d \times \mathbb{R} :\; \tfrac{1}{2} < \tilde{t} < \tfrac{3}{2},\;\; -\tilde{t} < \tilde{y}_j < \tilde{t} \;\; \forall j \in [d] \Big\}.$$

We define the uncertainty set as the preimage

$$\mathcal{C}_{\mathrm{op}} := \{\, c \in \mathbb{R}^n :\; T(c) \in B_{\mathrm{op}} \,\}. \tag{19}$$

Since T is linear and B_{op} is an open polyhedron, $\mathcal{C}_{\mathrm{op}}$ is also an open polyhedron. Moreover, it is full-dimensional because it is open and nonempty (e.g., $((0,1), 0, 0) \in \mathcal{C}_{\mathrm{op}}$).

We claim that

$$\mathcal{D} = \varnothing \text{ is a global SDD for } (X, \mathcal{C}_{\mathrm{op}}) \quad \Longleftrightarrow \quad P \subseteq Q. \tag{20}$$

Since $P \subseteq Q$ is coNP-hard, this proves the theorem.

Step 1: Containment is equivalent to a cone inclusion. Recall that $\mathrm{cone}\{\bar{v}_1, \ldots, \bar{v}_M\} = \{(tz, t) : t \geq 0,\; z \in Q\}$. In particular, for any $\tilde{t} > 0$ we have

$$(\tilde{y}, \tilde{t}) \in \mathrm{cone}\{\bar{v}_1, \ldots, \bar{v}_M\} \quad \Longleftrightarrow \quad \tilde{y}/\tilde{t} \in Q.$$

Therefore,

$$B_{\mathrm{op}} \subseteq \mathrm{cone}\{\bar{v}_1, \ldots, \bar{v}_M\} \quad \Longleftrightarrow \quad (-1, 1)^d \subseteq Q.$$

Because Q is closed, $(-1,1)^d \subseteq Q$ holds if and only if $[-1,1]^d \subseteq Q$, i.e., $P \subseteq Q$.

Step 2: If $P \subseteq Q$, then $\mathcal{D} = \varnothing$ is a global SDD. Assume $P \subseteq Q$. By Step 1 we have $B_{\mathrm{op}} \subseteq \mathrm{cone}\{\bar{v}_1, \ldots, \bar{v}_M\}$, and since B_{op} is open this implies

$$B_{\mathrm{op}} \subseteq \mathrm{int}\big(\mathrm{cone}\{\bar{v}_1, \ldots, \bar{v}_M\}\big).$$

Next, note that w uniquely determines (r, s) via the equalities $w + r = 2 \cdot \mathbf{1}$ and $\bar{V}w - s = \beta$. Thus minimizing $c^\top z$ over X is equivalent to minimizing $T(c)^\top w$ over the projected feasible set

$$W := \{w \in \mathbb{R}^{n_0} :\; \exists (r, s) \geq 0 \text{ s.t. } (w, r, s) \in X\}.$$

Equivalently, with the change of variables $u := w - \mathbf{1} = (x, t)$, this projected set corresponds to the bounded polytope

$$\widetilde{X} = \{(x, t) \in \mathbb{R}^d \times \mathbb{R} :\; v_i^\top x + t \geq 0 \;\forall i \in [M],\;\; -1 \leq (x, t) \leq 1\}.$$

At $u_0 := (0,0) \in \widetilde{X}$, the box constraints are slack and the active constraints are $v_i^\top x + t \geq 0$, i.e., $-\bar{v}_i^\top u \leq 0$. Hence $\Lambda(u_0) = \mathrm{cone}\{\bar{v}_1, \ldots, \bar{v}_M\}$. If $(\tilde{y}, \tilde{t}) \in \mathrm{int}(\Lambda(u_0))$, then u_0 is the ***unique*** minimizer for this cost: for any $u \in \widetilde{X} \setminus \{u_0\}$ the direction $u - u_0$ is a nonzero feasible direction at u_0, so $(\tilde{y}, \tilde{t})^\top (u - u_0) > 0$ and hence $(\tilde{y}, \tilde{t})^\top u > (\tilde{y}, \tilde{t})^\top u_0$.

Now fix any $c \in \mathcal{C}_{\mathrm{op}}$. Then $T(c) \in B_{\mathrm{op}} \subseteq \mathrm{int}(\Lambda(u_0))$, so the unique minimizer of $\min\{T(c)^\top u : u \in \widetilde{X}\}$ is u_0. By (18), the minimizer set of $\min_{z \in X} c^\top z$ is therefore the singleton $\{z_0\}$. Thus $X^\star(c) = \{z_0\}$ for all $c \in \mathcal{C}_{\mathrm{op}}$, and the constant rule $\widehat{X}(\emptyset) := \{z_0\}$ makes $\mathcal{D} = \varnothing$ a global SDD.

Step 3: If $P \not\subseteq Q$, then no zero-query global SDD exists. Now assume $P \not\subseteq Q$. By Step 1 there exists $(\tilde{y}, \tilde{t}) \in B_{\mathrm{op}}$ such that $(\tilde{y}, \tilde{t}) \notin \mathrm{cone}\{\bar{v}_1, \ldots, \bar{v}_M\}$. Let $c_1 := ((\tilde{y}, \tilde{t}), 0, 0) \in \mathcal{C}_{\mathrm{op}}$, so $T(c_1) = (\tilde{y}, \tilde{t})$. Since $(\tilde{y}, \tilde{t}) \notin \Lambda(u_0)$, we have $u_0 \notin \widetilde{X}^\star((\tilde{y}, \tilde{t}))$, and thus $z_0 \notin X^\star(c_1)$.

On the other hand, $c_0 := ((0,1), 0, 0) \in \mathcal{C}_{\mathrm{op}}$ and by the argument in the proof of Theorem 9 (Step 2 there), we have $X^\star(c_0) = \{z_0\}$. Therefore $X^\star(c_1) \neq X^\star(c_0)$, i.e., the optimal solution set is not constant over $c \in \mathcal{C}_{\mathrm{op}}$.

Since $|\mathcal{D}| = 0$, any candidate decoder $\widehat{X} : \mathbb{R}^{|\mathcal{D}|} \to \mathcal{P}(X)$ is necessarily constant. It cannot match two distinct optimal solution sets, so no such map can satisfy Definition 1. This establishes (20).

Consequence for minimum-size global SDDs. Finally, if one could compute the size of a minimum global SDD (or output such a dataset) in polynomial time, then one could decide whether the optimum size equals 0, which is exactly the coNP-hard decision problem in (20). ■

B.4. Proof of Property 8(i)

Proof Let $\mathcal{D} \subseteq \mathcal{D}'$ and fix $c \in \mathcal{C}$. Write $s := s(c; \mathcal{D})$ and $s' := s(c; \mathcal{D}')$. By the definition of the fiber,

$$\mathcal{C}(\mathcal{D}', s') = \{c' \in \mathcal{C} : q^\top c' = q^\top c \ \ \forall q \in \mathcal{D}'\}.$$

Since $\mathcal{D} \subseteq \mathcal{D}'$, any $c' \in \mathcal{C}(\mathcal{D}', s')$ satisfies $q^\top c' = q^\top c$ for all $q \in \mathcal{D}$, and hence $c' \in \mathcal{C}(\mathcal{D}, s)$. Therefore $\mathcal{C}(\mathcal{D}', s') \subseteq \mathcal{C}(\mathcal{D}, s)$.

Now assume $\mathcal{D}$ is pointwise sufficient at c. Then there exists $x^\star \in \mathcal{X}$ such that $x^\star \in \mathcal{X}^\star(c'')$ for all $c'' \in \mathcal{C}(\mathcal{D}, s)$. By the containment above, the same $x^\star$ is optimal for all $c'' \in \mathcal{C}(\mathcal{D}', s')$, so $\mathcal{D}'$ is also pointwise sufficient at c. ■

B.5. Proof of Property 8(ii)

Proof Let $\mathcal{D}$ be a global SDD for $(\mathcal{X}, \mathcal{C})$ in the sense of Definition 1. By definition, there exists a mapping $\widehat{X} : \mathbb{R}^{|\mathcal{D}|} \to \mathcal{P}(\mathcal{X})$ such that for every $c \in \mathcal{C}$,

$$\widehat{X}\big(s(c; \mathcal{D})\big) \ = \ \mathcal{X}^\star(c).$$

Fix any $c \in \mathcal{C}$ and write $s := s(c; \mathcal{D})$. For any $c' \in \mathcal{C}(\mathcal{D}, s)$ we have $s(c'; \mathcal{D}) = s$ by definition of the fiber, and therefore

$$\widehat{X}(s) \ = \ \widehat{X}\big(s(c'; \mathcal{D})\big) \ = \ \mathcal{X}^\star(c').$$

In particular, $\widehat{X}(s)$ is nonempty, and any choice of $x^\star \in \widehat{X}(s)$ satisfies $x^\star \in \mathcal{X}^\star(c')$ for all $c' \in \mathcal{C}(\mathcal{D}, s)$. Thus the single decision $x^\star$ is optimal for all costs in the fiber, meaning that $\mathcal{D}$ is pointwise sufficient at c in the sense of Definition 6. ■

Appendix C. Proofs and technical details for Section 4

C.1. Closed-form FI for ellipsoids

Recall the face-intersection subproblem

$$\mathrm{FI}(\delta; Q, s) \;:=\; \min\{\delta^\top c :\; c \in C,\; Q^\top c = s\}.$$

Proposition 33 *Let $C = \{c \in \mathbb{R}^d :\; (c-\bar{c})^\top \Sigma^{-1}(c-\bar{c}) \le R^2\}$ with $\Sigma \succ 0$. Fix $Q \in \mathbb{R}^{d\times k}$ with $\mathrm{rank}(Q) = k$ and $s \in \mathbb{R}^k$, and assume $C(Q,s) := \{c \in C :\; Q^\top c = s\} \neq \emptyset$. Define*

$$c^\perp := \bar{c} + \Sigma Q(Q^\top \Sigma Q)^{-1}(s - Q^\top \bar{c}), \qquad M^\perp := \Sigma - \Sigma Q(Q^\top \Sigma Q)^{-1} Q^\top \Sigma \succeq 0,$$

and

$$\rho := \sqrt{R^2 - (c^\perp - \bar{c})^\top \Sigma^{-1}(c^\perp - \bar{c})}.$$

Then for any $\delta \in \mathbb{R}^d$,

$$\min_{c \in C(Q,s)} \delta^\top c = \delta^\top c^\perp - \rho\sqrt{\delta^\top M^\perp \delta}.$$

If $\delta^\top M^\perp \delta > 0$, a minimizer is

$$c^{\mathrm{out}}(\delta) = c^\perp - \rho\,\frac{M^\perp \delta}{\sqrt{\delta^\top M^\perp \delta}}.$$

If $\delta^\top M^\perp \delta = 0$, then $\delta^\top c$ is constant over $C(Q,s)$.

Proof Let $\Sigma^{1/2}$ be the symmetric square root of Σ (so $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$). Change variables

$$z := \Sigma^{-1/2}(c - \bar{c}) \quad\Longleftrightarrow\quad c = \bar{c} + \Sigma^{1/2} z.$$

Then the ellipsoid constraint becomes $\|z\|_2 \le R$. The equality constraint becomes

$$Q^\top c = s \quad\Longleftrightarrow\quad Q^\top(\bar{c} + \Sigma^{1/2} z) = s \quad\Longleftrightarrow\quad (\Sigma^{1/2}Q)^\top z = s - Q^\top \bar{c}.$$

Define $\widetilde{Q} := \Sigma^{1/2} Q$ and $\widetilde{s} := s - Q^\top \bar{c}$, and also $\widetilde{\delta} := \Sigma^{1/2}\delta$. Up to the constant $\delta^\top \bar{c}$, the FI problem is equivalent to

$$\min\{\widetilde{\delta}^\top z :\; \|z\|_2 \le R,\; \widetilde{Q}^\top z = \widetilde{s}\}.$$

Let

$$z_0 := \widetilde{Q}(\widetilde{Q}^\top \widetilde{Q})^{-1}\widetilde{s},$$

the unique minimum-ℓ_2-norm solution of $\widetilde{Q}^\top z = \widetilde{s}$. Every feasible z can be written uniquely as $z = z_0 + v$ with $v \in \ker(\widetilde{Q}^\top)$. Since $z_0 \in \mathrm{span}(\widetilde{Q})$ and $\ker(\widetilde{Q}^\top) = \mathrm{span}(\widetilde{Q})^\perp$, we have the orthogonal decomposition $\|z\|_2^2 = \|z_0\|_2^2 + \|v\|_2^2$. Thus feasibility is equivalent to $\|z_0\|_2 \le R$ (which holds since the fiber is nonempty) and

$$\|v\|_2 \le \rho_z := \sqrt{R^2 - \|z_0\|_2^2}.$$

Let $P := I - \widetilde{Q}(\widetilde{Q}^\top \widetilde{Q})^{-1}\widetilde{Q}^\top$ denote the orthogonal projector onto $\ker(\widetilde{Q}^\top)$. Then for $v \in \ker(\widetilde{Q}^\top)$, $\widetilde{\delta}^\top v = (P\widetilde{\delta})^\top v$. Therefore,

$$\min_{\substack{v \in \ker(\widetilde{Q}^\top) \\ \|v\|_2 \le \rho_z}} \widetilde{\delta}^\top v = \min_{\|v\|_2 \le \rho_z} (P\widetilde{\delta})^\top v = -\rho_z\,\|P\widetilde{\delta}\|_2,$$

with minimizer $v^\star = -\rho_z \frac{P\widetilde{\delta}}{\|P\widetilde{\delta}\|_2}$ when $P\widetilde{\delta} \neq 0$ (and any v when $P\widetilde{\delta} = 0$).

Hence, the optimal value in z-space is

$$\widetilde{\delta}^\top z_0 - \rho_z \, \|P\widetilde{\delta}\|_2,$$

and the optimal c is $c = \bar{c} + \Sigma^{1/2}(z_0 + v^\star)$.

It remains to express everything back in the original variables. First, note that $\widetilde{Q}^\top \widetilde{Q} = Q^\top \Sigma Q$ and

$$\Sigma^{1/2} z_0 = \Sigma^{1/2} \widetilde{Q}(\widetilde{Q}^\top \widetilde{Q})^{-1} \widetilde{s} = \Sigma Q (Q^\top \Sigma Q)^{-1}(s - Q^\top \bar{c}).$$

Thus

$$c^\perp = \bar{c} + \Sigma^{1/2} z_0 = \bar{c} + \Sigma Q (Q^\top \Sigma Q)^{-1}(s - Q^\top \bar{c}).$$

Second,

$$\|z_0\|_2^2 = z_0^\top z_0 = (c^\perp - \bar{c})^\top \Sigma^{-1} (c^\perp - \bar{c}),$$

so $\rho_z = \rho$ as defined in the statement.

Finally,

$$\|P\widetilde{\delta}\|_2^2 = \widetilde{\delta}^\top P \widetilde{\delta} = \delta^\top \Big(\Sigma - \Sigma Q (Q^\top \Sigma Q)^{-1} Q^\top \Sigma \Big) \delta = \delta^\top M^\perp \delta.$$

Also,

$$\Sigma^{1/2} P \widetilde{\delta} = \Big(\Sigma - \Sigma Q (Q^\top \Sigma Q)^{-1} Q^\top \Sigma \Big) \delta = M^\perp \delta.$$

Substituting these identities yields the claimed closed-form expressions for the minimum value and the optimizer. ■

C.2. Proof of Lemma 16

Proof Because B is an optimal basis for $c^{\text{in}} \in \mathcal{C}_k$, we have $c^{\text{in}} \in \Lambda(B)$, i.e. $(c^{\text{in}})^\top \delta(B, j) \geq 0$ for all $j \in N$. Since we are in the `Else` branch, there exists $c^{\text{out}} \in \mathcal{C}_k$ with $(c^{\text{out}})^\top \delta(B, j_0) = m_{\min} < 0$ for at least one j_0. Define the segment $c_\alpha := (1-\alpha)c^{\text{in}} + \alpha c^{\text{out}}$, $\alpha \in [0, 1]$. Because $\mathcal{C}_k$ is convex, $c_\alpha \in \mathcal{C}_k$ for all $\alpha \in [0, 1]$.

By the facet-hit rule, for every $j \in N$ with $(c^{\text{out}})^\top \delta(B, j) < 0$, the value α_j is the first point along $c_\alpha = (1-\alpha)c^{\text{in}} + \alpha c^{\text{out}}$ at which $(c_\alpha)^\top \delta(B, j)$ becomes zero. Since

$$\alpha^\star = \min_{j : (c^{\text{out}})^\top \delta(B,j) < 0} \alpha_j,$$

we have $\alpha^\star \leq \alpha_j$ for every facet violated by c^{out}. Therefore each such facet is still nonnegative at $\alpha^\star$. Moreover, because $j^\star$ attains the minimum, $\alpha^\star = \alpha_{j^\star}$, and hence

$$(c_{\alpha^\star})^\top \delta(B, j^\star) = 0.$$

For any j not violated by c^{out}, both endpoint values are nonnegative, so linearity along the segment gives

$$(c_{\alpha^\star})^\top \delta(B, j) \geq 0.$$

Hence, $(c_{\alpha^\star})^\top \delta(B, j) \geq 0, \forall j \in N$. Thus $c^{\text{hit}} := c_{\alpha^\star} \in \mathcal{C}_k \cap \Lambda(B)$ and lies on the face $\Lambda(B) \cap \{\delta(B, j^\star)\}^\perp$.

By Equation (2) (with $x^\star = x(B)$ and $\mathcal{C} = \mathcal{C}_k$), this implies $\delta(B, j^\star) \in \Delta(\mathcal{X}, \mathcal{C}_k)$. Since $\mathcal{C}_k \subseteq \mathcal{C}$, we also have $\Delta(\mathcal{X}, \mathcal{C}_k) \subseteq \Delta(\mathcal{X}, \mathcal{C})$. Finally, by Theorem 3,

$$\delta(B, j^\star) \in \operatorname{span} \Delta(\mathcal{X}, \mathcal{C}) = \operatorname{dir}(\mathcal{X}^\star(\mathcal{C})),$$

and in particular $\delta(B, j^\star) \in \operatorname{dir}(\mathcal{X}^\star(\mathcal{C}))$. ■

C.3. Proof of Lemma 17

Proof Assume for contradiction that $q_{k+1} \in \operatorname{span}(Q_k)$. Then $(q_{k+1})^\top c'$ is constant over $\mathcal{C}_k = \{c' \in \mathcal{C} : Q_k^\top c' = s_k\}$. In particular, $(q_{k+1})^\top c^{\text{in}} = (q_{k+1})^\top c^{\text{out}}$. But $c^{\text{in}} \in \Lambda(B)$ implies $(q_{k+1})^\top c^{\text{in}} \geq 0$, while the facet-hit rule guarantees $(q_{k+1})^\top c^{\text{out}} < 0$. This contradiction shows $q_{k+1} \notin \operatorname{span}(Q_k)$, and therefore $\operatorname{rank}(Q_{k+1}) = \operatorname{rank}(Q_k) + 1$. Moreover, since $d^\star = \dim(\operatorname{dir}(\mathcal{X}^\star(\mathcal{C})))$ and $q_{k+1} \in \operatorname{dir}(\mathcal{X}^\star(\mathcal{C}))$ by Lemma 16, the algorithm makes at most $d^\star$ queries. ■

C.4. Proof of Theorem 18

Proof At termination we have $m_{\min} \geq 0$, hence for every $j \in N$,

$$\min_{c' \in \mathcal{C}_k} (c')^\top \delta(B, j) \;\geq\; 0 \quad \Rightarrow \quad (c')^\top \delta(B, j) \geq 0 \;\; \forall c' \in \mathcal{C}_k.$$

By (4), this implies $\mathcal{C}_k \subseteq \Lambda(B)$. Therefore the fixed decision $x(B)$ is optimal for every $c' \in \mathcal{C}_k = \mathcal{C}(\mathcal{D}, s(c; \mathcal{D}))$, so $\mathcal{D}$ is pointwise sufficient at c.

By Lemma 17, the algorithm makes at most $d^\star$ new queries. Since each non-terminating iteration adds exactly one new query direction, the `while` loop executes at most $d^\star + 1$ iterations. ■

C.5. Proof of Property 19

Proof We bound the computational work in one iteration of Algorithm 1. Each iteration performs the following optimization subroutines.

(i) **One LP over $\mathcal{X}$.** Line 5 solves

$$\min\{(c^{\text{in}})^\top x : \; Ax = b, \; x \geq 0\}.$$

Since $\mathcal{X}$ is given in standard form, this LP can be solved in time polynomial in the bit complexity of the input.

(ii) **$d - m$ face-intersection subproblems over $\mathcal{C}$.** For each $j \in N$ (so $|N| = d - m$), line 6 of Algorithm 1 solves

$$\min\{\delta_j^\top c' : \; c' \in \mathcal{C}, \; Q_k^\top c' = s_k\}.$$

If $\mathcal{C}$ is a polytope given in H-representation, $\mathcal{C} = \{c : \; Gc \leq h\}$, then this is the LP

$$\min\{\delta_j^\top c' : \; Gc' \leq h, \; Q_k^\top c' = s_k\},$$

whose size is polynomial in the input (in particular, it has dimension d and $k \le d^\star$ equality constraints). Hence, each FI call can be solved in polynomial time. If instead $\mathcal{C}$ is an ellipsoid, Proposition 33 gives a closed-form expression for the optimal value and a minimizer, which can be computed in polynomial time (it amounts to solving a $k \times k$ linear system and basic matrix–vector operations).

Thus, each iteration runs in polynomial time and makes at most $d - m$ FI calls. Finally, by Theorem 18, Algorithm 1 executes at most $d^\star + 1 \le d + 1$ iterations and makes at most $d^\star$ oracle queries of the form $q^\top c$. Combining the per-iteration bound with the iteration bound yields the claimed overall polynomial running time. ■

C.6. A counterexample motivating the facet-hit rule

We give a simple example showing why selecting an arbitrary violated facet at a witness point can fail. Let $\mathcal{X} = [0,1]^2$ and consider the vertex $x = (0,0)$, whose optimality cone is

$$\Lambda = \{c \in \mathbb{R}^2 : \ c_1 \ge 0, \ c_2 \ge 0\},$$

with facet hyperplanes $c_1 = 0$ and $c_2 = 0$. Let $\varepsilon \in (0,1)$ and define two points $c^{\text{in}} = (1, \varepsilon)$ and $c^{\text{out}} = (-1, -1)$. Consider any convex fiber $\mathcal{C}_k$ whose intersection with Λ is the segment $\text{conv}\{c^{\text{in}}, c^{\text{out}}\}$ (for instance, take $\mathcal{C}_k$ to be exactly this segment).

Then c^{out} violates *both* inequalities $c_1 \ge 0$ and $c_2 \ge 0$. However, along the segment $c_\alpha = (1-\alpha)c^{\text{in}} + \alpha c^{\text{out}}$, the coordinate $c_{2,\alpha}$ hits 0 at a very small α, while $c_{1,\alpha}$ is still strictly positive; thus $\mathcal{C}_k \cap \Lambda$ reaches the boundary only through the facet $c_2 = 0$. In contrast, the segment intersects $c_1 = 0$ only after c_2 has already become negative, i.e., *outside* Λ. Therefore, querying the normal of the "wrong" violated facet $c_1 = 0$ does not correspond to a boundary that the fiber can reach while keeping x optimal. The facet-hit rule avoids this issue by explicitly selecting the *first* facet reached from an interior anchor point.

Appendix D. Proofs and technical details for Section 5

D.1. Proof of Lemma 20

Proof When processing c_i, the inner run of Algorithm 1 certifies pointwise sufficiency for c_i, so $\ell(\mathcal{D}_i, c_i) = 0$. For $t > i$, the cumulative procedure only enlarges the dataset, $\mathcal{D}_t \supseteq \mathcal{D}_i$, and Property 8(i) implies $\ell(\mathcal{D}_t, c_i) = 0$ for all later t, hence also at $t = n$. ■

D.2. Proof of Lemma 21

Proof In Algorithm 2, the dataset only changes at indices in T by definition. Removing an index $t \notin T$ removes an iteration that would have run Algorithm 1 with initialization $\mathcal{D}_{t-1}$ and returned the same dataset $\mathcal{D}_t = \mathcal{D}_{t-1}$. Therefore, deleting all non-hard iterations leaves the dataset entering each hard iteration unchanged, and thus reproduces the same sequence of dataset updates and the same final dataset. ■

D.3. Proof of Lemma 22

Proof By Lemma 17, each time Algorithm 1 appends a new query, that direction is linearly independent of the previously queried directions in that run. Since Algorithm 2 warm-starts each run at $\mathcal{D}_{i-1}$, this implies by induction over $i = 1, \dots, n$ that the cumulative datasets $\mathcal{D}_i$ remain linearly independent.

By definition, $i \in T$ implies $\mathcal{D}_i \neq \mathcal{D}_{i-1}$, hence at least one new independent direction was added while processing c_i, so $|T| \leq |\mathcal{D}_n|$.

Finally, by Lemma 16, every query direction that Algorithm 1 can append lies in $\mathrm{dir}(\mathcal{X}^\star(\mathcal{C}))$. Since $\mathcal{D}_n$ is the union of all appended directions across the cumulative run, we have $\mathcal{D}_n \subseteq \mathrm{dir}(\mathcal{X}^\star(\mathcal{C}))$. Therefore, linear independence yields $|\mathcal{D}_n| = \dim \mathrm{span}(\mathcal{D}_n) \leq \dim \mathrm{dir}(\mathcal{X}^\star(\mathcal{C})) = d^\star$. ■

D.4. Proof of Theorem 23

We apply the stable sample compression bound of Hanneke and Kontorovich (2021, Corollary 11).

We view any queried dataset D as inducing a binary prediction rule $h_D : \mathcal{C} \to \{0, 1\}$ defined by $h_D(c) := \ell(D, c) \in \{0, 1\}$. Thus $R(D) = \Pr_{C \sim P_c}[h_D(C) = 1] = \Pr_{C \sim P_c}[\ell(D, C) = 1]$ is exactly the associated 0–1 risk. Equivalently, one may regard this as a supervised distribution over (C, Y) with $C \sim P_c$ and $Y = 0$ almost surely.

A stable compression scheme. Let $S = (c_1, \dots, c_n)$ be the training sequence and let $(\mathcal{D}_n, T)$ be the output of Algorithm 2 on S. Define a compression function κ by letting $\kappa(S)$ be the subsequence of hard samples $(c_i)_{i \in T}$ (in their original order). Define a reconstruction function ρ that maps any subsequence S' to the prediction rule $h_{\mathcal{D}'}$ induced by the final dataset $\mathcal{D}'$ returned by running Algorithm 2 on S'.

By Lemma 21, running Algorithm 2 on $\kappa(S)$ reproduces the same final dataset $\mathcal{D}_n$, so $\rho(\kappa(S)) = h_{\mathcal{D}_n}$. Moreover, Lemma 21 also implies the *stability* property of Hanneke and Kontorovich (2021, Definition 8): removing any subset of the non-hard samples (i.e., elements of $S \setminus \kappa(S)$) does not affect the reconstructed output.

Zero empirical risk. Lemma 20 gives $\ell(\mathcal{D}_n, c_i) = 0$ for all $i = 1, \dots, n$, i.e., the empirical 0–1 risk of $\rho(\kappa(S)) = h_{\mathcal{D}_n}$ on S is zero.

Apply Corollary 11 of Hanneke and Kontorovich, 2021, with probability at least $1 - \delta$ over the draw of $S \sim P_c^n$,

$$R(\mathcal{D}_n) \;=\; R(h_{\mathcal{D}_n}) \;=\; R(\rho(\kappa(S))) \;\leq\; \frac{4}{n}\Big(6|\kappa(S)| + \ln\frac{e}{\delta}\Big) \;=\; \frac{4}{n}\Big(6|T| + \ln\frac{e}{\delta}\Big).$$

Finally, Lemma 22 implies $|T| \leq d^\star$, which gives the last sentence of Theorem 23.

D.5. Proof of Theorem 24

We give an explicit construction in which the intrinsic dimension $d^\star$ and the ambient dimension can be chosen independently.

Feasible region. Fix integers $d \geq d^\star \geq 2$. Let $m := d$, set $A := [I_d\ I_d] \in \mathbb{R}^{d\times 2d}$ and $b := \mathbf{1} \in \mathbb{R}^d$, and define the lifted polytope

$$\mathcal{X} := \{z = (x, s) \in \mathbb{R}^{2d} : Az = b,\ z \geq 0\} = \{(x, s) : x + s = \mathbf{1},\ x \geq 0,\ s \geq 0\},$$

which is an extended formulation of the hypercube $[0, 1]^d$ obtained by introducing slack variables. Then $\mathcal{X}$ is bounded and nondegenerate: at every extreme point, for each $j \in [d]$ exactly one of x_j, s_j equals 1 and the other equals 0, so every extreme point has exactly $m = d$ strictly positive components.

Prior set and a rare-types distribution. Let $\mu \in \mathbb{R}^d$ be defined coordinatewise by

$$\mu_j \ = \ \begin{cases} 0.99, & j \in \{1, \ldots, d^\star\}, \\ 10, & j \in \{d^\star + 1, \ldots, d\}, \end{cases}$$

and define the lifted center $\bar{\mu} := (\mu, 0) \in \mathbb{R}^{2d}$. Let the convex prior set be the lifted radius-1 Euclidean ball

$$\mathcal{C} \ := \ \{(c, 0) \in \mathbb{R}^{2d} :\ \|c - \mu\|_2 \leq 1\}.$$

For each $i \in \{1, \ldots, d^\star\}$ define the lifted costs and query directions

$$c^{(i)} \ := \ (\mu - e_i, 0) \in \mathcal{C}, \qquad \delta_i \ := \ (-e_i, e_i) \in \mathbb{R}^{2d}.$$

Fix $\varepsilon \in (0, 1/4)$ and let $k := d^\star - 1$. Define a distribution P_c supported on $\{c^{(1)}, \ldots, c^{(d^\star)}\} \subseteq \mathcal{C}$ by

$$\mathbb{P}(c = c^{(1)}) \ = \ 1 - 2\varepsilon, \qquad \mathbb{P}(c = c^{(i)}) \ = \ \frac{2\varepsilon}{k}, \quad i = 2, \ldots, d^\star.$$

We call $c^{(1)}$ the *common* type and $\{c^{(2)}, \ldots, c^{(d^\star)}\}$ the *rare* types.

Step 1: Prove that $\dim \operatorname{dir}(\mathcal{X}^\star(\mathcal{C})) = d^\star$**.**

Lemma 34 *For every $(c, 0) \in \mathcal{C}$, every minimizer $z^\star = (x^\star, s^\star) \in \arg\min_{(x,s)\in\mathcal{X}} (c, 0)^\top (x, s)$ satisfies $x^\star_j = 0$ and $s^\star_j = 1$ for all $j > d^\star$. Moreover, $\mathcal{X}^\star(\mathcal{C}) \supseteq \{(0, \mathbf{1}),\ (e_1, \mathbf{1} - e_1), \ldots, (e_{d^\star}, \mathbf{1} - e_{d^\star})\}$. Consequently,*

$$\dim \operatorname{dir}(\mathcal{X}^\star(\mathcal{C})) \ = \ d^\star.$$

Proof Fix $(c, 0) \in \mathcal{C}$ and any feasible $(x, s) \in \mathcal{X}$. Since $x + s = \mathbf{1}$, we have

$$(c, 0)^\top (x, s) = c^\top x.$$

Thus, the LP objective depends only on $x \in [0, 1]^d$. For each $j > d^\star$ we have $c_j \geq \mu_j - 1 = 9 > 0$, hence any minimizer must set $x^\star_j = 0$ (and therefore $s^\star_j = 1$) for all $j > d^\star$.

Next, $\bar{\mu} = (\mu, 0) \in \mathcal{C}$ and μ has strictly positive coordinates, so the unique minimizer is $(x, s) = (0, \mathbf{1})$. For each $i \leq d^\star$, the cost $c^{(i)} = (\mu - e_i, 0)$ has $c^{(i)}_i = -0.01 < 0$ and all other $c^{(i)}_j > 0$, so the unique minimizer is $(x, s) = (e_i, \mathbf{1} - e_i)$. This proves the stated inclusion of $\mathcal{X}^\star(\mathcal{C})$.

Therefore $\operatorname{dir}(\mathcal{X}^\star(\mathcal{C}))$ contains $\operatorname{span}\{(e_i, -e_i) : i = 1, \ldots, d^\star\}$, so $\dim \operatorname{dir}(\mathcal{X}^\star(\mathcal{C})) \geq d^\star$. On the other hand, we already showed that every optimizer has $x_j = 0$ for $j > d^\star$, hence every difference of reachable optima lies in $\operatorname{span}\{(e_i, -e_i) : i = 1, \ldots, d^\star\}$, giving $\dim \operatorname{dir}(\mathcal{X}^\star(\mathcal{C})) \leq d^\star$. ■

Step 2: Without querying δ_i, type i cannot be certified.

Lemma 35 *Fix any $i \in \{2, \dots, d^\star\}$ and let $\mathcal{D} \subseteq \mathbb{R}^{2d}$ be any dataset that does not contain δ_i. If $\mathcal{D} \subseteq \{\delta_1, \dots, \delta_{d^\star}\}$, then*

$$\bar{\mu} \in \mathcal{C}\big(\mathcal{D}, s(c^{(i)}; \mathcal{D})\big).$$

Consequently, $\mathcal{D}$ is not *pointwise sufficient at $c^{(i)}$.*

Proof Assume $\mathcal{D} \subseteq \{\delta_1, \dots, \delta_{d^\star}\}$ and $\delta_i \notin \mathcal{D}$. Every query in $\mathcal{D}$ is of the form $\delta_j = (-e_j, e_j)$ with $j \neq i$. For such j,

$$\delta_j^\top \bar{\mu} = (-e_j, e_j)^\top (\mu, 0) = -\mu_j = -(\mu - e_i)_j = (-e_j, e_j)^\top (\mu - e_i, 0) = \delta_j^\top c^{(i)}.$$

Hence $\bar{\mu}$ is consistent with the same measurements as $c^{(i)}$, i.e., $\bar{\mu} \in \mathcal{C}(\mathcal{D}, s(c^{(i)}; \mathcal{D}))$.

But by Lemma 34, $x^\star(\bar{\mu}) = 0$ (so the unique optimizer is $(0, \mathbf{1})$) while $x^\star(c^{(i)}) = e_i$ (unique optimizer $(e_i, \mathbf{1} - e_i)$). Thus, the fiber contains two costs with different unique minimizers, so no single decision can be optimal for all costs in the fiber. Therefore $\mathcal{D}$ is not pointwise sufficient at $c^{(i)}$. ■

Step 2': On type i, the pointwise routine adds only δ_i.

Lemma 36 *Fix $i \in \{1, \dots, d^\star\}$. Run Algorithm 1 on $c = c^{(i)} = (\mu - e_i, 0)$ with initialization $D_{\text{init}} \subseteq \{\delta_1, \dots, \delta_{d^\star}\}$ such that $\delta_i \notin D_{\text{init}}$. Then the call performs exactly one augmentation and returns $D = D_{\text{init}} \cup \{\delta_i\}$. In particular, during this call, the algorithm cannot add any δ_j with $j \neq i$.*

Proof Let Q_k be the matrix whose columns are the directions in D_{init} and set $s_k := Q_k^\top c^{(i)}$. Since each $\delta_j = (-e_j, e_j)$, the fiber

$$\mathcal{C}_k := \{c' \in \mathcal{C} : \; Q_k^\top c' = s_k\}$$

fixes the coordinates $c'_j = (\mu - e_i)_j = \mu_j$ for every $\delta_j \in D_{\text{init}}$ and leaves coordinate i unconstrained; all $c' \in \mathcal{C}_k$ have the form $(\tilde{c}, 0)$ with $\tilde{c} \in \mathbb{R}^d$.

(a) The first LP solve yields the vertex $(e_i, \mathbf{1} - e_i)$ and its cone. With $c_{\text{in}} = c^{(i)} = (\mu - e_i, 0)$, we have $(\mu - e_i)_i < 0$ and $(\mu - e_i)_j > 0$ for all $j \neq i$, so the LP over $\mathcal{X}$ has the unique minimizer

$$z^\star = (x^\star, s^\star) = (e_i, \mathbf{1} - e_i).$$

For the matrix $A = [I_d \; I_d]$, this vertex corresponds to the unique feasible basis

$$B(i) := \{x_i\} \cup \{s_j : \; j \neq i\}, \qquad N(i) := \{s_i\} \cup \{x_j : \; j \neq i\}.$$

For $j \neq i$, increasing the nonbasic variable x_j from 0 decreases s_j from 1 to 0, hence

$$\delta(B(i), x_j) = (e_j, -e_j).$$

Increasing the nonbasic variable s_i from 0 decreases x_i from 1 to 0, hence

$$\delta(B(i), s_i) = (-e_i, e_i) = \delta_i.$$

Therefore, the optimality cone (4) takes the explicit form

$$\Lambda(B(i)) = \Big\{(u, v) \in \mathbb{R}^{2d} : \; (u_j - v_j) \geq 0 \; \forall j \neq i, \;\; (-u_i + v_i) \geq 0\Big\}.$$

Since every $c' \in \mathcal{C}_k$ has $v = 0$, we have

$$\mathcal{C}_k \subseteq \{(u, 0) : u \in \mathbb{R}^d\}, \qquad \Lambda(B(i)) \cap \{(u, 0)\} = \{(u, 0) : \; u_j \geq 0 \; \forall j \neq i, \; u_i \leq 0\}.$$

(b) Among facets of $\Lambda(B(i))$, only the facet for δ_i can be hit from within $\mathcal{C}_k$. Because coordinate i is free in $\mathcal{C}_k$, the point $\tilde{c} := (\mu + e_i, 0)$ belongs to $\mathcal{C}_k$ (and to $\mathcal{C}$) but has $\tilde{c}_i > 0$, hence $\tilde{c} \notin \Lambda(B(i)) \cap \{(u,0)\}$. Therefore $\mathcal{C}_k \not\subseteq \Lambda(B(i))$, the containment test fails, and Algorithm 1 enters the ELSE branch, producing some witness $c_{\text{out}} \in \mathcal{C}_k \setminus \Lambda(B(i))$ and considering the segment $c_\alpha := (1-\alpha)c_{\text{in}} + \alpha c_{\text{out}}$.

We claim that for every $j \neq i$,

$$\big(\mathcal{C} \cap \Lambda(B(i))\big) \;\cap\; \{(u,v) :\; (u_j - v_j) = 0\} \;=\; \emptyset.$$

Indeed, take any $c = (u,v) \in \mathcal{C} \cap \Lambda(B(i))$. Since $c \in \mathcal{C}$ we have $v = 0$ and $\|u - \mu\|_2 \leq 1$. Moreover, $c \in \Lambda(B(i))$ implies $u_i \leq 0$. Since $\mu_i = 0.99 > 0$,

$$(u_i - \mu_i)^2 \geq (0 - \mu_i)^2 = \mu_i^2.$$

Thus,

$$\sum_{t \neq i} (u_t - \mu_t)^2 \leq 1 - (u_i - \mu_i)^2 \leq 1 - \mu_i^2,$$

and hence for every $j \neq i$,

$$|u_j - \mu_j| \leq \sqrt{1 - \mu_i^2} \quad \Rightarrow \quad u_j \geq \mu_j - \sqrt{1 - \mu_i^2}.$$

With $\mu_j \in \{0.99, 10\}$ and $\sqrt{1 - \mu_i^2} = \sqrt{1 - 0.99^2} < 0.15$, we get $u_j > 0$ for all $j \neq i$. Therefore $u_j - v_j = u_j > 0$ for all $j \neq i$, proving the claim.

Now let $\alpha^\star$ be the first parameter where the segment leaves $\operatorname{relint}(\Lambda(B(i)))$. Then $c_{\alpha^\star} \in \mathcal{C}_k \cap \Lambda(B(i))$ lies on a facet hyperplane of $\Lambda(B(i))$. By the claim above, it cannot lie on any facet $(u_j - v_j) = 0$ with $j \neq i$; hence, it must lie on the facet $(-u_i + v_i) = 0$, whose normal is exactly $\delta(B(i), s_i) = \delta_i$. Therefore, the facet-hit rule appends $q_{k+1} = \delta_i$.

(c) After adding δ_i, the fiber becomes a singleton and the routine terminates. Appending $\delta_i = (-e_i, e_i)$ adds the constraint $\delta_i^\top (c', 0) = \delta_i^\top (\mu - e_i, 0)$, i.e., $-u_i = (1 - \mu_i)$ and hence $u_i = \mu_i - 1$. Then $(u_i - \mu_i)^2 = 1$, and the radius-1 constraint $\|u - \mu\|_2^2 \leq 1$ forces $\sum_{t \neq i} (u_t - \mu_t)^2 = 0$, hence $u = \mu - e_i$ and $c' = c^{(i)}$. Therefore the updated fiber is the singleton $\{c^{(i)}\}$, the containment test succeeds, and Algorithm 1 terminates after exactly one augmentation, returning $D = D_{\text{init}} \cup \{\delta_i\}$. ■

Step 3: A coupon-collector lower bound for Algorithm 2. Let $I \subseteq \{2, \ldots, d^\star\}$ be the set of rare indices that appear at least once among the n i.i.d. samples. By Lemma 36, Algorithm 2 learns δ_i if and only if $i \in I$. By Lemma 35, the final dataset $\mathcal{D}_n$ fails on every rare type $i \notin I$. Therefore

$$R(\mathcal{D}_n) \;\geq \sum_{i \in \{2,\ldots,d^\star\} \setminus I} \mathbb{P}(c = c^{(i)}) \;=\; \frac{2\varepsilon}{k} \big|\{2, \ldots, d^\star\} \setminus I\big|.$$

Let N be the number of rare samples among $c_1, \ldots, c_n$:

$$N \;=\; \sum_{t=1}^{n} \mathbf{1}\{c_t \neq c^{(1)}\} \;\sim\; \text{Binomial}(n, 2\varepsilon).$$

Since $|I| \leq N$, the event $N < k/2$ implies $|\{2, \ldots, d^\star\} \setminus I| \geq k/2$, and hence $R(\mathcal{D}_n) \geq \varepsilon$. It remains to lower bound $\mathbb{P}(N < k/2)$. If $n \leq k/(8\varepsilon)$, then $\mathbb{E}[N] = 2\varepsilon n \leq k/4$, and Markov's inequality gives

$$\mathbb{P}\Big(N \geq \frac{k}{2}\Big) \;\leq\; \frac{\mathbb{E}[N]}{k/2} \;\leq\; \frac{k/4}{k/2} \;=\; \frac{1}{2}.$$

Hence $\mathbb{P}(N < k/2) \geq 1/2$, and on this event we have $R(\mathcal{D}_n) \geq \varepsilon$. This concludes the proof of Theorem 24.

Appendix E. Proofs and technical details for Section 6

E.1. Ellipsoidal lifting

Throughout this appendix, we assume the shifted ellipsoidal prior

$$\mathcal{C} := \{c \in \mathbb{R}^d : \; (c - c_0)^\top \Sigma^{-1} (c - c_0) \leq 1\}, \qquad \Sigma \succ 0, \; c_0 \in \mathbb{R}^d.$$

For any orthonormal basis $U \in \mathbb{R}^{d \times t}$ we define the lifting matrix

$$\mathcal{L}_U := \Sigma U (U^\top \Sigma U)^{-1},$$

and the associated canonical lifting map $\mathrm{lift}_U : \mathbb{R}^t \to \mathbb{R}^d$ by

$$\mathrm{lift}_U(s) := c_0 + \mathcal{L}_U s.$$

Lemma 37 *For any $s \in \mathbb{R}^t$, $\mathrm{lift}_U(s)$ is the unique solution to*

$$\min_{c \in \mathbb{R}^d} \; \frac{1}{2}(c - c_0)^\top \Sigma^{-1} (c - c_0) \quad \textit{s.t.} \quad U^\top (c - c_0) = s.$$

In particular, it satisfies $U^\top(\mathrm{lift}_U(s) - c_0) = s$ and

$$(\mathrm{lift}_U(s) - c_0)^\top \Sigma^{-1} (\mathrm{lift}_U(s) - c_0) = s^\top (U^\top \Sigma U)^{-1} s.$$

Consequently, $\mathrm{lift}_U(s) \in \mathcal{C}$ whenever $s \in U^\top(\mathcal{C} - c_0)$. When $\Sigma = I$, we have $\mathrm{lift}_U(s) = c_0 + Us$.

Proof Let $z := c - c_0$. The problem becomes

$$\min_{z \in \mathbb{R}^d} \; \frac{1}{2} z^\top \Sigma^{-1} z \quad \text{s.t.} \quad U^\top z = s,$$

which is exactly the centered-ellipsoid case after the change of variables. The Lagrangian is $\mathcal{L}(z, \lambda) = \frac{1}{2} z^\top \Sigma^{-1} z - \lambda^\top (U^\top z - s)$. Stationarity gives $\Sigma^{-1} z - U\lambda = 0$, hence $z = \Sigma U \lambda$. Imposing the constraint yields $s = U^\top z = U^\top \Sigma U \lambda$, so $\lambda = (U^\top \Sigma U)^{-1} s$ and $z = \Sigma U (U^\top \Sigma U)^{-1} s = \mathcal{L}_U s$. The claimed identities follow by direct substitution. ■

The lifting map is introduced for a simple but important reason: To use such a prediction in the original optimization problem, we must map it back to a full cost vector in $\mathbb{R}^d$. When $\mathcal{C}$ is not centered at the origin, a purely linear compressed predictor would typically have range contained in a linear subspace through the origin and hence may not lie in $\mathcal{C}$. The canonical lifting map provides a principled way to "complete" a low-dimensional coordinate into a cost vector that is feasible for the prior set $\mathcal{C}$.

E.2. SPO+ formula and its subgradient

This subsection recalls the convex surrogate loss $\ell_{\mathrm{SPO}+}$ introduced by Elmachtoub and Grigas (2022) for contextual linear optimization. While the true SPO loss directly measures downstream decision regret, it is generally nonconvex and can be discontinuous in the prediction $\hat{c}$, which makes direct empirical risk minimization challenging. The SPO+ loss is a convex upper bound on the SPO loss and, crucially, it admits an oracle-based evaluation. Both $\ell_{\mathrm{SPO}+}(\hat{c}, c)$ and a stochastic subgradient can be computed using linear-optimization oracle $x^\star(\cdot)$.

$$\ell_{\mathrm{SPO}+}(\hat{c}, c) \;:=\; \max_{x\in\mathcal{X}} (c - 2\hat{c})^\top x \;+\; 2\hat{c}^\top x^\star(c) \;-\; c^\top x^\star(c). \tag{21}$$

The maximization term in (21) is the support function $\sigma_{\mathcal{X}}(c - 2\hat{c}) := \max_{x\in\mathcal{X}} (c - 2\hat{c})^\top x$, and thus $\hat{c} \mapsto \ell_{\mathrm{SPO}+}(\hat{c}, c)$ is convex but generally nonsmooth (e.g., when multiple maximizers exist). We use the deterministic oracle $x^\star(\cdot)$ to select a canonical optimizer, which leads to a well-defined stochastic subgradient in (22).

Let $x^0 = x^\star(c)$ and let $x^1 \in \arg\max_{x\in\mathcal{X}} (c - 2\hat{c})^\top x$. Since $\arg\max_{x\in\mathcal{X}} (c - 2\hat{c})^\top x = \arg\min_{x\in\mathcal{X}} (2\hat{c} - c)^\top x$, we can take $x^1 = x^\star(2\hat{c} - c)$. By Danskin's theorem applied to $\hat{c} \mapsto \max_{x\in\mathcal{X}} (c - 2\hat{c})^\top x$, we obtain the subgradient

$$\partial_{\hat{c}} \max_{x\in\mathcal{X}} (c - 2\hat{c})^\top x \;\ni\; -2x^1.$$

The remaining terms $2\hat{c}^\top x^0 - c^\top x^0$ contribute subgradient $2x^0$ in $\hat{c}$. Therefore,

$$2(x^0 - x^1) \;\in\; \partial_{\hat{c}}\, \ell_{\mathrm{SPO}+}(\hat{c}, c), \tag{22}$$

which shows that computing a stochastic subgradient requires at most two oracle calls per sample: one at c and one at $2\hat{c} - c$.

E.3. Subgradient chain rule for the compressed parametrization

We next show how to backpropagate a stochastic subgradient of $\ell_{\mathrm{SPO}+}$ through the compressed parametrization used in Stage II. Under our model-compression setup, the lifting map is an affine transformation $\hat{c} = c_0 + \mathcal{L}_U g$, so the subgradient propagation from $\hat{c}$ to the low-dimensional coordinate g reduces to a simple multiplication by $\mathcal{L}_U^\top$.

Fix an orthonormal basis $U \in \mathbb{R}^{d\times t}$ and let $\mathcal{L}_U = \Sigma U (U^\top \Sigma U)^{-1}$ be the corresponding lifting matrix. Write the lifted prediction as $\hat{c} = \mathrm{lift}_U(g) = c_0 + \mathcal{L}_U g$ with $g \in \mathbb{R}^t$. Define $\phi(\hat{c}) = \ell_{\mathrm{SPO}+}(\hat{c}, c)$ (convex in $\hat{c}$ for fixed c) and $\psi(g) = \phi(c_0 + \mathcal{L}_U g)$. For any $v \in \partial\phi(c_0 + \mathcal{L}_U g)$, the standard convex chain rule for composition with an affine map gives $\mathcal{L}_U^\top v \in \partial\psi(g)$. If $g = g_\theta(\xi)$ is differentiable in θ, then for any $u \in \partial\psi(g_\theta(\xi))$ we have $\nabla_\theta \psi(g_\theta(\xi)) = (\nabla_\theta g_\theta(\xi))^\top u$. Combining these two facts, we obtain

$$\partial_\theta \ell_{\mathrm{SPO}+}(c_0 + \mathcal{L}_U g_\theta(\xi), c) \;\ni\; (\nabla_\theta g_\theta(\xi))^\top \mathcal{L}_U^\top v, \quad v := 2\big(x^\star(c) - x^\star\big(2(c_0 + \mathcal{L}_U g_\theta(\xi)) - c\big)\big). \tag{23}$$

As a concrete example, if $g_\theta(\xi) = B_\theta \xi$ with parameter $B_\theta \in \mathbb{R}^{t\times p}$, then (23) implies that a valid stochastic subgradient for B_θ at sample (ξ, c) is $(\mathcal{L}_U^\top v)\, \xi^\top$.

E.4. Algorithm 3: Compressed SPO+ training in Stage II

Algorithm 3 Compressed SPO+ training (Stage II)

Input: Stage II sample $S = \{(\xi_i, c_i)\}_{i=1}^n$, compressed dataset $\hat{\mathcal{D}}$, basis $U \in \mathbb{R}^{d\times t}$ for $\hat{W} = \operatorname{span}(\hat{\mathcal{D}})$, stepsizes $\{\eta_k\}_{k\geq 0}$

Compute $\mathcal{L}_U \leftarrow \Sigma U(U^\top \Sigma U)^{-1}$ and define $\operatorname{lift}_U(s) = c_0 + \mathcal{L}_U s$
for $k = 0, 1, 2, \ldots$ **do**
  Sample (ξ_k, c_k) uniformly from S
  Predict a centered coordinate in $\mathbb{R}^t$: $\hat{s}_k \leftarrow g_\theta(\xi_k)$
  Lift to $\mathbb{R}^d$: $\hat{c}_k \leftarrow \operatorname{lift}_U(\hat{s}_k) = c_0 + \mathcal{L}_U \hat{s}_k$
  $x^0 \leftarrow x^\star(c_k)$
  $x^1 \leftarrow x^\star(2\hat{c}_k - c_k)$
  $v_k \leftarrow 2(x^0 - x^1)$
  $\theta \leftarrow \theta - \eta_k(\nabla_\theta g_\theta(\xi_k))^\top \mathcal{L}_U^\top v_k$
end for

E.5. SPO generalization in the decision-sufficient subspace (Proof of Theorem 25)

We now prove the Stage II generalization bound in Theorem 25. The proof has three ingredients: (i) we bound the complexity of the induced decision class $x^\star \circ \mathcal{H}_{U_\star,d^\star}$ via its Natarajan dimension, (ii) we plug this bound into the Natarajan-dimension generalization theorem for the SPO loss due to El Balghiti et al. (2023), and (iii) we show that the decision-relevant subspace $W_\star = \operatorname{dir}(\mathcal{X}^\star(\mathcal{C}))$ is lossless for decisions under the canonical lift, so restricting to the compressed class incurs no approximation error.

Step 1: Natarajan dimension of the compressed decision class. We begin by translating compressed affine predictors into a multiclass linear prediction problem over labels $\mathcal{X}^\angle$. This is the only place where the intrinsic dimension $d^\star$ enters the analysis.

Lemma 38 (Natarajan dimension bound) *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a bounded polytope with extreme points $\mathcal{X}^\angle$. Let $\mathcal{H}^{\mathrm{aff}}_{U_\star,d^\star}$ be the class of compressed affine predictors*

$$\mathcal{H}^{\mathrm{aff}}_{U_\star,d^\star} := \Big\{ f_{B,b}(\xi) := bc_0 + \mathcal{L}_{U_\star} B\xi \;:\; B \in \mathbb{R}^{d^\star \times p},\ b \in \mathbb{R} \Big\},$$

and let $\mathcal{F}^{\mathrm{aff}}_{U_\star,d^\star} := x^\star \circ \mathcal{H}^{\mathrm{aff}}_{U_\star,d^\star}$ be the induced decision class. Then the Natarajan dimension satisfies $d_N(\mathcal{F}^{\mathrm{aff}}_{U_\star,d^\star}) \leq d^\star p + 1$.

Proof For any $x \in \mathcal{X}^\angle$,

$$f_{B,b}(\xi)^\top x = bc_0^\top x + \big(\mathcal{L}_{U_\star} B\xi\big)^\top x = bc_0^\top x + \Big\langle (B),\ \Big((\mathcal{L}_{U_\star}^\top x)\xi^\top\Big) \Big\rangle .$$

Define the feature map

$$\Psi^{\mathrm{aff}}(\xi, x) := \begin{bmatrix} \big((\mathcal{L}_{U_\star}^\top x)\xi^\top\big) \\ c_0^\top x \end{bmatrix} \in \mathbb{R}^{d^\star p+1}, \qquad w_{B,b} := \begin{bmatrix} (B) \\ b \end{bmatrix} \in \mathbb{R}^{d^\star p+1}.$$

Then $\mathcal{F}^{\mathrm{aff}}_{U_\star,d^\star}$ is a subset of the multiclass linear hypothesis class

$$\mathcal{H}_{\Psi^{\mathrm{aff}}} := \{\xi \mapsto \operatorname*{arg\,min}_{x\in\mathcal{X}^\angle} \langle w, \Psi^{\mathrm{aff}}(\xi, x)\rangle :\ w \in \mathbb{R}^{d^\star p+1}\},$$

because each $f_{B,b}$ induces the decision rule $x^\star(f_{B,b}(\xi)) = \arg\min_{x\in\mathcal{X}^\angle}\langle w_{B,b}, \Psi^{\mathrm{aff}}(\xi, x)\rangle$. Finally, by Shalev-Shwartz and Ben-David (2014, Theorem 29.7), the Natarajan dimension of $\mathcal{H}_{\Psi^{\mathrm{aff}}}$ is at most $d^\star p + 1$, and the same bound holds for its subset $\mathcal{F}^{\mathrm{aff}}_{U_\star,d^\star}$. ■

Step 2: Uniform SPO generalization in the compressed class. We now combine Lemma 38 with the SPO generalization bound of El Balghiti et al. (2023). This yields a uniform convergence guarantee over $\mathcal{H}_{U_\star,d^\star}$ with leading complexity term scaling as $d^\star p$.

Lemma 39 *For any $\delta \in (0,1)$, with probability at least $1-\delta$ over an i.i.d. sample $S = \{(\xi_i, c_i)\}_{i=1}^n$, we have*

$$R_{\mathrm{SPO}}(f) \le \hat{R}_{\mathrm{SPO}}(f) + 2\,\omega_{\mathcal{X}}(\mathcal{C})\sqrt{\frac{2(d^\star p+1)\log(n|\mathcal{X}^\angle|^2)}{n}} + \omega_{\mathcal{X}}(\mathcal{C})\sqrt{\frac{\log(1/\delta)}{2n}},$$

simultaneously for all $f \in \mathcal{H}_{U_\star,d^\star}$, where $\omega_{\mathcal{X}}(\mathcal{C}) := \sup_{c\in\mathcal{C}}\left(\max_{x\in\mathcal{X}} c^\top x - \min_{x\in\mathcal{X}} c^\top x\right)$.

Proof This follows from the Natarajan-dimension generalization bound for SPO loss in El Balghiti et al. (2023). The SPO loss is uniformly bounded by $\omega_{\mathcal{X}}(\mathcal{C})$ when $c \in \mathcal{C}$. Using Lemma 38 and the inclusion $\mathcal{H}_{U_\star,d^\star} \subseteq \mathcal{H}^{\mathrm{aff}}_{U_\star,d^\star}$ (take $b = 1$), we have $d_N(x^\star \circ \mathcal{H}_{U_\star,d^\star}) \le d_N(\mathcal{F}^{\mathrm{aff}}_{U_\star,d^\star}) \le d^\star p + 1$, which yields the stated bound. ■

Step 3: The decision-relevant subspace implies lossless compression. To prove the first statement in Theorem 25, we show that if $U_\star$ spans $W_\star = \mathrm{dir}(\mathcal{X}^\star(\mathcal{C}))$, then compressing any cost vector to $W_\star$ and lifting it back to $\mathcal{C}$ leaves the oracle decision unchanged.

Lemma 40 (Lossless compression by $W_\star$) *Assume $W_\star := \mathrm{dir}(\mathcal{X}^\star(\mathcal{C}))$ and let $U_\star \in \mathbb{R}^{d\times d^\star}$ be an orthonormal basis of $W_\star$. Recall that $\mathcal{L}_{U_\star} := \Sigma U_\star (U_\star^\top \Sigma U_\star)^{-1}$ and $\mathrm{lift}_{U_\star}(s) := c_0 + \mathcal{L}_{U_\star} s$.*

Fix a deterministic tie-breaking rule so that $x^\star(\cdot)$ is single-valued. Then for any $\hat{c} \in \mathcal{C}$, letting

$$\tilde{c} := \mathrm{lift}_{U_\star}\big(U_\star^\top(\hat{c} - c_0)\big),$$

we have $x^\star(\hat{c}) = x^\star(\tilde{c})$.

Consequently, for any predictor $f : \Xi \to \mathcal{C}$, the compressed predictor $\hat{f}(\xi) := \mathrm{lift}_{U_\star}\big(U_\star^\top(f(\xi) - c_0)\big)$ induces the same decisions and satisfies $\ell_{\mathrm{SPO}}(f(\xi), c) = \ell_{\mathrm{SPO}}(\hat{f}(\xi), c)$ for all $c \in \mathcal{C}$.

Proof Fix $\hat{c} \in \mathcal{C}$ and set $s := U_\star^\top(\hat{c} - c_0)$ and $\tilde{c} := \mathrm{lift}_{U_\star}(s) = c_0 + \mathcal{L}_{U_\star} s$. By the definition of $\mathcal{L}_{U_\star}$,

$$U_\star^\top(\tilde{c} - c_0) = U_\star^\top \mathcal{L}_{U_\star} s = U_\star^\top \Sigma U_\star (U_\star^\top \Sigma U_\star)^{-1} s = s = U_\star^\top(\hat{c} - c_0). \qquad (\star)$$

Moreover, since $\hat{c} \in \mathcal{C}$, Lemma 37 implies $\tilde{c} \in \mathcal{C}$.

We claim that $\mathcal{X}^\star(\hat{c}) = \mathcal{X}^\star(\tilde{c})$. Take any $x \in \mathcal{X}^\star(\hat{c})$ and $y \in \mathcal{X}^\star(\tilde{c})$. Since optimal faces are convex hulls of optimal extreme points and $W_\star = \mathrm{dir}(\mathcal{X}^\star(\mathcal{C}))$ is a linear subspace, we have $x - y \in W_\star$. By $(\star)$, $\hat{c} - \tilde{c}$ is orthogonal to $W_\star$, hence

$$(\hat{c} - \tilde{c})^\top (x - y) = 0.$$

On the other hand, optimality gives

$$\hat{c}^\top x \le \hat{c}^\top y, \qquad \tilde{c}^\top y \le \tilde{c}^\top x.$$

The orthogonality identity is equivalent to

$$\hat{c}^\top y - \hat{c}^\top x = \tilde{c}^\top y - \tilde{c}^\top x.$$

The left-hand side is nonnegative, while the right-hand side is nonpositive, so both are zero. Thus $y \in \mathcal{X}^\star(\hat{c})$ and $x \in \mathcal{X}^\star(\tilde{c})$. Since x and y were arbitrary, $\mathcal{X}^\star(\hat{c}) = \mathcal{X}^\star(\tilde{c})$. The deterministic tie-breaking rule then gives $x^\star(\hat{c}) = x^\star(\tilde{c})$.

The predictor claim follows by applying the above argument pointwise to $\hat{c} = f(\xi) \in \mathcal{C}$. ■

With Lemmas 39 and 40 in hand, we can now complete the proof of Theorem 25.

Proof (of Theorem 25)

Part (1) follows from Lemma 40 by taking $f = f_\star \in \mathcal{H}(\mathcal{C})$: the compressed predictor $\hat{f}_\star(\xi) = \mathrm{lift}_{U_\star}(U_\star^\top(f_\star(\xi) - c_0))$ induces the same decisions as $f_\star$, hence has the same population SPO risk, and is a risk minimizer in $\mathcal{H}_{U_\star,d^\star}(\mathcal{C})$.

For part (2), apply Lemma 39, which gives a uniform generalization bound over the larger class $\mathcal{H}_{U_\star,d^\star}$ (and therefore also over its subset $\mathcal{H}_{U_\star,d^\star}(\mathcal{C})$).. ■

E.6. Learning decision-sufficient representation from contextual samples

In the main text, for the clean Stage-II theorem, we assumed access to the decision-relevant subspace $W_\star = \mathrm{dir}(\mathcal{X}^\star(\mathcal{C}))$. Here, we explain how to obtain $\hat{W}$ from labeled contextual samples $\{(\xi_i, c_i)\}_{i=1}^N$ via conditional mean regression, using Algorithm 2. This yields an explicit additional misspecification term under a margin condition.

Recall that for contextual linear optimization with SPO loss, the Bayes-optimal decision rule is $x^\star(\mu(\xi))$, where $\mu(\xi) := \mathbb{E}[c \mid \xi]$. Since $\mathcal{C}$ is convex and $c \in \mathcal{C}$ a.s., we have $\mu(\xi) \in \mathcal{C}$ for all ξ. If we could draw i.i.d. samples from the (unobserved) distribution of $\mu(\xi)$, then we could run Algorithm 2 on those samples and, by our certificate (Theorem 23), obtain a dataset that is pointwise sufficient *with high probability* under the distribution of $\mu(\xi)$. In practice, we only observe noisy costs c, so we proceed in two steps:

1. Estimate μ from contextual samples via a $\mathcal{C}$-valued regression model $\hat{\mu}$, and
2. Treat the predictions $\hat{\mu}(\xi)$ on fresh contexts as pseudo-cost samples and run Algorithm 2.

At a high level, we use contextual samples to approximately simulate the i.i.d. samples we would like to have from the distribution of $\mu(\xi)$, so that we can obtain a high-probability, pointwise decision-sufficient dataset under that distribution. The process is presented in Algorithm 4.

A centered linear conditional-mean model. A convenient choice for $\hat{\mu}$ is multi-response ordinary least squares (OLS). Because the prior set is the ellipsoid $\mathcal{C} = \{c : (c-c_0)^\top \Sigma^{-1}(c-c_0) \le 1\}$, it is natural to write the conditional-mean model in centered form around c_0:

$$c - c_0 \;=\; A_\mu \xi + \epsilon, \qquad \mathbb{E}[\epsilon \mid \xi] = 0, \qquad \mu(\xi) = \mathbb{E}[c \mid \xi] = c_0 + A_\mu \xi. \tag{24}$$

Equivalently, we regress the centered response $y := c - c_0$ onto ξ, and then set $\hat{\mu}(\xi) := c_0 + \hat{A}_\mu \xi$.

Algorithm 4 Learning decision-sufficient representation from contextual samples (Stage I)

Input: Regression sample $\{(\xi_i, c_i)\}_{i=1}^{n_\mu}$, discovery contexts $\{\xi_j^{\mathrm{disc}}\}_{j=1}^{n_{\mathrm{disc}}}$.

1: Fit a regression model $\hat{\mu}$ for $\mu(\xi) := \mathbb{E}[c \mid \xi]$ (the "centered linear" model below)

2: Form pseudo-costs $\hat{c}_j \leftarrow \hat{\mu}(\xi_j^{\mathrm{disc}})$ for $j = 1, \ldots, n_{\mathrm{disc}}$

3: Run Algorithm 2 on $\{\hat{c}_j\}_{j=1}^{n_{\mathrm{disc}}}$ and return $(\hat{\mathcal{D}}, T)$

E.6.1. A CONCRETE BOUND FOR ORDINARY LEAST SQUARES

Suppose $\|\xi\|_2 \le 1$ almost surely and the population design covariance satisfies $\Sigma_\xi := \mathbb{E}[\xi\xi^\top] \succeq \kappa I_p$ for some $\kappa > 0$. More generally, if $\|\xi\|_2 \le C_\xi$ almost surely, the same proof yields the same bound up to an additional multiplicative factor C_ξ.

Assume the regression noise $\epsilon := c - \mu(\xi)$ is conditionally mean-zero and σ-subgaussian in every direction, i.e., for all $\lambda \in \mathbb{R}$ and all $u \in \mathbb{R}^d$ with $\|u\|_2 = 1$,

$$\mathbb{E}\Big[\exp\big(\lambda\, u^\top \epsilon\big) \mid \xi\Big] \;\le\; \exp(\lambda^2\sigma^2/2).$$

Let $\hat{A}_\mu$ be the (multi-response) OLS estimator based on n_μ i.i.d. samples $\{(\xi_i, c_i)\}_{i=1}^{n_\mu}$ by regressing $y_i := c_i - c_0$ on ξ_i, and define

$$\hat{\mu}(\xi) := c_0 + \hat{A}_\mu \xi, \qquad \varepsilon_\mu^2 := \mathbb{E}_\xi\big[\|\hat{\mu}(\xi) - \mu(\xi)\|_2^2\big] = \mathbb{E}_\xi\big[\|(\hat{A}_\mu - A_\mu)\xi\|_2^2\big].$$

Lemma 41 *Fix $\delta_\mu \in (0,1)$ and assume*

$$n_\mu \;\ge\; \frac{8}{\kappa}\log\frac{2p}{\delta_\mu}.$$

Then, with probability at least $1 - \delta_\mu$ over the regression sample, the OLS estimator is well-defined and

$$\|\hat{A}_\mu - A_\mu\|_F \;\le\; C_{\mathrm{reg}} \cdot \frac{\sigma}{\sqrt{\kappa}}\sqrt{\frac{d\Big(p + \log\frac{4d}{\delta_\mu}\Big)}{n_\mu}}, \qquad \textit{where one may take } C_{\mathrm{reg}} = 4\sqrt{2}. \tag{25}$$

Consequently,

$$\sup_{\|\xi\|_2 \le 1} \|\hat{\mu}(\xi) - \mu(\xi)\|_2 \;\le\; C_{\mathrm{reg}} \cdot \frac{\sigma}{\sqrt{\kappa}}\sqrt{\frac{d\Big(p + \log\frac{4d}{\delta_\mu}\Big)}{n_\mu}}, \tag{26}$$

and, in particular,

$$\varepsilon_\mu \;\le\; C_{\mathrm{reg}} \cdot \frac{\sigma}{\sqrt{\kappa}}\sqrt{\frac{d\Big(p + \log\frac{4d}{\delta_\mu}\Big)}{n_\mu}}. \tag{27}$$

Proof Write the regression sample in centered form $y_i = c_i - c_0 = A_\mu \xi_i + \epsilon_i$, where $\mathbb{E}[\epsilon_i \mid \xi_i] = 0$ and ϵ_i is σ-subgaussian in every direction. Let

$$\hat{\Sigma} \;:=\; \frac{1}{n_\mu}\sum_{i=1}^{n_\mu} \xi_i \xi_i^\top \quad \text{and} \quad X \in \mathbb{R}^{n_\mu \times p} \text{ be the design matrix with rows } \xi_i^\top.$$

Then $X^\top X = n_\mu \hat{\Sigma}$.

This proof follows a standard non-asymptotic OLS argument: we first lower bound the minimum eigenvalue of the empirical Gram matrix via a matrix Chernoff bound (Tropp, 2012), and then control the self-normalized noise term $\|(X^\top X)^{-1/2} X^\top \epsilon^{(j)}\|_2$ using sub-Gaussian concentration together with an ε-net (sphere covering) argument (Vershynin, 2018).

Step 1: a lower bound on $\lambda_{\min}(\hat{\Sigma})$. Set $Y_i := \xi_i \xi_i^\top$, so each $Y_i \succeq 0$ and $\lambda_{\max}(Y_i) = \|\xi_i\|_2^2 \le 1$ a.s. Moreover,

$$\mathbb{E}[Y_i] \;=\; \mathbb{E}[\xi \xi^\top] \;=\; \Sigma_\xi \;\succeq\; \kappa I_p.$$

Applying the matrix Chernoff bound (Tropp, 2012, Theorem 1.1) with $\delta = 1/2$ yields

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_{i=1}^{n_\mu} Y_i \right) \le \frac{1}{2}\, \lambda_{\min}\left(\sum_{i=1}^{n_\mu} \mathbb{E} Y_i \right) \right) \;\le\; p \left[\frac{e^{-1/2}}{(1/2)^{1/2}} \right]^{n_\mu \kappa} \;\le\; p\, e^{-n_\mu \kappa / 8}.$$

Under the assumed sample size condition, $p\, e^{-n_\mu \kappa/8} \le \delta_\mu/2$; hence with probability at least $1 - \delta_\mu/2$,

$$\lambda_{\min}(X^\top X) = n_\mu \lambda_{\min}(\hat{\Sigma}) \;\ge\; \frac{n_\mu \kappa}{2}.$$

In particular, $X^\top X$ is invertible and OLS is well-defined on this event.

Step 2: row-wise OLS coefficient error. Let $\beta_j^\top$ denote the j-th row of A_μ and $\hat{\beta}_j^\top$ the j-th row of $\hat{A}_\mu$. Then the scalar response model for coordinate j is $y_{i,j} = \beta_j^\top \xi_i + \epsilon_{i,j}$, and the OLS error satisfies

$$\hat{\beta}_j - \beta_j \;=\; (X^\top X)^{-1} X^\top \epsilon^{(j)},$$

where $\epsilon^{(j)} := (\epsilon_{1,j}, \ldots, \epsilon_{n_\mu, j}) \in \mathbb{R}^{n_\mu}$.

Define the normalized noise vector

$$g_j \;:=\; (X^\top X)^{-1/2} X^\top \epsilon^{(j)} \in \mathbb{R}^p.$$

Conditioned on X, for any $u \in \mathbb{R}^p$ with $\|u\|_2 = 1$,

$$u^\top g_j = u^\top (X^\top X)^{-1/2} X^\top \epsilon^{(j)} = a^\top \epsilon^{(j)}, \quad \text{where } a := X (X^\top X)^{-1/2} u \in \mathbb{R}^{n_\mu}.$$

Note $\|a\|_2^2 = u^\top (X^\top X)^{-1/2} X^\top X (X^\top X)^{-1/2} u = \|u\|_2^2 = 1$. Since $\{\epsilon_{i,j}\}_{i=1}^{n_\mu}$ are independent and each is σ-subgaussian, we have for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\Big[\exp\big(\lambda\, u^\top g_j\big) \mid X \Big] = \prod_{i=1}^{n_\mu} \mathbb{E}[\exp(\lambda a_i \epsilon_{i,j}) \mid X] \le \prod_{i=1}^{n_\mu} \exp(\lambda^2 \sigma^2 a_i^2 / 2) = \exp(\lambda^2 \sigma^2 / 2).$$

Thus, conditional on X, $u^\top g_j$ is σ-subgaussian for every unit vector u.

Let $\mathcal{N}$ be a $1/2$-net of the Euclidean unit sphere in $\mathbb{R}^p$ with $|\mathcal{N}| \le 5^p$ (Vershynin, 2018). For any fixed $u \in \mathcal{N}$ and any $t > 0$, subgaussian tails imply $\mathbb{P}\big(|u^\top g_j| \ge \sigma\sqrt{2t} \mid X\big) \le 2e^{-t}$. Taking a union bound over $\mathcal{N}$ and choosing $t := p \log 5 + \log(2/\delta_j)$ gives

$$\mathbb{P}\left(\max_{u \in \mathcal{N}} |u^\top g_j| \ge \sigma\sqrt{2t} \;\Big|\; X \right) \le |\mathcal{N}|\, 2e^{-t} \;\le\; \delta_j.$$

On the complementary event, the standard net argument yields $\|g_j\|_2 \le 2\max_{u\in\mathcal{N}} |u^\top g_j|$, so with conditional probability at least $1-\delta_j$,

$$\|g_j\|_2 \;\le\; 2\sigma\sqrt{2t} = 2\sigma\sqrt{2\Big(p\log 5 + \log(2/\delta_j)\Big)} \le 4\sigma\sqrt{p+\log(2/\delta_j)},$$

where we used $\log 5 \le 2$ and $\log(2/\delta_j) \ge 0$.

Set $\delta_j := \delta_\mu/(2d)$, so $\log(2/\delta_j) = \log(4d/\delta_\mu)$. Then, with probability at least $1-\delta_\mu/2$ over the noise (and conditional on X), the above bound holds simultaneously for all $j = 1,\dots,d$ by a union bound.

Step 3: combine and translate to prediction error. On the intersection of the events from Steps 1 and 2 (which has probability at least $1-\delta_\mu$), for each j,

$$\|\hat\beta_j - \beta_j\|_2 = \|(X^\top X)^{-1/2} g_j\|_2 \le \frac{\|g_j\|_2}{\sqrt{\lambda_{\min}(X^\top X)}} \le \frac{4\sigma\sqrt{p+\log(4d/\delta_\mu)}}{\sqrt{n_\mu\kappa/2}} = \frac{4\sqrt{2}\,\sigma}{\sqrt{\kappa}}\sqrt{\frac{p+\log(4d/\delta_\mu)}{n_\mu}}.$$

Therefore,

$$\|\hat A_\mu - A_\mu\|_F^2 = \sum_{j=1}^{d} \|\hat\beta_j - \beta_j\|_2^2 \le \frac{32\,\sigma^2}{\kappa}\cdot\frac{d\big(p+\log(4d/\delta_\mu)\big)}{n_\mu}.$$

Taking square roots yields (25). Moreover, for every ξ with $\|\xi\|_2 \le 1$,

$$\|\hat\mu(\xi) - \mu(\xi)\|_2 = \|(\hat A_\mu - A_\mu)\xi\|_2 \le \|\hat A_\mu - A_\mu\|_F\,\|\xi\|_2 \le \|\hat A_\mu - A_\mu\|_F,$$

which implies (26). Finally, since $\|\xi\|_2 \le 1$ a.s.,

$$\varepsilon_\mu^2 = \mathbb{E}_\xi\big[\|(\hat A_\mu - A_\mu)\xi\|_2^2\big] \le \|\hat A_\mu - A_\mu\|_F^2,$$

which implies (27). ■

E.6.2. THE COMPRESSED PREDICTOR

Let $\hat W := \operatorname{span}(\hat{\mathcal{D}})$ and let $\hat U \in \mathbb{R}^{d\times t}$ be an orthonormal basis of $\hat W$, where $t := \dim(\hat W)$ is the learned representation dimension. Define the lifted compressed conditional-mean predictor

$$\tilde\mu(\xi) \;:=\; \operatorname{lift}_{\hat U}\Big(\hat U^\top\big(\hat\mu(\xi) - c_0\big)\Big) = c_0 + \mathcal{L}_{\hat U}\hat U^\top\big(\hat\mu(\xi) - c_0\big). \tag{28}$$

Membership in the compressed linear class. Under the centered linear model (24) and the OLS definition $\hat\mu(\xi) = c_0 + \hat A_\mu\xi$, we have $\hat\mu(\xi) - c_0 = \hat A_\mu\xi$, hence

$$\tilde\mu(\xi) = c_0 + \mathcal{L}_{\hat U}\,(\hat U^\top \hat A_\mu)\,\xi. \tag{29}$$

Therefore $\tilde\mu \in \mathcal{H}_{\hat U,t}$ for the choice $B = \hat U^\top\hat A_\mu \in \mathbb{R}^{t\times p}$, where $\mathcal{H}_{\hat U,t} := \{f_B(\xi) = c_0 + \mathcal{L}_{\hat U}B\xi : B\in\mathbb{R}^{t\times p}\}$.

E.6.3. MARGIN CONDITION

The deterministic oracle $x^\star(\cdot)$ is locally constant on the interior of each optimality cone of $\mathcal{X}$, and may change discontinuously when the cost vector lies on the boundary between two cones. To formalize this, define the *cone-boundary set*

$$\mathcal{B}_{\mathcal{X}} \;:=\; \left\{ c \in \mathbb{R}^d :\; \exists\, x \neq x' \in \mathcal{X}^\angle \text{ s.t. } x, x' \in \arg\min_{x \in \mathcal{X}} c^\top x \right\}.$$

Our transfer argument from regression error to decision error relies on two standard assumptions: (i) the conditional mean $\mu(\xi)$ rarely falls close to $\mathcal{B}_{\mathcal{X}}$ (a standard margin condition), and (ii) the Stage-I pseudo-cost predictions stay within the prior set $\mathcal{C}$ and avoid $\mathcal{B}_{\mathcal{X}}$ almost surely, so that optimal extreme points are unique, and oracle tie-breaking plays no role.

Assumption 42 (Margin condition) *There exist constants $C_{\mathrm{marg}} > 0$ and $\alpha > 0$ such that, for all $\eta > 0$,*

$$\mathbb{P}_\xi\big[\mathrm{dist}\big(\mu(\xi), \mathcal{B}_{\mathcal{X}}\big) \le \eta\big] \;\le\; C_{\mathrm{marg}}\, \eta^\alpha.$$

Moreover, the Stage-I regression predictor satisfies

$$\mathbb{P}_\xi[\hat{\mu}(\xi) \in \mathcal{C}] \;=\; 1, \qquad \mathbb{P}_\xi[\tilde{\mu}(\xi) \in \mathcal{B}_{\mathcal{X}}] \;=\; 0,$$

where $\tilde{\mu}$ is defined in (28).

E.6.4. STAGE-I REPRESENTATION ERROR BOUND

We first relate the regression error of $\hat{\mu}$ to the probability that the induced plug-in decision disagrees with the Bayes rule. Fix $\eta > 0$. On the event that $\mu(\xi)$ lies at distance $> \eta$ from the cone boundary $\mathcal{B}_{\mathcal{X}}$, the optimal extreme point is constant throughout the ball $\mathbb{B}(\mu(\xi), \eta)$. Hence, if the regression estimate is η-accurate and the learned dataset $\hat{\mathcal{D}}$ is pointwise sufficient at $\hat{\mu}(\xi)$, then the lifted predictor $\tilde{\mu}(\xi)$ induces the same unique decision as $\mu(\xi)$. The lemma below formalizes this decomposition.

Lemma 43 *Assume $c \in \mathcal{C}$ almost surely, and let $\tilde{\mu}$ be defined in (28). Under Assumption 42, define for $\eta > 0$*

$$\tau_\mu(\eta) \;:=\; \mathbb{P}_\xi\big[\|\hat{\mu}(\xi) - \mu(\xi)\|_2 > \eta\big].$$

Then, for any $\eta > 0$,

$$\mathbb{P}_\xi\big[x^\star(\tilde{\mu}(\xi)) \neq x^\star(\mu(\xi))\big] \;\le\; \mathbb{P}_\xi\big[\hat{\mathcal{D}} \text{ is not pointwise sufficient at } \hat{\mu}(\xi)\big] + \tau_\mu(\eta) + C_{\mathrm{marg}}\, \eta^\alpha. \quad (30)$$

In particular, using $\mathbb{E}[\|\hat{\mu}(\xi) - \mu(\xi)\|_2^2] \le \varepsilon_\mu^2$ and Markov's inequality,

$$\mathbb{P}_\xi\big[x^\star(\tilde{\mu}(\xi)) \neq x^\star(\mu(\xi))\big] \;\le\; \mathbb{P}_\xi\big[\hat{\mathcal{D}} \text{ is not pointwise sufficient at } \hat{\mu}(\xi)\big] + \frac{\varepsilon_\mu^2}{\eta^2} + C_{\mathrm{marg}}\, \eta^\alpha. \quad (31)$$

Optimizing the right-hand side over η yields

$$\mathbb{P}_\xi\big[x^\star(\tilde{\mu}(\xi)) \neq x^\star(\mu(\xi))\big] \;\le\; \mathbb{P}_\xi\big[\hat{\mathcal{D}} \text{ is not pointwise sufficient at } \hat{\mu}(\xi)\big] + C_{\mathrm{marg}}^{\frac{2}{\alpha+2}}\, C_\alpha\, \varepsilon_\mu^{\frac{2\alpha}{\alpha+2}}, \quad (32)$$

where $C_\alpha := (1 + \alpha/2)\,(2/\alpha)^{\alpha/(\alpha+2)}$.

Proof Define the events

$$A(\xi) := \{\hat{\mathcal{D}} \text{ is pointwise sufficient at } \hat{\mu}(\xi)\},$$
$$B(\xi) := \{\|\hat{\mu}(\xi) - \mu(\xi)\|_2 \le \eta\},$$
$$C(\xi) := \{\mathrm{dist}(\mu(\xi), \mathcal{B}_{\mathcal{X}}) > \eta\},$$
$$D(\xi) := \{\tilde{\mu}(\xi) \notin \mathcal{B}_{\mathcal{X}}\}.$$

Note that $\mathbb{P}_\xi[D(\xi)^c] = 0$ by Assumption 42.

Step 1: $\tilde{\mu}(\xi)$ is fiber-equivalent to $\hat{\mu}(\xi)$ under $\hat{\mathcal{D}}$. Since $\hat{\mu}(\xi) \in \mathcal{C}$ a.s. and $\hat{U}$ spans $\hat{\mathcal{D}}$, we have $\hat{U}^\top(\hat{\mu}(\xi) - c_0) \in \hat{U}^\top(\mathcal{C} - c_0)$ and thus Lemma 37 implies $\tilde{\mu}(\xi) \in \mathcal{C}$. Moreover, using the definition of $\tilde{\mu}$ and the lifting operator $\mathrm{lift}_{\hat{U}}(s) = c_0 + \mathcal{L}_{\hat{U}} s$ (see (8)), we have

$$\hat{U}^\top(\tilde{\mu}(\xi) - c_0) = \hat{U}^\top \mathcal{L}_{\hat{U}}\, \hat{U}^\top(\hat{\mu}(\xi) - c_0) = \left(\hat{U}^\top \Sigma \hat{U}\right)\left(\hat{U}^\top \Sigma \hat{U}\right)^{-1} \hat{U}^\top(\hat{\mu}(\xi) - c_0) = \hat{U}^\top(\hat{\mu}(\xi) - c_0).$$

Therefore, for any $q \in \hat{\mathcal{D}} \subseteq \mathrm{span}(\hat{U})$ we can write $q = \hat{U}a$ for some $a \in \mathbb{R}^t$, and thus

$$q^\top \tilde{\mu}(\xi) - q^\top \hat{\mu}(\xi) = a^\top \hat{U}^\top \left(\tilde{\mu}(\xi) - \hat{\mu}(\xi)\right) = a^\top \left(\hat{U}^\top(\tilde{\mu}(\xi) - c_0) - \hat{U}^\top(\hat{\mu}(\xi) - c_0)\right) = 0.$$

In particular, $\tilde{\mu}(\xi)$ and $\hat{\mu}(\xi)$ lie in the same fiber $\mathcal{C}\left(\hat{\mathcal{D}}, s(\hat{\mu}(\xi); \hat{\mathcal{D}})\right)$.

Step 2: On $A(\xi) \cap B(\xi) \cap C(\xi) \cap D(\xi)$, the oracle decisions coincide. On $A(\xi)$, pointwise sufficiency at $\hat{\mu}(\xi)$ means that there exists some decision $x^{\mathrm{ps}}(\xi) \in \mathcal{X}$ such that

$$x^{\mathrm{ps}}(\xi) \in \mathcal{X}^\star(c') \qquad \forall\, c' \in \mathcal{C}\left(\hat{\mathcal{D}}, s(\hat{\mu}(\xi); \hat{\mathcal{D}})\right).$$

Since $\tilde{\mu}(\xi)$ is in the same fiber, we have $x^{\mathrm{ps}}(\xi) \in \mathcal{X}^\star(\tilde{\mu}(\xi))$ and also $x^{\mathrm{ps}}(\xi) \in \mathcal{X}^\star(\hat{\mu}(\xi))$.

Next, on $B(\xi) \cap C(\xi)$ we have $\hat{\mu}(\xi) \in \mathbb{B}(\mu(\xi), \eta)$ while $\mathbb{B}(\mu(\xi), \eta)$ is contained in the interior of a single normal cone. Hence, the optimal extreme point is unique throughout this ball, and in particular $\mathcal{X}^\star(\hat{\mu}(\xi)) = \{x^\star(\hat{\mu}(\xi))\}$ and $x^\star(\hat{\mu}(\xi)) = x^\star(\mu(\xi))$.

Finally, on $D(\xi)$ we have $\tilde{\mu}(\xi) \notin \mathcal{B}_{\mathcal{X}}$, so $\mathcal{X}^\star(\tilde{\mu}(\xi)) = \{x^\star(\tilde{\mu}(\xi))\}$ is also a singleton. Since $x^{\mathrm{ps}}(\xi)$ is optimal for both $\hat{\mu}(\xi)$ and $\tilde{\mu}(\xi)$, uniqueness forces $x^\star(\tilde{\mu}(\xi)) = x^{\mathrm{ps}}(\xi) = x^\star(\hat{\mu}(\xi)) = x^\star(\mu(\xi))$.

Step 3: Conclude by a union bound. Thus, on $A(\xi) \cap B(\xi) \cap C(\xi) \cap D(\xi)$ we have $x^\star(\tilde{\mu}(\xi)) = x^\star(\mu(\xi))$, and therefore

$$\mathbb{P}_\xi\left[x^\star(\tilde{\mu}(\xi)) \neq x^\star(\mu(\xi))\right] \le \mathbb{P}_\xi[A(\xi)^c] + \mathbb{P}_\xi[B(\xi)^c] + \mathbb{P}_\xi[C(\xi)^c] + \mathbb{P}_\xi[D(\xi)^c].$$

The first term is the pointwise-sufficiency failure probability at $\hat{\mu}(\xi)$. The second term is exactly $\tau_\mu(\eta)$. The third term is controlled by Assumption 42, giving $\mathbb{P}[C(\xi)^c] \le C_{\mathrm{marg}}\eta^\alpha$. Finally, $\mathbb{P}[D(\xi)^c] = 0$ by Assumption 42, which yields (30). The Markov-based bound (31) follows from $\tau_\mu(\eta) \le \varepsilon_\mu^2/\eta^2$. Optimizing $C_{\mathrm{marg}}\eta^\alpha + \varepsilon_\mu^2/\eta^2$ over $\eta > 0$ yields (32). ■

Combining the tail-form transfer bound (30), the certificate guarantee of Theorem 23 for Algorithm 2, and the bounded-design OLS control of Lemma 41 yields the following main finite-sample bound on the representation-induced error of Stage I. The first term below captures the probability that the learned dataset $\hat{\mathcal{D}}$ fails to be pointwise sufficient at a fresh pseudo-cost (scaling as $\widetilde{O}(|T|/n_{\mathrm{disc}})$), while the second term converts regression error to decision error through the uniform prediction-radius bound (26) and the cone-boundary margin.

Theorem 44 (Stage-I representation error under bounded-design OLS) *Suppose Stage I runs Algorithm 4 with discovery sample size n_{disc} and returns $(\hat{\mathcal{D}}, T)$, where T is the compression subsequence produced by Algorithm 2. Under Assumption 42, suppose in addition that the bounded-design OLS conditions of Lemma 41 hold. Fix $\delta_\mu, \delta \in (0,1)$ and define*

$$r_{\mu,\delta_\mu} \;:=\; C_{\mathrm{reg}} \cdot \frac{\sigma}{\sqrt{\kappa}} \sqrt{\frac{d\Big(p + \log\frac{4d}{\delta_\mu}\Big)}{n_\mu}}.$$

Then, with probability at least $1-\delta_\mu-\delta$ over the regression sample and the i.i.d. discovery contexts,

$$\mathbb{P}_\xi\big[x^\star(\tilde{\mu}(\xi)) \neq x^\star(\mu(\xi))\big] \;\le\; \frac{4}{n_{\mathrm{disc}}}\left(6|T| + \log(e/\delta)\right) \;+\; C_{\mathrm{marg}}\, r_{\mu,\delta_\mu}^{\alpha}. \tag{33}$$

In particular, since $|T| \le |\hat{\mathcal{D}}| \le d^\star$,

$$\frac{4}{n_{\mathrm{disc}}}\left(6|T| + \log(e/\delta)\right) \;\le\; \frac{4}{n_{\mathrm{disc}}}\left(6d^\star + \log(e/\delta)\right).$$

Moreover, letting $f_\star(\xi) := \mu(\xi)$ be a Bayes-optimal SPO predictor and letting $\hat{f}_\star \in \arg\min_{f \in \mathcal{H}_{\hat{U},t}} R_{\mathrm{SPO}}(f)$ be the best predictor restricted to the learned representation, we have

$$0 \;\le\; R_{\mathrm{SPO}}(\hat{f}_\star) - R_{\mathrm{SPO}}(f_\star) \;\le\; \omega_{\mathcal{X}}(\mathcal{C}) \cdot \left[\frac{4}{n_{\mathrm{disc}}}\left(6d^\star + \log(e/\delta)\right) + C_{\mathrm{marg}}\, r_{\mu,\delta_\mu}^{\alpha}\right]. \tag{34}$$

Proof On the regression event from Lemma 41, equation (26) implies that

$$\|\hat{\mu}(\xi) - \mu(\xi)\|_2 \;\le\; r_{\mu,\delta_\mu} \qquad \text{for every } \xi \text{ with } \|\xi\|_2 \le 1.$$

Since $\|\xi\|_2 \le 1$ almost surely, this gives

$$\tau_\mu\big(r_{\mu,\delta_\mu}\big) = \mathbb{P}_\xi\big[\|\hat{\mu}(\xi) - \mu(\xi)\|_2 > r_{\mu,\delta_\mu}\big] = 0.$$

Applying (30) with $\eta = r_{\mu,\delta_\mu}$ therefore yields

$$\mathbb{P}_\xi\big[x^\star(\tilde{\mu}(\xi)) \neq x^\star(\mu(\xi))\big] \;\le\; \mathbb{P}_\xi\big[\hat{\mathcal{D}} \text{ is not pointwise sufficient at } \hat{\mu}(\xi)\big] + C_{\mathrm{marg}}\, r_{\mu,\delta_\mu}^{\alpha}.$$

On the same regression event, Theorem 23 applied to the pseudo-cost sample $\{\hat{c}_j\}_{j=1}^{n_{\mathrm{disc}}}$ gives, with probability at least $1-\delta$ over the discovery contexts,

$$\mathbb{P}_\xi\big[\hat{\mathcal{D}} \text{ is not pointwise sufficient at } \hat{\mu}(\xi)\big] \;\le\; \frac{4}{n_{\mathrm{disc}}}\left(6|T| + \log(e/\delta)\right).$$

Combining the two displays and taking a union bound over the regression and discovery samples proves (33).

For the SPO misspecification bound, note that for any predictor f,

$$R_{\mathrm{SPO}}(f) - R_{\mathrm{SPO}}(f_\star) = \mathbb{E}_\xi\Big[\mu(\xi)^\top x^\star(f(\xi)) - \mu(\xi)^\top x^\star(\mu(\xi))\Big] \;\ge\; 0,$$

and for each ξ the bracketed difference is at most $\omega_{\mathcal{X}}(\mathcal{C})$ and is zero whenever the two decisions coincide. Hence $R_{\mathrm{SPO}}(f) - R_{\mathrm{SPO}}(f_\star) \le \omega_{\mathcal{X}}(\mathcal{C})\, \mathbb{P}_\xi[x^\star(f(\xi)) \neq x^\star(\mu(\xi))]$.

Apply this to the specific candidate $f = \tilde{\mu}$. By (29) we have $\tilde{\mu} \in \mathcal{H}_{\hat{U},t}$, so the optimality of $\hat{f}_\star$ in $\mathcal{H}_{\hat{U},t}$ yields $R_{\mathrm{SPO}}(\hat{f}_\star) \le R_{\mathrm{SPO}}(\tilde{\mu})$. Combining these facts with (33) gives (34). ■

Stage I representation error rate under OLS. In Algorithm 4, we use n_μ contextual samples for regression and n_{disc} contextual samples for discovery. Theorem 44 yields the additional Stage I representation error

$$\widetilde{O}\Big(n_\mu^{-\alpha/2} + n_{\text{disc}}^{-1}\Big).$$

Under a constant-fraction split $n_\mu = \Theta(n_I)$ and $n_{\text{disc}} = \Theta(n_I)$, this becomes

$$\widetilde{O}\Big(n_I^{-1} + n_I^{-\alpha/2}\Big) = \widetilde{O}\Big(n_I^{-\min\{1,\alpha/2\}}\Big).$$

Therefore, if $\alpha > 1$ and Stage I and Stage II use comparable sample sizes (e.g., $n_I = \Theta(n_{\text{train}})$), the additional Stage I representation error is lower-order than the $n_{\text{train}}^{-1/2}$ term in Stage II. Consequently, the overall statistical rate is governed by Stage II, whose dominant term depends on the intrinsic dimension $d^\star$ rather than the ambient dimension d; see Theorem 25. This improves sample efficiency for CLO. Moreover, our decision-sufficient representation learning framework reduces the number of trainable parameters in Stage II from dp to $d^\star p$ as an additional advantage.

Remark on the generic L_2-to-decision transfer. If one uses only the mean-squared regression error ε_μ and applies the Markov-based inequality (32), then the same argument gives the weaker Stage I representation rate

$$\widetilde{O}\Big(n_\mu^{-\alpha/(\alpha+2)} + n_{\text{disc}}^{-1}\Big).$$

Under the same constant-fraction split, this becomes $\widetilde{O}(n_I^{-\alpha/(\alpha+2)})$, so the corresponding sufficient condition for Stage I to be lower-order than the $n_{\text{train}}^{-1/2}$ Stage II term is $\alpha > 2$.

E.7. Remark: does vanilla SPO+ implicitly adapt to $d^\star$ without explicit compression?

A natural question is whether vanilla full-dimensional SPO+ training (e.g., over linear predictors $f_B(\xi) = B\xi \in \mathbb{R}^d$) might implicitly ignore directions in $W_\star^\perp$ and therefore enjoy an intrinsic $d^\star p$ dependence. In general, existing uniform-convergence analyses do not imply such a guarantee: the hypothesis class can output cost vectors that induce any extreme point in $\mathcal{X}^\angle$, so the worst-case Natarajan-dimension bound remains dp (El Balghiti et al., 2023). Moreover, $W_\star$ is defined from the decision structure over the prior set $\mathcal{C}$. Unconstrained predictions may leave $\mathcal{C}$, and outside $\mathcal{C}$ the optimizer can depend on directions beyond $W_\star$. Explicit compression makes the intended dimension reduction robust and transparent, rather than relying on optimization dynamics or problem-specific invariances.

E.8. Numerical experiment

We provide a small synthetic shortest-path CLO experiment to illustrate the sample-efficiency gains suggested by the intrinsic $d^\star$ dependence in Theorem 25. We consider a monotone shortest-path instance on a 5×5 grid ($g = 5$), so the cost dimension is $d = 2g(g-1) = 40$. The feasible polytope has $|\mathcal{X}^\angle| = \binom{2(g-1)}{g-1} = 70$ extreme points, each corresponding to a monotone path. Contexts are drawn i.i.d. as $\xi \sim \mathcal{N}(0, I_p)$ with $p = 5$. We take $\mathcal{C} = \{c \in \mathbb{R}^d : \|c - c_0\|_2 \le 1\}$, where c_0 assigns cost 10 on a fixed low-cost corridor and 100 elsewhere. This forces all shortest paths to remain within the corridor, and by enumeration on the 5×5 grid, the resulting intrinsic dimension is $d^\star = 7$.

We compare (i) **full-d SPO+**: a linear predictor $\hat{c}(\xi) \in \mathbb{R}^d$ trained by SGD on the SPO+ surrogate, and (ii) **ours (learn $\hat{W}$ then SPO+)**: first learn a subspace $\hat{W}$ online from observed contexts and costs, and then train a reduced predictor in the learned subspace. We use $n_{\text{train}} = 300$ labeled context–cost pairs for Stage II and an independent test set of size $n_{\text{test}} = 2000$, repeating over 10 random trials and reporting mean $\pm$ 90% confidence intervals.

Stage I: Figure 1 reports the learned dimension $t = \dim(\hat{W})$. Stage II: Figure 2 plots the test SPO risk versus the number of labeled samples used to train the predictor. Consistent with our theory, restricting training to the learned subspace improves performance at a fixed sample size, and $\hat{W}$ quickly stabilizes near the true intrinsic dimension $d^\star = 7$.

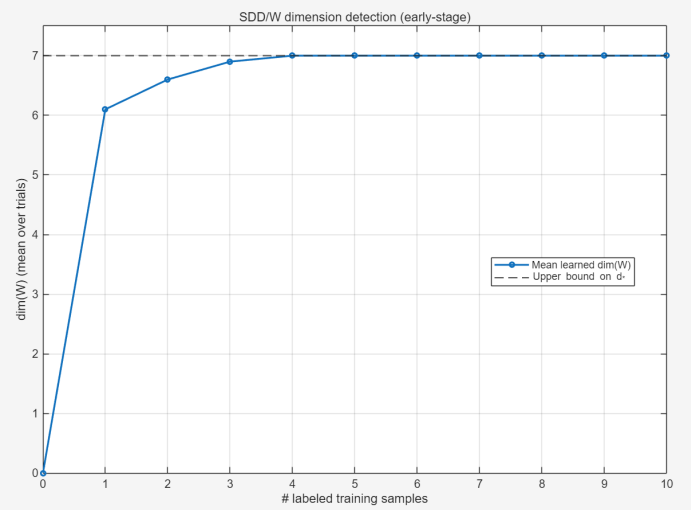

Figure 1: Stage I. Learned dimension $t = \dim(\hat{W})$ (mean over 10 trials).

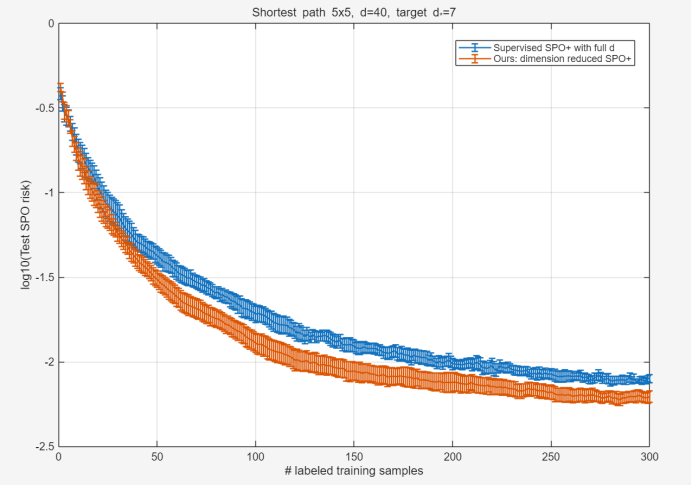

Figure 2: Stage II. SPO risk vs. number of labeled samples (mean $\pm$ 90% CIs over 10 trials).