Which Directions Matter? Sparse Design for Affine Robust Optimization

Pedro Chumpitaz-Flores[*1] **My Duong**[*1] **Juan S. Borrero**[1] **Kaixun Hua**[1]

[1]*University of South Florida*, Tampa, FL, USA

Abstract

Robust machine learning and optimization rely on the uncertainty model choice. We investigate which uncertainty directions a model must cover when defined by a finite dictionary and a budget constraint. Selecting a subset forms an atomic uncertainty set with a closed form support function, yielding tractable robust programs for affine objectives. We propose a data-driven selection rule based on a coverage objective over evaluation directions, including gradients, adversarial perturbations, or shifts observed on held out data. We prove this objective is monotone and submodular, supporting a greedy method with a $(1 - 1/e)$ approximation guarantee and a matching hardness barrier. We also provide a certificate bounding the loss from the selected subset and a radius calibration rule with out-of-sample control.

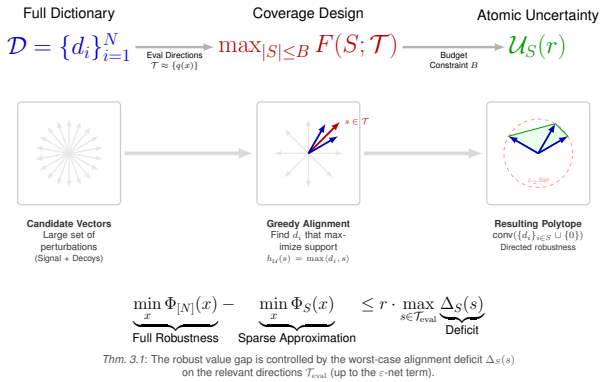

Figure 1: **Sparse directional design via greedy support alignment.** From a dictionary $\mathcal{D}$ of candidate perturbations containing signal and noise (left), the strategy (center) greedily identifies atoms d_i maximizing support alignment with evaluation directions $\mathcal{T}$, such as gradients. This defines the atomic uncertainty set $\mathcal{U}_S(r)$ as a sparse polytope (right). Section 3 establishes selection submodularity and optimality, while the performance certificate (Theorem 3.1) ensures minimizing the alignment deficit Δ_S on $\mathcal{T}$ recovers the robust value.

1 INTRODUCTION

Directional uncertainty often involves large families of perturbation directions, yet only a small subset influences the robust optimum. This paper investigates: Given a finite dictionary $\mathcal{D} = \{d_i\}_{i=1}^N$ and a budget B, which directions recover the robustness of the full model?

We address this through an atomic directional uncertainty family $\mathcal{U}_S(r)$ indexed by subsets $S \subseteq [N]$, a submodular support coverage surrogate for budget-constrained selection, and certificates linking directional coverage to the robust value gap. We further provide finite-sample radius calibration for out-of-sample feasibility.

Our approach builds on robust optimization via uncertainty set design and support function reformulations [Ben-Tal and Nemirovski, 1998, El Ghaoui and Lebret, 1997, Bertsimas and Sim, 2004, Bertsimas et al., 2011], motivated by certified learning under structured perturbations [Szegedy et al., 2013, Goodfellow et al., 2014, Madry et al., 2018, Wong and Kolter, 2018, Raghunathan et al., 2018, Cohen et al., 2019, Engstrom et al., 2017]. Algorithmically, we employ monotone submodular maximization with cardinality constraints [Nemhauser et al., 1978, Krause and Golovin, 2014] and integrate a calibration layer inspired by conformal prediction coverage guarantees [Vovk et al., 2005, Shafer and Vovk, 2008, Lei et al., 2018, Romano et al., 2019]. The dictionary $\mathcal{D}$ comprises application-specific candidates, including gradients, adversarial directions at training points, domain shifts, or stress directions used in robust optimization. Our experiments instantiate $\mathcal{D}$ using synthetic shifts and classifier feature directions to study the micro-budget

[*]Equal contribution.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:1327–1356.

regime $B \ll N$.

Contributions. We present a framework for designing budget-constrained directional uncertainty sets.

- **Conceptually:** We frame directional robustness as a subset design problem over a finite dictionary, where a small direction set recovers the full model performance.
- **Technically:** We introduce an atomic uncertainty family with a closed-form support function, deriving certificates that relate directional coverage to the robust value gap and providing finite-sample radius calibration for out-of-sample feasibility.
- **Empirically:** We demonstrate that greedy selection based on support coverage outperforms random baselines in the micro-budget regime, recovering full model behavior with a minimal fraction of directions.

2 PROBLEM SETUP AND PERFORMANCE GAP

2.1 ATOMIC SETS AND DICTIONARIES

We consider a finite dictionary $\mathcal{D} = \{d_1, \dots, d_N\} \subset \mathbb{R}^p$ augmented with the null atom $d_0 := \mathbf{0}$, and denote the index set by $[N]_0 := \{0, 1, \dots, N\}$. For any subset $S \subseteq [N]$, let $S_0 := S \cup \{0\}$. We define the atomic set dependent on a radius $r > 0$ as

$$U_S(r) \;:=\; \Big\{ u = \sum_{i \in S} \alpha_i d_i :\; \alpha_i \geq 0,\; \sum_{i \in S} \alpha_i \leq r \Big\}, \quad (1)$$

which implies that the support function takes the closed form

$$h_{U_S}(s) \;=\; \sup_{u \in U_S(r)} \langle u,\, s \rangle \;=\; r \max_{i \in S_0} \langle d_i,\, s \rangle \,. \quad (2)$$

Although (1) relies on non-negative coefficients $\alpha_i \geq 0$, symmetric uncertainty sets such as ℓ_1-balls are accommodated by including antipodal pairs $\{d, -d\}$ in the dictionary $\mathcal{D}$.

2.2 AFFINE ROBUST OPTIMIZATION

We address a baseline robust optimization problem defined over a nonempty and compact domain $X \subseteq \mathbb{R}^n$.

Assumption 2.1 (Affine structure in u and regularity)**.** There exist continuous functions $b : X \to \mathbb{R}$ and $q : X \to \mathbb{R}^p$ such that the objective satisfies $f(x, u) = b(x) + \langle u,\, q(x) \rangle$ for all pairs (x, u). Accordingly, the robust objective is

$$\begin{aligned} \Phi_S(x) \;&:=\; \sup_{u \in U_S(r)} f(x, u) \\ &=\; b(x) + h_{U_S}\big(q(x)\big) \\ &=\; b(x) + r \max_{i \in S_0} \langle d_i,\, q(x) \rangle \,. \end{aligned} \quad (3)$$

Under the compactness of X, minimizers of Φ_S are guaranteed to exist.

Remark 2.2 (Optional convexity for tractability). If X is convex, b is convex, and q is affine, then the function $x \mapsto \Phi_S(x)$ is convex. This structure covers standard linear and second order cone programming instances by linearizing the maximum in the objective.

Given an out-of-sample evaluation metric $\mathrm{Risk}(x)$, such as the hold out violation rate or the conditional value at risk of the regret, the subset design problem seeks to minimize

$$\min_{\substack{S \subseteq [N] \\ |S| \leq B}} \mathrm{Risk}\big(x_S^\star\big) \qquad \text{where} \quad x_S^\star \in \operatorname*{Arg\,min}_{x \in X} \; \Phi_S(x). \quad (4)$$

2.3 PERFORMANCE GAP AND CERTIFICATES

To quantify the loss of robustness, we define the full model corresponding to $S = [N]$ with objective $\Phi_{[N]}(x) = b(x) + r \max_{i \in [N]_0} \langle d_i,\, q(x) \rangle$. The following result characterizes the difference between the full and restricted objectives.

Proposition 2.3 (Pointwise value gap)**.** *For any $x \in X$ and $S \subseteq [N]$, the value gap is given by*

$$\begin{aligned} &\Phi_{[N]}(x) - \Phi_S(x) \\ &\qquad = r \Big(\max_{i \in [N]_0} \langle d_i,\, q(x) \rangle - \max_{i \in S_0} \langle d_i,\, q(x) \rangle \Big) \\ &\qquad \geq 0. \end{aligned} \quad (5)$$

In particular, if $x_S^\star$ is a minimizer of Φ_S, then

$$\begin{aligned} &\min_x \Phi_{[N]}(x) - \min_x \Phi_S(x) \\ &\qquad \leq \Phi_{[N]}(x_S^\star) - \Phi_S(x_S^\star) \\ &\qquad = r\, \Delta_S\big(q(x_S^\star)\big), \end{aligned} \quad (6)$$

where the deficit is defined as $\Delta_S(s) := \max_{i \in [N]_0} \langle d_i,\, s \rangle - \max_{i \in S_0} \langle d_i,\, s \rangle$.

Corollary 2.4 (Sufficient condition for equality of values)**.** *If there exists a solution $x_S^\star \in \operatorname{Arg\,min} \Phi_S$ such that the deficit satisfies $\Delta_S\big(q(x_S^\star)\big) = 0$, then the optimal values coincide:*

$$\min_x \Phi_{[N]}(x) \;=\; \min_x \Phi_S(x).$$

We note that this does not imply coincidence of the minimizers except under additional properties such as uniqueness or strict convexity.

Remark 2.5 (Critical directions). We refer to the vectors $s_t := q(x_t)$ as revealed directions, where x_t is a solution of $\min_x \Phi_{S_t}(x)$ obtained during the algorithm described in Section 4.2. Monitoring the value $\Delta_{S_t}(s_t)$ allows us to certify the condition of Corollary 2.4 at run time.

3 THEORETICAL AND STATISTICAL GUARANTEES

3.1 DETERMINISTIC ERROR BOUNDS

Metric definitions. To establish a deterministic global bound on the value gap, we first recall the definition of the alignment deficit $\Delta_S(s) := \max_{i\in[N]_0}\langle d_i, s\rangle - \max_{i\in S_0}\langle d_i, s\rangle \geq 0$, where $S_0 = S \cup \{0\}$ and $[N]_0 = \{0, 1, \ldots, N\}$ with $d_0 = \mathbf{0}$. We introduce the pseudometric induced by the dictionary, defined as

$$d_{\mathcal{D}}(s, s') := \max_{i\in[N]_0} |\langle d_i,\, s - s'\rangle|. \tag{7}$$

Based on this metric, a finite set $\mathcal{T} \subset \mathbb{R}^p$ constitutes an ε-net of $q(X)$ if, for every $x \in X$, there exists an element $s \in \mathcal{T}$ such that $d_{\mathcal{D}}\big(q(x), s\big) \leq \varepsilon$. This geometric structure bridges the gap between discrete directional coverage and global robustness.

Theorem 3.1 (ε-net bridge theorem)**.** *Under Assumption 2.1 and compactness of X, for any subset $S \subseteq [N]$ and any ε-net $\mathcal{T}$ of $q(X)$ in $d_{\mathcal{D}}$, the value gap satisfies*

$$\min_x \Phi_{[N]}(x) - \min_x \Phi_S(x) \;\leq\; r\Big(\max_{s\in\mathcal{T}} \Delta_S(s) + 2\varepsilon\Big).$$

Sketch (Full proof in App. D.3). For each x, the support function $h_U(q(x))$ is r-Lipschitz in $d_{\mathcal{D}}$, meaning $|h_U(s) - h_U(s')| \leq r\, d_{\mathcal{D}}(s, s')$. Taking $s \in \mathcal{T}$ with $d_{\mathcal{D}}(q(x), s) \leq \varepsilon$ and applying the inequality to both the full support and the restricted support yields $\Phi_{[N]}(x) - \Phi_S(x) \leq r\big(\Delta_S(s) + 2\varepsilon\big)$. Minimizing over x and maximizing over $s \in \mathcal{T}$ concludes the proof. □

Remark 3.2 (Constructing $\mathcal{T}$ via Lipschitzness of q). Let $\|\cdot\|$ be a norm on $\mathbb{R}^p$ with dual $\|\cdot\|_*$ and suppose q is L_q-Lipschitz on $(X, \|\cdot\|)$, such that $\|q(x) - q(x')\| \leq L_q\|x - x'\|$. Since $d_{\mathcal{D}}(s, s') \leq \|D\|_*\, \|s - s'\|$ with $\|D\|_* := \max_{i\in[N]_0} \|d_i\|_*$, a grid in X with step

$$\Delta x \;\leq\; \frac{\varepsilon}{\|D\|_*\, L_q}$$

ensures that the image $\mathcal{T} = \{q(x) : x \text{ on the grid}\}$ is an ε-net in $d_{\mathcal{D}}$. In the linear case $q(x) = M^\top x$, we have $L_q = \|M^\top\|$ (operator norm induced by $\|\cdot\|$). An *adaptive* version starts from the revealed directions $s_t = q(x_t)$ and refines partitions until the $d_{\mathcal{D}}$-diameter is at most ε.

In operational contexts, we maintain a set of evaluation directions $\mathcal{T}$ during the greedy procedure to compute the certificate $\widehat{\mathrm{gap}}(S; \mathcal{T}, \varepsilon) := r(\max_{s\in\mathcal{T}} \Delta_S(s) + 2\varepsilon)$. By Theorem 3.1, this quantity deterministically upper-bounds the global value gap. While Theorem 4.5 ensures that the greedy strategy improves the average coverage $F(S; \mathcal{T})$, this certificate controls the worst-case deficit directly. In practice, we evaluate this certificate on sets built from revealed directions $\{q(x_t)\}$ and augmented with hold-out collections, reporting both the worst-case proxy and the average coverage.

3.2 FINITE SAMPLE FEASIBILITY

We complement the deterministic certificates with finite-sample out-of-sample guarantees assuming independent and identically distributed samples $\{u_j\}_{j=1}^m$ from the operational environment. Throughout this analysis, we use the atomic set definition $U_S(r) := \{u = \sum_{i\in S} \alpha_i d_i : \alpha_i \geq 0, \sum_{i\in S} \alpha_i \leq r\}$.

Directional statistic. For each subset $S \subseteq [N]$, define the directional score as

$$Z_S(u) := \max_{i\in S_0} \langle d_i,\, u\rangle.$$

Calibrating the radius r as a high quantile of Z_S allows us to control the coverage level used by the pessimization oracle. Let $\widehat{Q}_\beta(Z_S)$ denote the empirical β-quantile of the scores $\{Z_S(u_j)\}_{j=1}^m$. For a confidence level $1 - \delta$, define

$$\eta := \sqrt{\frac{1}{2m}\, \log\Big(\frac{2\sum_{k=0}^B \binom{N}{k}}{\delta}\Big)}.$$

Assume $\eta \leq \alpha$. We then define the calibrated radius

$$\widehat{r}(S) \;:=\; \widehat{Q}_{\,1-\alpha+\eta}(Z_S). \tag{8}$$

Theorem 3.3 (Finite-sample directional calibration)**.** *Assume $\eta \leq \alpha$. With probability at least $1 - \delta$ over the sampling, the calibration (8) ensures that, simultaneously for every subset $S \subseteq [N]$ with $|S| \leq B$,*

$$\mathbb{P}\big(Z_S(U) > \hat{r}(S)\big) \leq \alpha.$$

The full proof is provided in Appendix D.4.

3.3 RADIUS CALIBRATION PROCEDURE

The integration of this statistical layer into the greedy algorithm follows a sequential process. First, upon completion of the greedy selection which yields S, we compute the scores $Z(u_j) = \max_{i\in S_0}\langle d_i, u_j\rangle$ for all samples $j = 1, \ldots, m$. Second, we calibrate the radius $\widehat{r}$ according to (8). If the subset size is fixed to exactly B, the combinatorial term $\sum_{k=0}^B \binom{N}{k}$ may be replaced by $\binom{N}{B}$. Finally, we solve the robust optimization problem $\min_{x\in X} \Phi_S(x)$ using the uncertainty set $U_S(\widehat{r})$ to obtain the solution $x_S^\star$. This procedure allows us to report both the finite-sample guarantee in Theorem 3.3 and the value gap performance certificate.

Conformal alternative. The statistical correction term scales logarithmically as $O(\sqrt{\log \sum \binom{N}{k}/m})$, which becomes modest when the sample size m is in the low thousands. For moderate sample sizes, a split-conformal calibration approach offers an alternative. By splitting the data into training (for selection) and an independent calibration set, one can define $\widehat{r}$ as the split-conformal quantile of Z, yielding exact marginal coverage for the event $Z_S(U) \le \widehat{r}$ without the logarithmic term, although this sacrifices uniformity over all data-dependent subsets.

3.4 VALIDATION PROTOCOL

The validation protocol requires partitioning the available data into training and hold-out sets. The training set is used to determine the subset S and, if applicable, to pre-calibrate the radius. The hold-out set is reserved to estimate the risk metric $\mathrm{Risk}(x_S^\star)$ for reporting purposes. Beyond the theoretical guarantees provided in Section 3.2, this protocol generates empirical summaries such as means, conditional value at risk, and violation rates. In our experiments, whenever calibration data is available, we apply the post-selection radius calibration step described above before reporting the final performance metrics.

4 GREEDY SELECTION ALGORITHM

4.1 SUPPORT COVERAGE OBJECTIVE

We begin by defining a finite set of evaluation directions $\mathcal{T} \subset \mathbb{R}^p$, which may consist of revealed directions $\{s_t\}$ or a validation set. The support coverage functional is defined as the average maximum alignment over this set:

$$F(S;\mathcal{T}) \;:=\; \frac{1}{|\mathcal{T}|}\sum_{s\in\mathcal{T}} \max_{i\in S_0} \langle d_i,\, s\rangle\,. \tag{9}$$

Proposition 4.1 (Submodularity and monotonicity of F)**.** *For any fixed direction $s \in \mathcal{T}$, the function $g_s(S) := \max_{i\in S_0}\langle d_i, s\rangle$ is monotone and submodular with respect to the set S. Since F is a non-negative linear combination of such functions, $F(\cdot;\mathcal{T})$ retains both monotonicity and submodularity.*

This functional induces the surrogate design problem, denoted as $(\mathsf{P}_{\mathrm{cov}})$, which captures the directional coverage properties of a subset S subject to a budget constraint:

$$\max_{S\subseteq[N]} \Big\{F(S;\mathcal{T}) \;:\; |S| \le B\Big\}. \tag{$\mathsf{P}_{\mathrm{cov}}$}$$

This problem abstracts the combinatorial core of our framework, where a larger value of $F(S;\mathcal{T})$ corresponds to superior average support coverage on the directions relevant to the robust optimization task.

4.2 ORACLE ASSISTED GREEDY METHOD

To address the design problem efficiently, we propose an iterative procedure that alternates between optimizing the robust objective and expanding the set of evaluation directions. At each iteration, the algorithm identifies a worst case perturbation for the current solution using an oracle on the full model. This revealed direction is added to the evaluation set, guiding the subsequent greedy selection of the dictionary atom that maximizes the marginal gain in support coverage.

Algorithm 1 Oracle assisted greedy for designing $U_S(r)$

Input: Dictionary $\mathcal{D} = \{d_i\}_{i=1}^N$, budget B, radius r *(later OOS calibrated to $\widehat{r}$ in Section 3.2)*, optional validation set $\mathcal{V}$.
$S \leftarrow \emptyset$, $\mathcal{T} \leftarrow \emptyset$.
for $t = 1, \dots, B$ **do**
  Solve $x_t \in \mathrm{Arg\,min}_{x\in X}\ \Phi_S(x)$ (convex if applicable).
  Oracle on the full dictionary: obtain $u_t \in U_{[N]}(r)$ that approximately maximizes $f(x_t, u)$ (optional witness)
  With $f(x,u) = b(x) + \langle u, q(x)\rangle$, set $s_t \leftarrow q(x_t)$ (used for greedy scoring), $A_t = \arg\max_{i\in[N]_0}\langle d_i, s_t\rangle$, and update $\mathcal{T} \leftarrow \mathcal{T}\cup\{s_t\}$.
  Choose $i_t \in [N]\setminus S$ that maximizes the marginal gain of $F(S\cup\{i\};\mathcal{T})$.
  $S \leftarrow S\cup\{i_t\}$.
  (Certifiable stopping) If $\Delta_S(s_t) = 0$ and the best marginal gain in F is zero, **stop**.
  (Optional) Early validation: stop if $\mathrm{Risk}(x_t)$ on $\mathcal{V}$ does not improve by more than ε.
end for
OOS calibration: compute $\widehat{r}$ as in Section 3.2 and re-solve $\min_{x\in X}\Phi_S(x)$ with $U_S(\widehat{r})$ to obtain $x_S^\star$.
Output: S, $U_S(\widehat{r})$, and $x_S^\star$.

Remark 4.2 (Worst case greedy baseline). As a baseline that targets worst case coverage on the current direction set $\mathcal{T}$, one can replace the selection rule in Algorithm 1 by

$$i_t \;\in\; \arg\min_{i\in[N]\setminus S}\ \max_{s\in\mathcal{T}} \Delta_{S\cup\{i\}}(s).$$

This rule directly optimizes the certificate term $\max_{s\in\mathcal{T}}\Delta_S(s)$, and we use it as a comparison point in experiments.

4.3 OPTIMALITY GUARANTEES

This section addresses the computational complexity of the design problem and the guarantees of the greedy approach. It first establishes computational hardness.

Theorem 4.3 (NP hardness of directional coverage design)**.** *Given a finite dictionary $\mathcal{D} = \{d_i\}_{i=1}^N \subset \mathbb{R}^p$, a finite set $\mathcal{T} \subset \mathbb{R}^p$, and a budget B, problem ($\mathsf{P}_{\mathrm{cov}}$) is NP hard. This holds even when the coordinates of $\mathcal{T}$ and $\mathcal{D}$ are restricted to $\{0, 1\}$.*

Sketch (Full proof in App. D.6). The proof reduces from the classical *Max Coverage* problem. Let $U = \{1, \dots, m\}$ be the ground set and $\{A_i\}_{i=1}^N$ be subsets of U. An instance of ($\mathsf{P}_{\mathrm{cov}}$) is constructed by setting $p = m$ and identifying $\mathbb{R}^p$ with $\mathbb{R}^m$. Each d_i is the indicator vector of A_i, and $\mathcal{T}$ is the set of canonical basis vectors. In this construction, maximizing $F(S; \mathcal{T})$ is equivalent to maximizing the cardinality of the union of the selected subsets, which is exactly Max Coverage. □

The reduction is gap preserving, which yields a hardness of approximation result.

Theorem 4.4 (Hardness of approximation)**.** *For every $\varepsilon > 0$, unless $P = NP$, there is no polynomial time algorithm that, given an instance of ($\mathsf{P}_{\mathrm{cov}}$), produces a subset S with $|S| \le B$ satisfying*

$$F(S; \mathcal{T}) \;\ge\; (1 - 1/e + \varepsilon)\, F(S^\star; \mathcal{T}),$$

where $S^\star$ is an optimal solution.

Sketch (Full proof in App. D.7). Max Coverage is NP hard to approximate within any factor strictly larger than $1 - 1/e$. Since the reduction preserves objective values up to a constant factor, any algorithm that exceeds this ratio for ($\mathsf{P}_{\mathrm{cov}}$) would contradict the known hardness of Max Coverage. □

Despite these hardness results, the submodularity of F implies that the greedy algorithm attains the optimal approximation factor.

Theorem 4.5 ($(1 - 1/e)$ guarantee for greedy on F)**.** *Under the constraint $|S| \le B$, the greedy algorithm that iteratively adds the index with the largest marginal increase in $F(\cdot; \mathcal{T})$ satisfies*

$$F(S_{\mathrm{greedy}}; \mathcal{T}) \;\ge\; (1 - 1/e)\, F(S^\star; \mathcal{T}),$$

where $S^\star$ maximizes $F(\cdot; \mathcal{T})$ subject to the budget constraint.

Corollary 4.6 (Approximation optimality of the greedy scheme)**.** *Combining Theorem 4.4 with Theorem 4.5, the greedy algorithm is approximation optimal for the directional coverage design problem under the standard assumption $P \neq NP$. No polynomial time algorithm achieves a strictly better worst case approximation factor.*

Remark 4.7 (Selection objective and certification). The selection step optimizes the average coverage surrogate F because it is monotone and submodular, which gives the tight greedy guarantee in Theorem 4.5. The robust value gap, however, is certified through the worst case deficit $\max_{s \in \mathcal{T}} \Delta_S(s)$ via the ε net bridge in Theorem 3.1. This distinction is explicit in the reported metrics: the algorithm optimizes F and reports the certificate term a posteriori.
Appendix E provides a finite probe set bridge between these quantities. In particular, Proposition E.1 bounds the worst case deficit on $\mathcal{T}$ by the surrogate gap $F([N]; \mathcal{T}) - F(S; \mathcal{T})$ plus a probe set occupancy term and a resolution term, and Corollary E.2 combines this bound with Theorem 3.1 to obtain a robust value gap bound in terms of the surrogate gap.

4.4 COMPUTATIONAL COMPLEXITY

The proposed method is tractable for standard convex problems. By the support representation in (2), the objective

$$\Phi_S(x) = b(x) + r \max_{i \in S_0} \langle d_i,\, q(x) \rangle$$

allows the minimization in line 3 of Algorithm 1 to be formulated as a convex problem whose size grows linearly with $|S|$ (for example, by linearizing the maximum in LP or SOCP instances). The pessimization step over $[N]$ reduces to support evaluation or an equivalent dual computation, and the selection step computes marginal gains of F over $\mathcal{T}$, which is inexpensive once the probe set is fixed.

5 COMPUTATIONAL EXPERIMENTS

Uncertainty directions come from a finite dictionary $\mathcal{D} = \{d_i\}_{i=1}^N \subset \mathbb{R}^p$. Given a budget B, each method selects a subset $S \subseteq \{1, \dots, N\}$ with $|S| \le B$, which defines the atomic uncertainty set $\mathcal{U}_S(r)$ (Section 2.1).

A common evaluation pipeline is used across settings. In Frozen Features, each method selects S and evaluates certified robustness at a fixed threat radius ε (Appendix H) without radius calibration. Unless noted, results are averaged over random seeds, and calibrated settings use fixed (α, δ). For each method, the evaluation direction set $\mathcal{T}_{\mathrm{eval}}$ combines directions revealed by the oracle based procedure

(Algorithm 1) with a held out pool when available. The reported metrics are the average coverage $F(S; \mathcal{T}_{\text{eval}})$ and the proxy worst case

$$G(S; \mathcal{T}_{\text{eval}}) := r_{\text{eval}} \max_{s \in \mathcal{T}_{\text{eval}}} \Delta_S(s),$$

where

$$\Delta_S(s) = \max_{i \in [N]_0} \langle d_i, s \rangle - \max_{i \in S_0} \langle d_i, s \rangle,$$

$$S_0 = S \cup \{0\},\ [N]_0 = \{0, 1, \ldots, N\},\ d_0 = 0.$$

The radius is set to $r_{\text{eval}} = \widehat{r}$ in calibrated settings and $r_{\text{eval}} = \varepsilon$ in Frozen Features.

Baselines include Greedy, Random, MaxNorm, and Greedy MaxGap, the minimax rule of Remark 4.2 that targets $\max_{s \in \mathcal{T}} \Delta_S(s)$ on the current direction set. In Section 5.1, we also report TopAct and a Mahalanobis ellipsoid fit. For real-world and dictionary-quality experiments, we also compare two subset/scenario-reduction baselines: CoreSet-kCenter [Sener and Savarese, 2018], a geometric subset-selection baseline adapted from the coreset active-learning method, and the robust-optimization scenario-reduction (RO-SR) baseline [Goerigk and Khosravi, 2023]. Random selects B atoms uniformly without replacement from $\mathcal{D}$, and we report the mean and standard deviation over K independent draws. Selection cost is reported through B, wall clock time, oracle calls, and $\rho = B/N$. Oracle methods use at most B oracle queries by construction. Runtime includes robust solves and selection overhead under fixed solver and hardware settings (Appendix H).

5.1 SYNTHETIC GEOMETRIC VALIDATION

This experiment isolates directional design in two dimensions. An X-shaped distribution is constructed and the dictionary is augmented with vertical decoy atoms that have large norm but low alignment with the evaluation direction. Uncertainty vectors $u \in \mathbb{R}^2$ are sampled near the diagonals, $\mathcal{D}$ is built from signal atoms and decoys, and the full dictionary support is compared with the restricted support induced by subsets S of size B in direction $x_{\text{eval}} = (1, 0)$ at unit radius. Greedy Coverage and Greedy MaxGap are evaluated on a fixed direction pool $\mathcal{T}$, together with Random nested subsets, MaxNorm, and TopAct. A Mahalanobis ellipsoid fit on calibration samples is included as a convex baseline.

Figure 2 reports the support error $z_{\text{full}} - z_S$ in direction x_{eval}. Greedy Coverage and TopAct reduce the error at small budgets by selecting atoms with large projection onto x_{eval}. Random improves more slowly because part of the budget is spent on decoys. MaxNorm performs poorly because decoys have large norms but contribute little support in direction x_{eval}. Greedy MaxGap reduces a proxy worst case criterion on $\mathcal{T}$ and typically requires larger budgets to match average coverage.

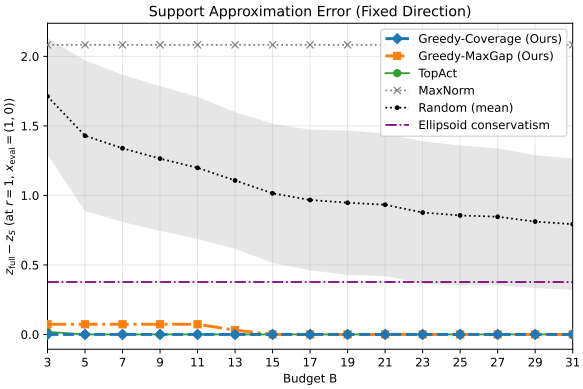

Figure 2: **Support approximation error.** Error versus budget B in direction x_{eval}.

Figure 3 illustrates the mechanism. With small B, the atomic polytope can be anisotropic because it uses a small set of selected atoms to match support in relevant directions. The ellipsoidal baseline summarizes the distribution globally and can be conservative in direction x_{eval} when decoys inflate dispersion.

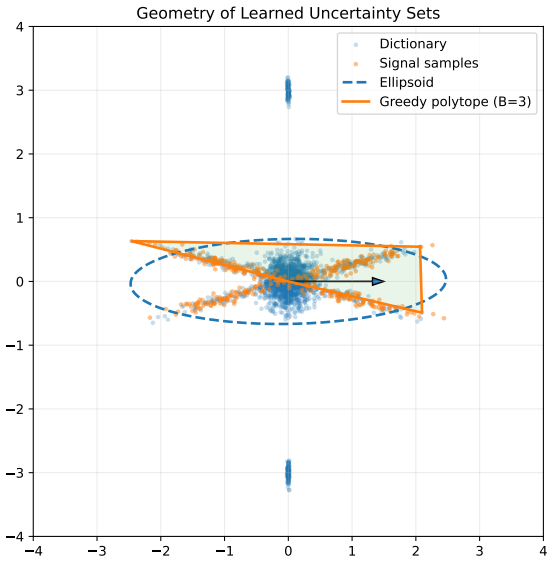

Figure 3: **Geometry of learned uncertainty sets.** Atomic polytope induced by S versus an ellipsoidal baseline.

Figure 4 reports $r_b(B)$. As B increases, the union bound correction grows with the number of candidate subsets, so $r_b(B)$ increases after the empirical quantile stabilizes. Appendix H compares DKW union and split calibration and reports sweeps over (α, δ) with test split violations. DKW union remains conservative, with mean violation gap $\text{viol}_d - \alpha \in [-0.0086, -0.0082]$ across the tested δ, while split calibration is mildly anti conservative, with mean gap in $[0.0037, 0.0048]$; see Tables 7 and 8 and Figure 7.

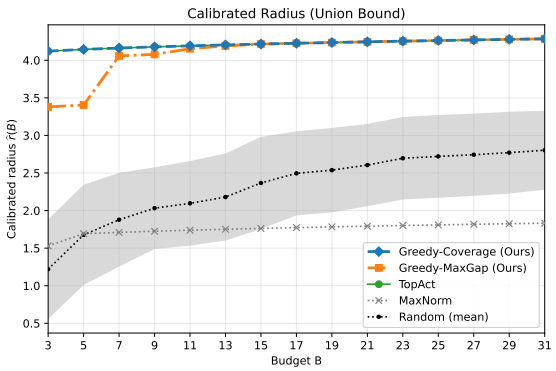

Figure 4: **Calibrated radius.** Calibrated $\widehat{r}(B)$ versus budget B under DKW union calibration.

5.2 REAL-WORLD ROBUST LEARNING UNDER GROUP SHIFT

In this section, we broaden the empirical evaluation to two complementary scenarios that probe different aspects of the method: **(1)** robust machine learning under group shift and **(2)** transaction-cost-aware robust portfolio allocation. We use two datasets, Waterbirds [Sagawa et al., 2019] for (1), which is a standard robust learning benchmark for group shifts, and Fama-French portfolios [Fama and French, 1993], a standard empirical asset-pricing dataset for (2). Here, we focus on the settings and results of (1), more details about experiment (2) are covered in Appendix H.4.

In the setting of group shifting, Greedy-Coverage matches the two added baselines while using only two of the four group directions. Table 1 shows that Greedy-Coverage matches the published subset-selection and scenario-reduction baselines, CoreSet-kCenter and RO-SR, at budget $B = 2$, achieving worst-group accuracy 0.7773, average accuracy 0.9258, and robust loss 0.5046. In contrast, Greedy-MaxGap, MaxNorm, TopAct, and Random perform substantially worse in worst-group accuracy and robust loss, indicating that preserving the support-function structure is more effective than generic score-based or random selection.

Table 1: Waterbirds sparse GroupDRO results. All metrics are evaluated on the full four-group Waterbirds test problem, not only on the selected groups.

Method	Budget	Worst-group acc. ↑	Avg. acc. ↑	Robust loss ↓
Greedy-Coverage	**2**	**0.7773**	**0.9258**	**0.5046**
CoreSet-kCenter	2	0.7773	0.9258	0.5046
RO-SR	2	0.7773	0.9258	0.5046
Greedy-MaxGap	2	0.1931	0.6562	2.5225
MaxNorm	2	0.1931	0.6562	2.5225
TopAct	2	0.1931	0.6562	2.5225
Random	2	0.2591	0.6419	7.6431

5.3 FROZEN-FEATURE SCALING AND DICTIONARY QUALITY

We study directional selection with large uncertainty dictionaries in high ambient dimension. To keep training convex, we train linear robust classifiers on frozen features extracted from a pretrained convolutional network. We consider binary tasks built from CIFAR-10 and CIFAR-100, and report results in two regimes: Hard tasks from CIFAR-10 and Expert tasks from CIFAR-100. The appendix lists all task pairs and additional diagnostics.

For each task and seed, we split the data into disjoint sets for fitting, direction selection, reporting, and final testing. We mine a candidate dictionary $\mathcal{D} = \{d_i\}_{i=1}^{N}$ by projected gradient ascent in feature space under an ℓ_2 radius ε. We first test whether the selection rule remains reliable when dictionary quality changes. Starting from a clean PGD-mined dictionary with 2000 directions, we evaluate three stress regimes: contamination, where irrelevant normalized directions are added; thinning, where clean directions are randomly removed; and weakening, where clean directions are perturbed with increasing noise.

Table 2 shows that Greedy-Coverage remains close to RO-SR and consistently improves over CoreSet-kCenter on support-function metrics. In the contamination setting, Greedy-Coverage and RO-SR preserve high coverage, while CoreSet-kCenter incurs larger support deficits. Unlike other baselines, Greedy-Coverage selects directions by preserving the support-function behavior that defines the affine robust objective. Thus, the certificate is not merely restating the optimized surrogate: it tracks the quality of the selected subset relative to the full dictionary and reveals when the dictionary itself has degraded.

Table 2: CIFAR dictionary-quality stress tests at budget $B = 10$. Coverage is the fraction of full-dictionary support preserved; deficits measure missing support, so higher coverage and lower deficits are better.

Stress test	Method	Coverage ↑	Mean def. ↓	P95 def. ↓	Max def. ↓
Contam.	**Greedy-Coverage**	**0.9899**	**0.0077**	**0.0197**	0.0415
	RO-SR	0.9897	0.0078	0.0204	**0.0363**
	CoreSet-kCenter	0.9718	0.0215	0.0425	0.0577
Thin	**Greedy-Coverage**	**0.9784**	**0.0165**	**0.0301**	0.0430
	RO-SR	0.9779	0.0168	**0.0301**	**0.0402**
	CoreSet-kCenter	0.9654	0.0264	0.0423	0.0549
Weak.	**Greedy-Coverage**	**0.9479**	**0.0397**	**0.0535**	0.0682
	RO-SR	0.9472	0.0403	0.0547	**0.0650**
	CoreSet-kCenter	0.9278	0.0551	0.0753	0.0875

We then evaluate scaling with dictionary size by measuring coverage as a function of the budget density $\rho = B/N$. Figure 5 reports coverage versus ρ for multiple values of N. At matched density, Greedy maintains coverage as N increases, while Random degrades. This gap grows with N, consistent with a haystack effect in which random selection allocates budget to directions that do not contribute to the

coverage objective. To express these trends as a cost, we fix a target coverage level and report the smallest budget B that reaches the target. Figure 6 shows that Greedy reaches coverage ≥ 0.990 with budgets that increase slowly with N, while Random does not reach the target within the tested range $B \leq 50$. Table 3 lists the corresponding budgets and reports an extrapolated estimate for Random, as it does not reach the target for $B \leq B_{\max}$.

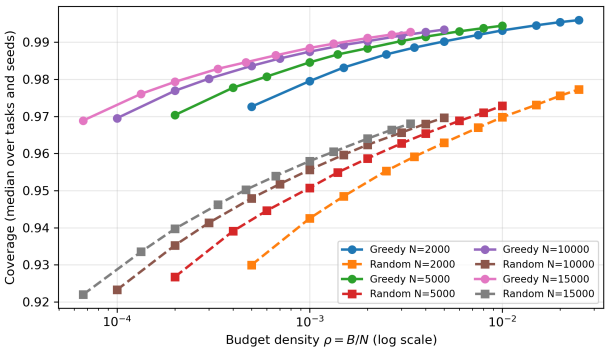

Figure 5: **Coverage versus budget density.** Median coverage over tasks and seeds as a function of $\rho = B/N$, shown for multiple dictionary sizes N.

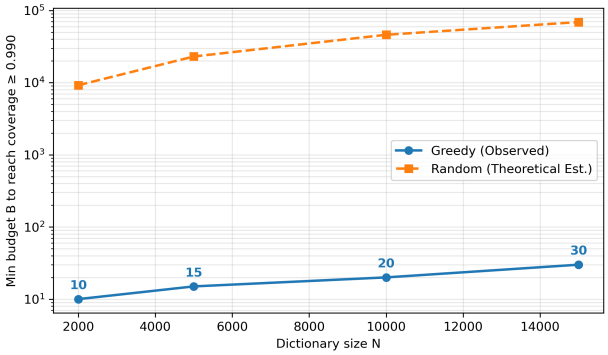

Figure 6: Minimum B needed to reach coverage ≥ 0.990 as a function of dictionary size N. Random does not reach the target for $B \leq 50$.

5.4 BUDGET EFFICIENCY AND RELIABILITY ANALYSIS

The synthetic construction follows Lozano and Borrero [2025] with modifications for uncertainty set selection. The dimensions are $n = q = 30$ and $m = p = 3$, and problem tightness is controlled by $\texttt{rhs} = 0.9$, where lower values correspond to tighter constraints. Each instance is generated from a random seed.

A total of 100 instances are generated, of which 64 instances (64.0%) are robust feasible under the full dictionary baseline, with zero violations. The onset budget and budget-wise efficiency analyses focus on this subset to isolate budget efficiency from baseline infeasibility. Appendix H.6 reports certification counts over all 100 instances.

Table 3: **Minimum budget for target coverage ($\tau = 0.99$).** Greedy reaches the target with small budgets ($B \leq 20$), while Random requires orders of magnitude more (extrapolated estimates *).

Dict. Size (N)	**Target**	$B_{\max}$	**Greedy** B	**Random** B **(Est.)**
2,000	0.99	50	**10**	$9,208^*$
5,000	0.99	50	**15**	$23,023^*$
10,000	0.99	50	**20**	$46,049^*$
15,000	0.99	50	**30**	$69,075^*$

* Extrapolated via coupon collector approximation.

Three deterministic policies (greedy-maxgap, greedy-coverage, and max-gamma) are compared with a random baseline. Efficiency is measured by the onset budget $B^\star$, the minimum budget that achieves zero violations.

Table 4 reports zero-violation rates across budgets, and Table 5 summarizes $B^\star$ over the same 64 instances. Success rates are not monotonic in B because each budget level is solved independently and the selected sets are not nested. Both greedy-maxgap and greedy-coverage reach a median onset of $B^\star = 2$, compared to 7 for max-gamma; within the evaluated budget grid, greedy-maxgap certifies 58 of 64 instances.

The random baseline has no deterministic guarantee at a fixed budget, so reliability is evaluated with 50 independent repetitions per instance. Table 6 reports the mean success rate across the 64 instances feasible under the full dictionary baseline, together with the standard deviation across instances. The dispersion remains large across budgets, including larger values of B.

6 DISCUSSION

How does informed selection compare to uninformed baselines? Across geometric validation (Section 5.1) and scaling benchmarks (Section 5.3), greedy policies reduce support approximation error by aligning selected atoms with the evaluation directions that drive the support function. Random selection spends budget on decoys or irrelevant constraints (Figure 2), weakening coverage and amplifying the haystack effect as the dictionary grows (Figure 5).

How should the ε net term in the value certificate be interpreted? Theorem 3.1 decomposes the robust value gap into a discrete alignment deficit on a finite probe set T and a discretization term 2ε. Practical performance depends on the alignment term $\max_{s \in T} \Delta_S(s)$ and on how well T approximates $q(X)$. Appendix H.3 introduces a proxy $\hat{\varepsilon}(T)$ and a stopping rule based on the empirical certificate

Table 4: Deterministic robustness across budgets: fraction of full-dictionary-feasible seeds with zero violations.

Policy	B=1	2	3	5	7	10	15	20	30
greedy-maxgap	40.6	51.6	56.2	59.4	56.2	54.7	53.1	42.2	39.1
greedy-coverage	40.6	54.7	59.4	62.5	62.5	54.7	45.3	34.4	34.4
max-gamma	28.1	26.6	29.7	37.5	46.9	42.2	35.9	29.7	31.2

Table 5: Onset of certified robustness ($B^\star$) on seeds feasible under the full dictionary baseline ($N = 64$). Robust Rate reports the number of instances with zero violations within the evaluated budget grid. Mean and standard deviation are computed over instances that achieve robustness.

Policy	Robust Rate	Median $B^\star$	Mean $\pm$ Std.
greedy-maxgap	58/64	2	6.3 ± 8.3
greedy-coverage	55/64	2	6.3 ± 9.5
max-gamma	53/64	7	9.7 ± 10.0

Table 6: Reliability of the Random baseline on instances feasible under the full dictionary baseline ($N = 64$). For each budget B, success is the fraction of repetitions with zero violations, and the table reports the mean and standard deviation across instances.

B	3	5	7	10	15
Reps/seed	50	50	50	50	50
Success (%)	22.9	37.2	46.7	60.1	69.8
Std	$\pm$ 22.8	$\pm$ 29.8	$\pm$ 33.1	$\pm$ 33.6	$\pm$ 33.1

$\max_{s \in T} \Delta_S(s) + 2\hat{\varepsilon}(T)$, making the bound usable with finite probe pools. Appendix G shows that this certificate remains informative under inexact selection and inexact oracle evaluations.

Does this efficiency translate to integer robust optimization? In integer programming (Section 5.4), Greedy Max-Gap certifies robustness for most robust feasible instances with smaller budgets than max-gamma (Table 5). This pattern indicates that feasibility is often driven by a small set of active constraints that the greedy objective identifies.

Why are success rates not monotonic? Success rates fluctuate because each budget is solved independently (Table 4). Larger budgets explore new combinations of scenarios rather than extending earlier selections, so rates can decrease at some levels and improve at others. A nested budget schedule would recover monotonicity if required.

Is random selection a viable heuristic? Table 6 shows substantial variance for random selection even at larger budgets, with inconsistent certification across runs. Random selection is a high variance heuristic, while greedy policies provide deterministic guarantees with small sets.

7 CONCLUSIONS

The paper presents a framework for directional uncertainty sets from a finite dictionary. It formulates design as a submodular coverage problem, yielding a greedy method with approximation guarantees and tractability. The framework also provides finite-sample guarantees for radius calibration. Experiments show robust direction selection from large dictionaries, stable coverage under dictionary contamination, thinning, and weakening, competitive sparse GroupDRO performance, and small-budget certification in integer robust optimization.

References

Aharon Ben-Tal and Arkadi Nemirovski. Robust convex optimization. *Mathematics of operations research*, 23(4): 769–805, 1998.

Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Operations research*, 52(1):35–53, 2004.

Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011.

Giuseppe Carlo Calafiore and Marco C Campi. The scenario approach to robust control design. *IEEE Transactions on automatic control*, 51(5):742–753, 2006.

Marco C Campi and Simone Garatti. The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization*, 19(3):1211–1230, 2008.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

Laurent El Ghaoui and Hervé Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on matrix analysis and applications*, 18(4):1035–1064, 1997.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. 2017.

Eugene F Fama and Kenneth R French. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56, 1993.

Uriel Feige. A threshold of ln n for approximating set cover. *Journal of the ACM (JACM)*, 45(4):634–652, 1998.

Marc Goerigk and Mohammad Khosravi. Optimal scenario reduction for one-and two-stage robust optimization with discrete uncertainty in the objective. *European Journal of Operational Research*, 310(2):529–551, 2023.

Marc Goerigk and Jannis Kurtz. Data-driven prediction of relevant scenarios for robust combinatorial optimization. *Computers & Operations Research*, 174:106886, 2025.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017.

Andreas Krause and Daniel Golovin. Submodular function maximization. *Tractability*, 3(71-104):3, 2014.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Leonardo Lozano and Juan S Borrero. On solving integer robust optimization problems with integer decision-dependent uncertainty sets: L. lozano, js borrero. *Mathematical Programming*, pages 1–33, 2025.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Baharan Mirzasoleiman, Amin Karbasi, Ashwinkumar Badanidiyuru, and Andreas Krause. Distributed submodular cover: Succinctly summarizing massive data. *Advances in Neural Information Processing Systems*, 28, 2015.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29, 2016.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations*, 2018.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. *International Conference on Learning Representations*, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *International conference on machine learning*, pages 1954–1963. PMLR, 2015.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International conference on machine learning*, pages 5286–5295. PMLR, 2018.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *Journal of machine learning research*, 10(7), 2009.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002.

Which Directions Matter? Sparse Design for Affine Robust Optimization (Supplementary Material)

Pedro Chumpitaz-Flores[*1] **My Duong**[*1] **Juan S. Borrero**[1] **Kaixun Hua**[1]

[1]*University of South Florida*, Tampa, FL, USA

A RELATED WORK

A.1 SCENARIO-BASED ROBUST OPTIMIZATION

Robust optimization studies worst case decision making under uncertainty and provides tractable reformulations for broad classes of programs through uncertainty sets and their support functions [Ben-Tal and Nemirovski, 1998, El Ghaoui and Lebret, 1997]. When uncertainty is complex, a common alternative is to approximate it with a finite set of representative scenarios. This scenario approach samples constraints and yields finite sample guarantees for chance constrained problems [Calafiore and Campi, 2006, Campi and Garatti, 2008]. Recent work studies scenario reduction for one and two stage robust optimization with discrete uncertainty and derives approximation guarantees that hold for any reduced uncertainty set [Goerigk and Khosravi, 2023]. Related work proposes data-driven rules to identify relevant scenarios that produce strong bounds early in iterative constraint or variable generation methods for robust combinatorial optimization [Goerigk and Kurtz, 2025]. These results provide certificates for probabilistic constraint satisfaction or for approximation quality and convergence speed under scenario reduction. They do not provide selection aware certificates that quantify how a reduced representation deviates from a richer directional robust model, such as computable bounds on the robust value gap induced by selecting a subset of directions.

A.2 ADVERSARIAL ROBUST LEARNING

Robust optimization has influenced robust learning in several forms. Robust support vector machines relate robust losses to regularization [Xu et al., 2009], and distributionally robust learning formalizes robustness to distribution shift [Namkoong and Duchi, 2016, Sinha et al., 2018]. The study of adversarial examples [Szegedy et al., 2013] led to adversarial training, which can be written as a robust optimization problem over norm bounded perturbations [Goodfellow et al., 2014, Madry et al., 2018]. To reduce cost or conservatism, structured perturbation families have been considered, including geometric transformations [Engstrom et al., 2017]. Certified defenses provide worst case guarantees for specific perturbation classes, including convex relaxation methods [Wong and Kolter, 2018], semidefinite bounds [Raghunathan et al., 2018], and randomized smoothing [Cohen et al., 2019]. Most of this literature treats the perturbation family as fixed or learned end to end. In contrast, the present setting selects a small subset from a large, pre specified dictionary of candidate directions and provides an explicit certificate for the approximation loss relative to the full directional model.

A.3 SUBMODULAR SUBSET SELECTION

The selection problem is closely connected to submodular optimization. Many coverage objectives are monotone and submodular, and greedy algorithms achieve approximation guarantees in this setting. The classical result of Nemhauser et al.

[*]Equal contribution.
[*]Equal contribution.

[1978] shows that greedy attains a $(1 - 1/e)$ factor for monotone submodular maximization under a cardinality constraint, and that this factor is tight under standard complexity assumptions. Submodularity has been used for data subset selection and active learning [Wei et al., 2015], as well as for large scale variants based on distributed and streaming computation [Mirzasoleiman et al., 2015]. Surveys also cover sensing and information gathering objectives [Krause and Golovin, 2014]. Related work on coresets selects representative points for learning objectives [Sener and Savarese, 2018]. The central difficulty here is not only to optimize a combinatorial objective, but to identify a coverage functional whose value directly upper bounds robust degradation. The selected objects are uncertainty directions rather than data points, while the resulting optimization follows the same greedy principle once the appropriate functional is defined.

A.4 CALIBRATION AND CONFORMAL GUARANTEES

After selecting directions, the radius of the uncertainty set must be calibrated. Small radii may under protect, whereas large radii may be overly conservative. Calibration of probabilistic outputs is widely studied in prediction systems [Zadrozny and Elkan, 2002, Guo et al., 2017]. Conformal prediction provides distribution free finite sample coverage guarantees by calibrating a score on held out data [Vovk et al., 2005, Shafer and Vovk, 2008], with developments for regression and quantile based intervals [Lei et al., 2018, Romano et al., 2019]. Conformal robust optimization yields finite sample coverage certificates for uncertainty sets, but typically does not address the representation problem of selecting which candidate directions to retain, nor provide a deterministic certificate for how selection changes the robust objective. The radius calibration in this work targets a prescribed out-of-sample coverage level for the robust solution, complementing the deterministic selection aware certificate.

B SCOPE AND LIMITATIONS.

Our framework focuses on affine robust optimization with finite dictionaries. While the design problem is NP-hard, our submodular formulation ensures the greedy strategy provides a strong approximation. The primary limitation is the dependence on the input dictionary; if the candidate pool lacks the true worst-case directions, the resulting set will be optimistic. Therefore, the method is best combined with high-recall mining procedures or domain knowledge. Future work will extend this approach to end-to-end dictionary learning and non-affine uncertainty structures.

C BROADER IMPACT.

This framework supports reliability oriented decision making in settings where structured perturbations matter, such as planning and risk sensitive prediction. It can increase conservatism and reduce performance for some groups if the dictionary reflects biased data or incomplete domain knowledge. The dictionary, budget, and calibration target encode which shifts are treated as relevant. We recommend documenting these choices, reporting selected directions, sensitivity to budget and radius, and evaluating coverage on held out data.

D PROOF OF THEOREMS

D.1 PROOF OF PROPOSITION 2.3

Proof. Recall

$$\Phi_S(x) = b(x) + r \max_{i \in S_0} \langle d_i, q(x) \rangle, \qquad \Phi_{[N]}(x) = b(x) + r \max_{i \in [N]_0} \langle d_i, q(x) \rangle.$$

Therefore, for every $x \in X$,

$$\begin{aligned} \Phi_{[N]}(x) - \Phi_S(x) &= r \left(\max_{i \in [N]_0} \langle d_i, q(x) \rangle - \max_{i \in S_0} \langle d_i, q(x) \rangle \right) \\ &= r\, \Delta_S(q(x)), \end{aligned}$$

which is the claimed identity. □

D.2 PROOF OF COROLLARY 2.4

Proof. Let $x_S^\star \in \arg\min_x \Phi_S(x)$. By Proposition 2.3,

$$\Phi_{[N]}(x_S^\star) - \Phi_S(x_S^\star) = r\,\Delta_S(q(x_S^\star)).$$

If $\Delta_S(q(x_S^\star)) = 0$, then $\Phi_{[N]}(x_S^\star) = \Phi_S(x_S^\star)$.

Also, $\Delta_S(\cdot) \geq 0$, so Proposition 2.3 implies

$$\Phi_{[N]}(x) \geq \Phi_S(x) \qquad \forall x \in X.$$

Hence

$$\min_x \Phi_{[N]}(x) \geq \min_x \Phi_S(x) = \Phi_S(x_S^\star) = \Phi_{[N]}(x_S^\star) \geq \min_x \Phi_{[N]}(x),$$

which proves

$$\min_x \Phi_{[N]}(x) = \min_x \Phi_S(x).$$

□

D.3 PROOF OF THEOREM 3.1

Proof. For any index set $A \subseteq [N]_0$, define

$$M_A(s) := \max_{i \in A} \langle d_i,\, s\rangle\,, \qquad s \in \mathbb{R}^p.$$

Then $\Delta_S(s) = M_{[N]_0}(s) - M_{S_0}(s)$.

We first record a Lipschitz property (with respect to d_D): for any $A \subseteq [N]_0$ and any $s, s' \in \mathbb{R}^p$,

$$|M_A(s) - M_A(s')| \leq \max_{i \in A} |\langle d_i,\, s - s'\rangle| \leq \max_{i \in [N]_0} |\langle d_i,\, s - s'\rangle| = d_D(s, s').$$

Indeed, for any $i \in A$, $\langle d_i,\, s\rangle \leq \langle d_i,\, s'\rangle + |\,\langle d_i,\, s - s'\rangle\,|$, so taking maxima over $i \in A$ gives $M_A(s) \leq M_A(s') + d_D(s, s')$; exchanging s, s' yields the reverse inequality.

Now fix an arbitrary $x \in X$. Since T is an ε-net of $q(X)$ in d_D, there exists $t = t(x) \in T$ such that

$$d_D(q(x), t) \leq \varepsilon.$$

Using the previous Lipschitz bound for both $A = [N]_0$ and $A = S_0$,

$$\begin{aligned}
\Delta_S(q(x)) &= M_{[N]_0}(q(x)) - M_{S_0}(q(x)) \\
&\leq \big(M_{[N]_0}(t) + \varepsilon\big) - \big(M_{S_0}(t) - \varepsilon\big) \\
&= \Delta_S(t) + 2\varepsilon \\
&\leq \max_{s \in T} \Delta_S(s) + 2\varepsilon.
\end{aligned}$$

By Proposition 2.3 (pointwise value gap), for every $x \in X$,

$$\Phi_{[N]}(x) - \Phi_S(x) = r\,\Delta_S(q(x)).$$

Therefore,

$$\Phi_{[N]}(x) - \Phi_S(x) \leq r\left(\max_{s \in T} \Delta_S(s) + 2\varepsilon\right).$$

This bound holds for every $x \in X$. In particular, if $x_S^\star \in \arg\min_x \Phi_S(x)$, then

$$\begin{aligned}
\min_x \Phi_{[N]}(x) - \min_x \Phi_S(x) &\leq \Phi_{[N]}(x_S^\star) - \Phi_S(x_S^\star) \\
&\leq r\left(\max_{s \in T} \Delta_S(s) + 2\varepsilon\right),
\end{aligned}$$

which proves the theorem. □

D.4 PROOF OF THEOREM 3.3

Proof. Let

$$M_B := \sum_{k=0}^{B} \binom{N}{k}.$$

Fix a subset $S \subseteq [N]$ with $|S| \leq B$. Let F_S denote the population CDF of $Z_S(U)$, and let $\widehat{F}_S$ be its empirical CDF based on $\{U_j\}_{j=1}^m$.

By the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality, for each fixed S,

$$\mathbb{P}\left(\sup_{z\in\mathbb{R}} \left|\widehat{F}_S(z) - F_S(z)\right| > \eta\right) \leq 2e^{-2m\eta^2} = \frac{\delta}{M_B}.$$

Define the event

$$\mathbb{E}_S := \left\{\sup_{z\in\mathbb{R}} \left|\widehat{F}_S(z) - F_S(z)\right| \leq \eta\right\}.$$

Then $\mathbb{P}(\mathbb{E}_S) \geq 1 - \delta/M_B$.

Now take a union bound over all subsets $S \subseteq [N]$ with $|S| \leq B$. The event

$$\mathbb{E} := \bigcap_{\substack{S\subseteq[N]\\|S|\leq B}} \mathbb{E}_S$$

satisfies

$$\mathbb{P}(\mathbb{E}) \geq 1 - \sum_{\substack{S\subseteq[N]\\|S|\leq B}} \frac{\delta}{M_B} \geq 1 - \delta.$$

We now prove the claimed guarantee on the event $\mathbb{E}$. Fix any $S \subseteq [N]$ with $|S| \leq B$, and set

$$r := \widehat{r}(S) = \widehat{Q}_{\,1-\alpha+\eta}(Z_S).$$

By definition of the empirical quantile,

$$\widehat{F}_S(r) \geq 1 - \alpha + \eta.$$

Since $\mathbb{E}$ implies $\widehat{F}_S(z) - F_S(z) \leq \eta$ for all z, we obtain

$$F_S(r) \geq \widehat{F}_S(r) - \eta \geq 1 - \alpha.$$

Therefore,

$$\mathbb{P}(Z_S(U) > r) = 1 - F_S(r) \leq \alpha.$$

Because this argument is valid for every S with $|S| \leq B$, the guarantee holds simultaneously for all such subsets on the event $\mathbb{E}$.

Finally, the subset returned by Algorithm 1 is data-dependent, but it belongs to the same finite family $\{S \subseteq [N] : |S| \leq B\}$. Hence the same simultaneous guarantee applies to the algorithm output. □

D.5 PROOF OF PROPOSITION 4.1

Proof. Fix $s \in T$ and define

$$g_s(S) := \max_{i\in S_0} \langle d_i, s\rangle.$$

Monotonicity is immediate: if $A \subseteq B$, then $A_0 \subseteq B_0$, hence

$$g_s(A) = \max_{i\in A_0} \langle d_i, s\rangle \leq \max_{i\in B_0} \langle d_i, s\rangle = g_s(B).$$

To show submodularity, let $A \subseteq B \subseteq [N]$ and $j \in [N] \setminus B$. Write $a_j := \langle d_j, s \rangle$. Then

$$g_s(A \cup \{j\}) - g_s(A) = \max\{g_s(A), a_j\} - g_s(A) = \max\{0, a_j - g_s(A)\},$$

and similarly

$$g_s(B \cup \{j\}) - g_s(B) = \max\{0, a_j - g_s(B)\}.$$

Since $g_s(A) \leq g_s(B)$, we get

$$\max\{0, a_j - g_s(A)\} \geq \max\{0, a_j - g_s(B)\},$$

which is the diminishing-returns property. Thus g_s is submodular.

Finally,

$$F(S; T) = \frac{1}{|T|} \sum_{s \in T} g_s(S)$$

is a nonnegative linear combination of monotone submodular functions, hence is itself monotone and submodular. □

D.6 PROOF OF THEOREM 4.3

Proof. We reduce from the classical *Max-Coverage* problem.

An instance of Max-Coverage consists of:

- a ground set $U = \{1, \ldots, m\}$,
- subsets $A_1, \ldots, A_N \subseteq U$,
- a budget B,

and asks for a set of indices $S \subseteq [N]$ with $|S| \leq B$ maximizing

$$\left| \bigcup_{i \in S} A_i \right|.$$

Given such an instance, construct an instance of the directional coverage problem as follows:

- Set the ambient dimension to $p = m$.
- For each $i \in [N]$, define $d_i \in \{0, 1\}^m$ as the indicator vector of A_i:

$$(d_i)_j := \begin{cases} 1, & j \in A_i, \\ 0, & j \notin A_i. \end{cases}$$

- Let

$$T := \{e_1, \ldots, e_m\} \subseteq \mathbb{R}^m,$$

  where e_j is the j-th canonical basis vector.

This reduction is computable in polynomial time, and all coordinates of d_i and T are in $\{0, 1\}$.

Now fix any $S \subseteq [N]$. For any $s = e_j \in T$,

$$\max_{i \in S_0} \langle d_i, e_j \rangle = \max_{i \in S_0} (d_i)_j.$$

Since each $(d_i)_j \in \{0, 1\}$ and $d_0 = 0$, we have

$$\max_{i \in S_0} (d_i)_j = \begin{cases} 1, & \text{if there exists } i \in S \text{ such that } j \in A_i, \\ 0, & \text{otherwise.} \end{cases}$$

Therefore,

$$\sum_{s\in T} \max_{i\in S_0} \langle d_i,\, s\rangle = \sum_{j=1}^{m} \mathbf{1}\left\{ j \in \bigcup_{i\in S} A_i \right\} = \left| \bigcup_{i\in S} A_i \right|.$$

Dividing by $|T| = m$ gives

$$F(S;T) = \frac{1}{m} \left| \bigcup_{i\in S} A_i \right|.$$

Thus, maximizing $F(S;T)$ subject to $|S| \le B$ is exactly equivalent (up to the constant factor $1/m$) to solving the Max-Coverage instance. Since Max-Coverage is NP-hard, the directional coverage design problem is NP-hard as well. □

D.7 PROOF OF THEOREM 4.4

Proof. We use the same reduction from Max-Coverage as in the proof of Theorem 4.3.

Let

$$\mathrm{cov}(S) := \left| \bigcup_{i\in S} A_i \right|$$

denote the Max-Coverage objective value. For the constructed instance of (P_{cov}), we proved that for every $S \subseteq [N]$,

$$F(S;T) = \frac{1}{m}\,\mathrm{cov}(S).$$

Hence the reduction preserves approximation ratios exactly: for any feasible S and any optimal solution $S^\star$,

$$\frac{F(S;T)}{F(S^\star;T)} = \frac{\mathrm{cov}(S)}{\mathrm{cov}(S^\star)}.$$

Suppose, for contradiction, that there exists a polynomial-time algorithm for (P_{cov}) achieving approximation factor $(1-1/e+\varepsilon)$ for some $\varepsilon > 0$. Applying it to the reduced instance and using the identity above would produce, in polynomial time, a solution S to Max-Coverage such that

$$\mathrm{cov}(S) \ge (1 - 1/e + \varepsilon)\ \mathrm{cov}(S^\star).$$

This contradicts the classical hardness-of-approximation barrier for Max-Coverage: for every $\varepsilon > 0$, no polynomial-time algorithm can achieve an approximation factor of $1 - 1/e + \varepsilon$ unless $P = NP$ (see, e.g., Feige [1998]).

Therefore, no polynomial-time algorithm can approximate (P_{cov}) within a factor strictly better than $1 - 1/e$ unless $P = NP$. □

D.8 PROOF OF THEOREM 4.5

Proof. By Proposition 4.1, the set function $F(\cdot;T)$ is monotone and submodular. Also, since $d_0 = 0$ and $S_0 = \{0\}$ when $S = \emptyset$, we have

$$F(\emptyset;T) = \frac{1}{|T|} \sum_{s\in T} \max_{i\in\{0\}} \langle d_i, s\rangle = 0.$$

Therefore, the classical Nemhauser–Wolsey–Fisher theorem for monotone submodular maximization under a cardinality constraint implies that the greedy algorithm returns a set S_{greedy} with

$$F(S_{\mathrm{greedy}};T) \ge (1 - 1/e) \max_{|S|\le B} F(S;T).$$

□

D.9 PROOF OF COROLLARY 4.6

Proof. Immediate from Theorem 4.4 and Theorem 4.5. □

E AVERAGE-TO-WORST BRIDGE ON THE FINITE PROBE SET

This section formalizes a sufficient condition under which improving the surrogate objective $F(S;T)$ also reduces the worst-case deficit $\max_{s\in T}\Delta_S(s)$ on the same probe set.

Recall the notation from Appendix C.3. For any index set $A \subseteq [N]_0$, define

$$M_A(s) := \max_{i\in A}\langle d_i, s\rangle, \qquad \Delta_S(s) := M_{[N]_0}(s) - M_{S_0}(s).$$

For a fixed nonempty finite set $T \subset \mathbb{R}^p$ and a subset $S \subseteq [N]$, define the average deficit

$$\bar{\Delta}_S(T) := \frac{1}{|T|}\sum_{s\in T}\Delta_S(s).$$

Using the definition of $F(\cdot;T)$, the surrogate gap satisfies the exact identity

$$F([N];T) - F(S;T) = \bar{\Delta}_S(T).$$

To convert this average quantity into a worst-case bound, define the local occupancy of the probe set under d_D:

$$\mu_\eta(T) := \min_{s\in T}\left|\{t \in T : d_D(s,t) \le \eta\}\right|, \qquad \eta \ge 0.$$

Since $T \neq \varnothing$ and $d_D(s,s) = 0$, we have $\mu_\eta(T) \ge 1$ for all $\eta \ge 0$.

Proposition E.1 (Average-to-worst bridge on T)**.** *Let $T \subset \mathbb{R}^p$ be nonempty and finite. Then, for any $S \subseteq [N]$ and any $\eta \ge 0$,*

$$\max_{s\in T}\Delta_S(s) \;\le\; \frac{|T|}{\mu_\eta(T)}\,\bar{\Delta}_S(T) + 2\eta \;=\; \frac{|T|}{\mu_\eta(T)}\big(F([N];T) - F(S;T)\big) + 2\eta.$$

Equivalently,

$$\max_{s\in T}\Delta_S(s) \;\le\; \inf_{\eta\ge 0}\left\{\frac{|T|}{\mu_\eta(T)}\,\bar{\Delta}_S(T) + 2\eta\right\}.$$

Proof. Appendix C.3 shows that, for any index set $A \subseteq [N]_0$, the map $M_A(\cdot)$ is 1-Lipschitz with respect to d_D:

$$|M_A(s) - M_A(s')| \le d_D(s,s').$$

Therefore, since $\Delta_S = M_{[N]_0} - M_{S_0}$,

$$|\Delta_S(s) - \Delta_S(s')| \le |M_{[N]_0}(s) - M_{[N]_0}(s')| + |M_{S_0}(s) - M_{S_0}(s')| \le 2\,d_D(s,s').$$

Hence $\Delta_S(\cdot)$ is 2-Lipschitz in d_D.

Let $s^\star \in \arg\max_{s\in T}\Delta_S(s)$ and define the d_D-ball on the probe set

$$\mathcal{N}_\eta(s^\star) := \{t \in T : d_D(t,s^\star) \le \eta\}.$$

For any $t \in \mathcal{N}_\eta(s^\star)$, the 2-Lipschitz property gives

$$\Delta_S(t) \ge \Delta_S(s^\star) - 2\eta.$$

Summing over $\mathcal{N}_\eta(s^\star)$ and using nonnegativity of Δ_S (since $S_0 \subseteq [N]_0$ implies $M_{[N]_0} \ge M_{S_0}$ pointwise),

$$|T|\,\bar{\Delta}_S(T) = \sum_{t\in T}\Delta_S(t) \ge \sum_{t\in\mathcal{N}_\eta(s^\star)}\Delta_S(t) \ge |\mathcal{N}_\eta(s^\star)|\big(\Delta_S(s^\star) - 2\eta\big).$$

By definition of $\mu_\eta(T)$, $|\mathcal{N}_\eta(s^\star)| \ge \mu_\eta(T)$, so

$$|T|\,\bar{\Delta}_S(T) \ge \mu_\eta(T)\big(\max_{s\in T}\Delta_S(s) - 2\eta\big).$$

Rearranging yields

$$\max_{s\in T}\Delta_S(s) \le \frac{|T|}{\mu_\eta(T)}\,\bar{\Delta}_S(T) + 2\eta.$$

Finally, identify $\bar{\Delta}_S(T)$ with the surrogate gap:

$$\bar{\Delta}_S(T) = \frac{1}{|T|}\sum_{s\in T}\big(M_{[N]_0}(s) - M_{S_0}(s)\big) = F([N];T) - F(S;T).$$

This proves the first display. The infimum form follows immediately by minimizing the right-hand side over $\eta \ge 0$. □

Corollary E.2 (Robust value-gap bound through the surrogate gap)**.** *Under the assumptions of Theorem 3.1, let T be an ε-net of $q(X)$ in d_D. Then, for any $S \subseteq [N]$ and any $\eta \ge 0$,*

$$\min_x \Phi[N](x) - \min_x \Phi_S(x) \;\le\; r\left(\frac{|T|}{\mu_\eta(T)}\big(F([N];T) - F(S;T)\big) + 2\eta + 2\varepsilon\right).$$

Equivalently,

$$\min_x \Phi[N](x) - \min_x \Phi_S(x) \;\le\; r\left(\inf_{\eta\ge 0}\left\{\frac{|T|}{\mu_\eta(T)}\big(F([N];T) - F(S;T)\big) + 2\eta\right\} + 2\varepsilon\right).$$

Proof. Theorem 3.1 gives

$$\min_x \Phi[N](x) - \min_x \Phi_S(x) \le r\big(\max_{s\in T}\Delta_S(s) + 2\varepsilon\big).$$

Apply Proposition E.1 to bound $\max_{s\in T}\Delta_S(s)$ and substitute. □

Practical interpretation. The factor $|T|/\mu_\eta(T)$ measures how densely the probe set covers itself under d_D at resolution η. When T is well distributed in the relevant directions, this factor is moderate, and improving the surrogate objective $F(S;T)$ reduces the worst-case certificate term on T, up to the additive resolution term 2η. Conversely, if T is highly nonuniform (small local occupancy near some directions), then average coverage can be less informative about the worst-case deficit; the bound makes this dependence explicit.

F DIRECTIONAL RELEVANCE IN THE ROBUST LP TEMPLATE

We first recall a minimal linear-programming instantiation of the robust template, and then use LP duality to isolate the directions of uncertainty that are *value-relevant* for the full directional model. This yields a dual-based notion of relevance that will serve as a structural benchmark for any restricted family $U_S(r)$.

Minimal LP template. Let $X = \{x : Ax \le b\}$ and $f(x,u) = c(u)^\top x$ with $c(u) = c_0 + Mu$. Then

$$\Phi_S(x) = c_0^\top x + r\max_{i\in S_0}\big\langle d_i,\, M^\top x\big\rangle,$$

and $\min_x \Phi_S(x)$ is an LP after introducing a variable to linearize the maximum. Pessimization over $[N]$ reduces to evaluating

$$i^\star \in \arg\max_{i\in[N]_0}\big\langle d_i,\, M^\top x\big\rangle.$$

F.1 SETUP AND NOTATION

Assume a compact polyhedral feasible region

$$X := \{x \in \mathbb{R}^n : Ax \le b\},$$

and an affine objective coefficient map

$$c(u) = c_0 + Mu,$$

so that the nominal objective is $f(x,u) = c(u)^\top x$. The full directional robust objective is

$$\Phi_{[N]}(x) = c_0^\top x + r\max_{i\in[N]_0}\big\langle d_i,\, M^\top x\big\rangle.$$

F.2 ROBUST LP TEMPLATE AND DUAL

We adopt the following standing setup.

Assumption F.1 (Robust LP template)**.** Let $X := \{x \in \mathbb{R}^n : Ax \le b\}$ be a nonempty, bounded polyhedron, with $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. The uncertain cost is

$$c(u) := c_0 + Mu, \qquad c_0 \in \mathbb{R}^n,\ M \in \mathbb{R}^{n\times p},$$

and the directional dictionary is $\mathcal{D} = \{d_i\}_{i=1}^N \subset \mathbb{R}^p$, with $d_0 := 0$ and $[N]_0 := \{0, 1, \dots, N\}$. For $S \subseteq [N]$ and $r > 0$,

$$U_S(r) := \Big\{u = \textstyle\sum_{i\in S} \alpha_i d_i :\ \alpha_i \ge 0,\ \sum_{i\in S} \alpha_i \le r\Big\}.$$

The *full* model corresponds to $S = [N]$.

The robust value of the full model is

$$z^\star_{[N]} := \min_{x\in X} \sup_{u\in U_{[N]}(r)} (c_0 + Mu)^\top x.$$

Using the support function of $U_{[N]}(r)$ and introducing an epigraph variable $t \in \mathbb{R}$ we obtain the equivalent LP

$$\begin{aligned} (\mathrm{P}[N]) \qquad z^\star_{[N]} \;=\; \min_{x,t}\ & c_0^\top x + rt \\ \text{s.t.}\ & Ax \le b, \\ & (Md_i)^\top x \le t, \quad \forall i \in [N]_0. \end{aligned} \tag{10}$$

We now write its dual in a form that exposes the role of the dictionary.

Proposition F.2 (Dual of the full robust LP)**.** *Under Assumption F.1, the dual of* (P$[N]$) *is*

$$\begin{aligned} (D[N]) \qquad z^\star_{[N]} \;=\; \max_{\lambda,\mu}\ & -\,b^\top \lambda \\ \textit{s.t.}\ & A^\top \lambda \;=\; -\,c_0 - M\Big(\textstyle\sum_{i\in[N]_0} \mu_i d_i\Big), \\ & \sum_{i\in[N]_0} \mu_i \;=\; r, \\ & \lambda \;\ge\; 0, \quad \mu_i \;\ge\; 0,\ \forall i \in [N]_0. \end{aligned} \tag{11}$$

Strong duality holds: (P$[N]$) *and* (D$[N]$) *have the same finite optimal value* $z^\star_{[N]}$.

Proof. Assign dual variables $\lambda \in \mathbb{R}^m_+$ to $Ax \le b$ and $\mu_i \ge 0$ to $(Md_i)^\top x - t \le 0$ for $i \in [N]_0$. The Lagrangian of (10) is

$$L(x, t; \lambda, \mu) = \big(c_0 + A^\top\lambda + M\textstyle\sum_{i\in[N]_0} \mu_i d_i\big)^\top x + \big(r - \sum_{i\in[N]_0} \mu_i\big)t - b^\top\lambda.$$

For $\inf_{x,t} L(x, t; \lambda, \mu)$ to be finite we need both coefficients of x and t to vanish. This yields the two equalities in (11); under them, $L(x, t; \lambda, \mu) \equiv -b^\top\lambda$. Maximizing over (λ, μ) subject to feasibility gives *(D$[N]$)*. Strong duality follows from standard LP duality under boundedness and feasibility of X. □

F.3 VALUE-RELEVANT DIRECTIONS AND THE RELEVANT FACE

Optimal dual solutions reveal which dictionary atoms are actually used in the worst-case cost at the robust optimum.

Definition F.3 (Value-relevant indices and face)**.** Let $(x^\star, t^\star)$ and $(\lambda^\star, \mu^\star)$ be optimal solutions of *(P$[N]$)* and *(D$[N]$)*, respectively. Define the set of *value-relevant indices*

$$I_{\mathrm{rel}} := \{i \in [N] : \mu^\star_i > 0\},$$

and the corresponding *relevant face*

$$C_{\mathrm{rel}} \;:=\; \mathrm{conv}\big(\{d_i : i \in I_{\mathrm{rel}}\} \cup \{0\}\big) \;\subseteq\; \mathrm{conv}\big(\{d_i : i \in [N]\} \cup \{0\}\big).$$

We also define the associated worst-case direction

$$u^\star \;:=\; \sum_{i \in [N]_0} \mu_i^\star d_i \in U_{[N]}(r), \qquad v^\star \;:=\; \frac{1}{r}\, u^\star \in C_{\mathrm{rel}}.$$

By complementary slackness, every index $i \in I_{\mathrm{rel}}$ corresponds to a binding constraint $(Md_i)^\top x^\star = t^\star$ in *(P[N])*. The vector $u^\star$ is a worst-case perturbation at $x^\star$, and $v^\star$ lies in an exposed face of $\mathrm{conv}(\mathcal{D} \cup \{0\})$ determined by $M^\top x^\star$.

F.4 DIRECTIONAL SUFFICIENCY VIA THE DUAL

For $S \subseteq [N]$, define the restricted uncertainty set $U_S(r)$ as in Assumption F.1, and denote by $z_S^\star$ the corresponding robust value:

$$z_S^\star \;:=\; \min_{x \in X} \; \sup_{u \in U_S(r)} \; (c_0 + Mu)^\top x.$$

The following theorem gives a dual-based notion of *directional sufficiency* for the full robust value.

Theorem F.4 (Global directional sufficiency via the dual)**.** *Adopt Assumption F.1 and let $(\lambda^\star, \mu^\star)$ be any optimal solution of* (D[N])*, with associated value-relevant set I_{rel} and face $C_{\mathrm{rel}} = \mathrm{conv}(\{d_i : i \in I_{\mathrm{rel}}\} \cup \{0\})$. Let $S \subseteq [N]$ and suppose that*

$$C_{\mathrm{rel}} \;\subseteq\; \mathrm{conv}\{d_i : i \in S_0\}. \tag{12}$$

Then the robust values of the full and restricted models coincide:

$$z_S^\star \;=\; z_{[N]}^\star.$$

Proof. Let $u^\star := \sum_{i \in [N]_0} \mu_i^\star d_i \in U_{[N]}(r)$ and $v^\star := u^\star / r \in C_{\mathrm{rel}}$. By (12), there exist coefficients $\{\alpha_j\}_{j \in S_0}$ with $\alpha_j \geq 0$ and $\sum_{j \in S_0} \alpha_j = 1$ such that

$$v^\star = \sum_{j \in S_0} \alpha_j d_j, \qquad \text{hence} \qquad u^\star = \, rv^\star = \, \sum_{j \in S_0} (r\alpha_j) d_j.$$

Define coefficients $\widetilde{\mu}_j := r\alpha_j$ for $j \in S_0$. Then $\widetilde{\mu}_j \geq 0$ and $\sum_{j \in S_0} \widetilde{\mu}_j = r$, and

$$c_0 + M \sum_{j \in S_0} \widetilde{\mu}_j d_j = \; c_0 + Mu^\star = \; c_0 + M \sum_{i \in [N]_0} \mu_i^\star d_i = \; -A^\top \lambda^\star,$$

where the last equality uses feasibility of $(\lambda^\star, \mu^\star)$ in *(D[N])*. Thus $(\lambda^\star, \widetilde{\mu})$ is feasible for the dual of the restricted model *(D[S])*, with objective value $-b^\top \lambda^\star = z_{[N]}^\star$. By strong duality for *(P[S])*–*(D[S])*,

$$z_S^\star = \max\{-b^\top \lambda : \; (\lambda, \mu) \text{ feasible in } (D[S])\} \;\geq\; -b^\top \lambda^\star \;=\; z_{[N]}^\star.$$

On the other hand, $U_S(r) \subseteq U_{[N]}(r)$ implies monotonicity of the robust value in the uncertainty set, whence $z_S^\star \leq z_{[N]}^\star$. Combining the two inequalities yields $z_S^\star = z_{[N]}^\star$. □

F.5 INSTANCE-WISE NO-FREE-LUNCH FOR INSUFFICIENT SUBSETS

We now show that if a subset S fails to represent the relevant convex geometry of the dictionary, then there is no uniform guarantee relating the restricted and full robust values: one can construct instances for which the value gap is arbitrarily large.

Theorem F.5 (Instance-wise no-free-lunch for insufficient subsets)**.** *Let $\mathcal{D} = \{d_i\}_{i=1}^N \subset \mathbb{R}^p$ be a finite dictionary and let $S \subset [N]$ be such that*

$$\operatorname{conv}\{d_i : i \in S\} \subsetneq \operatorname{conv}\{d_i : i \in [N]\}.$$

Then there exist a dimension $n \in \mathbb{N}$, a nonempty bounded polyhedron $X \subset \mathbb{R}^n$ and data (A, b, c_0, M, r) for the robust LP template of Assumption F.1 (with this X and $\mathcal{D}$) such that, for the corresponding full and restricted robust values $z^\star_{[N]}$ and $z^\star_S$, the following holds: for every $K > 0$ one can choose (A, b, c_0, M, r) with

$$z^\star_{[N]} - z^\star_S \geq K.$$

In particular, the family of restricted models based on $U_S(r)$ cannot provide a uniform approximation of the full directional model across all robust LP instances built on the same dictionary $\mathcal{D}$.

Proof. Since $\operatorname{conv}\{d_i : i \in S\} \subsetneq \operatorname{conv}\{d_i : i \in [N]\}$, the polytope $\operatorname{conv}\{d_i : i \in [N]\}$ has at least one extreme point that does not belong to $\operatorname{conv}\{d_i : i \in S\}$. Let d_{i_0} be such an extreme point. By the separating hyperplane theorem, there exists $q \in \mathbb{R}^p$ such that

$$\langle d_{i_0}, q\rangle > \max_{y \in \operatorname{conv}\{d_i : i \in S\}} \langle y, q\rangle = \max_{j \in S} \langle d_j, q\rangle.$$

In particular,

$$\Delta := \max_{i \in [N]} \langle d_i, q\rangle - \max_{j \in S} \langle d_j, q\rangle > 0.$$

We construct an instance in dimension $n = 1$. Let

$$X := \{x \in \mathbb{R} : x = 1\},$$

which is a nonempty bounded polyhedron (it can be written as $\{x : x \leq 1,\ -x \leq -1\}$). Take $c_0 := 0 \in \mathbb{R}$ and define $M \in \mathbb{R}^{1\times p}$ by $M := \begin{bmatrix} q^\top \end{bmatrix}$. For any $u \in \mathbb{R}^p$ and $x \in X$,

$$(c_0 + Mu)^\top x = (Mu)\, x = q^\top u.$$

For radius $r > 0$, the robust values of the full and restricted models are

$$z^\star_{[N]} = \sup_{u \in U_{[N]}(r)} q^\top u = r \max_{i \in [N]_0} \langle d_i, q\rangle, \qquad z^\star_S = \sup_{u \in U_S(r)} q^\top u = r \max_{j \in S_0} \langle d_j, q\rangle,$$

where $S_0 := S \cup \{0\}$ and $[N]_0 := \{0, 1, \ldots, N\}$. By definition of Δ and the inclusion of 0 in both index sets, we have

$$\max_{i \in [N]_0} \langle d_i, q\rangle - \max_{j \in S_0} \langle d_j, q\rangle \geq \max_{i \in [N]} \langle d_i, q\rangle - \max_{j \in S} \langle d_j, q\rangle = \Delta > 0.$$

Therefore

$$z^\star_{[N]} - z^\star_S = r\Big(\max_{i \in [N]_0} \langle d_i, q\rangle - \max_{j \in S_0} \langle d_j, q\rangle\Big) \geq r\,\Delta.$$

Given any prescribed $K > 0$, choosing $r := K/\Delta$ yields $z^\star_{[N]} - z^\star_S \geq K$, as claimed. □

G GREEDY WITH INEXACT ORACLE/SELECTION AND VALUE CERTIFICATE

In practice, the marginal gains of $F(\cdot; \mathcal{T})$ (see (9)) are computed with error and the selection rule can be suboptimal. We model this per iteration.

Assumption G.1 (Per-iteration inexact selection)**.** At iteration t, let $\Delta F_t^{\max}$ be the best possible marginal gain of $F(\cdot; \mathcal{T})$ among all indices $i \in [N] \setminus S_{t-1}$, that is,

$$\Delta F_t^{\max} := \max_{i \in [N] \setminus S_{t-1}} \big[F(S_{t-1} \cup \{i\}; \mathcal{T}) - F(S_{t-1}; \mathcal{T})\big].$$

We assume the chosen index i_t satisfies

$$\Delta F_t(\text{chosen}) \geq (1 - \rho)\, \Delta F_t^{\max} - \varepsilon_t,$$

with $\rho \in [0, 1)$ (multiplicative suboptimality) and $\varepsilon_t \geq 0$ (additive error).

Theorem G.2 (Inexact greedy guarantee for F)**.** *Under submodularity and monotonicity of $F(\cdot;\mathcal{T})$ and Assumption G.1, after B steps,*

$$F(S_{\text{aprx}};\mathcal{T}) \;\geq\; \left(1-\left(1-\tfrac{1-\rho}{B}\right)^{B}\right) F(S^{\star};\mathcal{T}) \;-\; \sum_{t=1}^{B}\left(1-\tfrac{1-\rho}{B}\right)^{B-t}\varepsilon_t,$$

where $S^{\star}$ maximizes $F(\cdot;\mathcal{T})$ with $|S^{\star}|\leq B$. In particular,

$$\left(1-\tfrac{1-\rho}{B}\right)^{B} \;\leq\; e^{-(1-\rho)} \quad\Rightarrow\quad F(S_{\text{aprx}};\mathcal{T}) \;\geq\; \left(1-e^{-(1-\rho)}\right) F(S^{\star};\mathcal{T}) \;-\; \sum_{t=1}^{B} e^{-\frac{1-\rho}{B}(B-t)}\,\varepsilon_t.$$

Proof. Let S_t be the subset selected after t iterations, with $S_0=\varnothing$, and let

$$G_t := F(S^{\star};\mathcal{T}) - F(S_t;\mathcal{T}).$$

We write $F(\cdot)$ for $F(\cdot;\mathcal{T})$ to simplify notation.

By monotonicity and submodularity, the standard greedy lower bound gives

$$\Delta F_t^{\max} \;\geq\; \frac{1}{B}\big(F(S^{\star}) - F(S_{t-1})\big) \;=\; \frac{1}{B}G_{t-1}.$$

(Indeed, adding the elements of $S^{\star}\setminus S_{t-1}$ one by one and using submodularity yields $F(S^{\star})-F(S_{t-1})\leq B\,\Delta F_t^{\max}$ since $|S^{\star}|\leq B$.)

By Assumption G.1,

$$F(S_t)-F(S_{t-1}) \;=\; \Delta F_t(\text{chosen}) \;\geq\; (1-\rho)\,\Delta F_t^{\max} - \varepsilon_t \;\geq\; \frac{1-\rho}{B}G_{t-1} - \varepsilon_t.$$

Rearranging,

$$G_t = F(S^{\star}) - F(S_t) \leq \Big(1-\frac{1-\rho}{B}\Big)G_{t-1} + \varepsilon_t.$$

Set

$$a := 1-\frac{1-\rho}{B} \in [0,1).$$

Then the recurrence is

$$G_t \leq a\,G_{t-1} + \varepsilon_t.$$

Unrolling for $t=1,\dots,B$ gives

$$G_B \leq a^B G_0 + \sum_{t=1}^{B} a^{B-t}\varepsilon_t.$$

Since $S_0=\varnothing$ and $F(\varnothing;\mathcal{T})=0$ (because $d_0=0$ is included), we have

$$G_0 = F(S^{\star};\mathcal{T}) - F(\varnothing;\mathcal{T}) = F(S^{\star};\mathcal{T}).$$

Therefore,

$$F(S_B;\mathcal{T}) = F(S^{\star};\mathcal{T}) - G_B \geq (1-a^B)F(S^{\star};\mathcal{T}) - \sum_{t=1}^{B} a^{B-t}\varepsilon_t.$$

Since $S_{\text{aprx}} = S_B$, this proves the first claim.

For the exponential form, use $(1-x)\leq e^{-x}$ with $x=(1-\rho)/B$:

$$a^B = \left(1-\frac{1-\rho}{B}\right)^{B} \leq e^{-(1-\rho)}, \qquad a^{B-t}\leq e^{-\frac{1-\rho}{B}(B-t)}.$$

Substituting these bounds yields the second inequality. □

Remark G.3 (Mapping oracle errors to (ρ, ε_t)). If at iteration t the marginal gains are evaluated with a *uniform* additive error δ_t *both* for $F(S_{t-1} \cup \{i\})$ *and* for $F(S_{t-1})$, then $\rho = 0$ and $\varepsilon_t \le 2\delta_t$ (the gain is a difference of two quantities, each with error $\le \delta_t$). If moreover only a multiplicative factor $(1 - \tilde{\rho})$ with respect to the best gain is guaranteed, one may take $\rho = \tilde{\rho}$.

Robust value certificate with inexact selection. Combining the pointwise value-gap identity (5)–(6) with a finite family $\mathcal{T}$ (Section 3.1), we obtain the **deterministic certificate**

$$\min_x \Phi_{[N]}(x) \; - \; \min_x \Phi_{S_{\text{aprx}}}(x) \; \le \; r\Big(\max_{s \in \mathcal{T}} \Delta_{S_{\text{aprx}}}(s) \; + \; 2\varepsilon \Big), \tag{13}$$

where ε is the grid granularity over $q(X)$ (Theorem 3.1). The right-hand side is computable at run time.

Certified stopping rule. Fix a threshold $\tau > 0$; stop when

$$\max_{s \in \mathcal{T}} \Delta_S(s) \; \le \; \frac{\tau}{2r} \qquad \text{and} \qquad \varepsilon \; \le \; \frac{\tau}{4r},$$

and report the certificate $\min \Phi_{[N]} - \min \Phi_S \le \tau$.

H EXPERIMENTAL DETAILS

This appendix provides implementation details for the experiments in Sec. 5, covering out-of-sample calibration, frozen feature classification, and scaling and runtime metrics.

Hardware and implementation details. All methods were implemented in `Python 3.10.12` and experiments ran on Ubuntu Linux with an Intel® Xeon® Gold 6230R processor (104 logical cores, 2.10 GHz) and 187 GiB of memory. Wall clock runtime is reported.

H.1 OUT-OF-SAMPLE CALIBRATION USAGE AND CONSERVATISM

Radius calibration applies to the synthetic geometric validation experiment. For a selected subset S, a radius $\widehat{r}(S)$ is estimated from a calibration set, reporting the guarantee of Theorem 3.3. In the frozen feature classification experiment, radius is not calibrated; the reported metric is the certified accuracy induced by the atomic model and the dictionary.

Calibration conservatism and out-of-sample feasibility. This subsection reports a calibration ablation for the synthetic pipeline. The analysis compares DKW union calibration against split calibration on a held out test split, detailing sensitivity to the target violation level α and confidence parameter δ.

Results are computed from `calibration_ablation.csv`, storing one row per $(B, \alpha, \delta, \text{method})$ configuration. Configurations include budgets $B \in \{9, 19, 29\}$, target levels $\alpha \in \{0.01, 0.03, 0.05, 0.1\}$, confidence levels $\delta \in \{10^{-5}, 10^{-4}, 10^{-3}\}$, and calibration methods {DKW union, Split}.

Table 7 summarizes the ablation by calibration method and confidence parameter. DKW union calibration produces higher correction terms and calibrated radii, with negative mean violation gaps $(\widehat{\text{viol}} - \alpha)$ across tested δ. Split calibration yields lower radii and positive mean violation gaps.

For DKW union, the mean correction term changes from 0.2316 at $\delta = 10^{-3}$ to 0.2367 at $\delta = 10^{-5}$, and the mean violation gap remains negative (-0.0082 to -0.0086). For split calibration, the mean correction term changes from 0.0616 to 0.0781, while the mean violation gap remains positive (0.0048 to 0.0037).

Table 8 summarizes these quantities by budget after averaging over tested (α, δ) pairs. DKW union radius increases with budget because the union correction grows with combinatorial family size, while split correction remains constant.

Figure 7 reports correction magnitude at the default confidence level, and Figure 8 reports the ablation over (α, δ) for both calibration methods. The left panel plots observed out-of-sample violation against the target level at default confidence. The right panel plots the violation gap as a function of the confidence parameter at fixed $\alpha = 0.05$.

Table 7: Calibration ablation summary by method and confidence parameter. Gap is observed violation minus target level.

Method	δ	Mean corr.	Mean radius	Mean viol.	Mean gap	Max gap	Min gap
DKW union	10^{-5}	0.2367	4.4114	0.0389	-0.0086	0.0000	-0.0254
DKW union	10^{-4}	0.2342	4.4089	0.0391	-0.0084	0.0000	-0.0250
DKW union	10^{-3}	0.2316	4.4064	0.0393	-0.0082	0.0000	-0.0248
Split	10^{-5}	0.0781	4.2529	0.0512	0.0037	0.0088	0.0008
Split	10^{-4}	0.0704	4.2451	0.0519	0.0044	0.0100	0.0016
Split	10^{-3}	0.0616	4.2364	0.0524	0.0048	0.0106	0.0026

Table 8: Calibration ablation summary by budget and calibration method, averaged over (α, δ) settings.

B	Method	Mean corr.	Mean radius	Mean viol.	Mean gap
9	DKW union	0.1809	4.3557	0.0429	-0.0046
9	Split	0.0700	4.2448	0.0518	0.0043
19	DKW union	0.2397	4.4144	0.0385	-0.0090
19	Split	0.0700	4.2448	0.0518	0.0043
29	DKW union	0.2818	4.4566	0.0360	-0.0115
29	Split	0.0700	4.2448	0.0518	0.0043

Table 9 reports configurations at $\alpha = 0.05$ for tested budgets and confidence levels. Metrics include raw calibration quantile, correction term, calibrated radius, observed violation, and violation gap.

The pairwise correction ratio between DKW union and split calibration, computed at matched (B, α, δ) settings, has mean 3.3715 and maximum 4.5388. The radius increase induced by DKW union averages 0.1641, while the change in violation gap $(\widehat{\text{viol}} - \alpha)$ averages -0.0127.

DKW union calibration provides feasibility control with higher correction terms and calibrated radii, while split calibration yields lower radii but can exceed the target violation level on the held out test split.

H.2 SYNTHETIC GEOMETRIC VALIDATION

A synthetic two dimensional dictionary with $N = 2000$ atoms and budgets $B \in \{2, 4, 8, 16, 32\}$ is utilized. Support approximation is evaluated along $x_{\text{fixed}} = (1, 0)$. A calibration set of size $m_{\text{cal}} = 1000$ is employed with $(\alpha, \delta) = (0.05, 10^{-5})$.

The dictionary is generated once and reused. It mixes structured directions, isotropic noise, and decoy directions orthogonal to x_{fixed}.

For any subset S, the following is reported:

$$z_S(r) = \sup_{u \in U_S(r)} \langle u, x_{\text{fixed}} \rangle = r \max\Big(0, \max_{i \in S} \langle d_i, x_{\text{fixed}} \rangle\Big),$$

and compared to $z_{[N]}(r)$. An ellipsoidal baseline fitted to the calibration sample is included.

H.2.1 Surrogate versus certificate alignment in the synthetic validation

This subsection quantifies the relationship between the surrogate selection objective and the reported worst deficit certificate proxy in the synthetic geometric validation. The selection step uses the average coverage objective

$$F(S; T_{\text{eval}}),$$

while the reporting includes the worst deficit proxy

$$G(S; T_{\text{eval}}) := r_{\text{eval}} \max_{s \in T_{\text{eval}}} \Delta_S(s).$$

A direct empirical comparison between these quantities clarifies when improvements in the surrogate objective align with improvements in the certificate proxy.

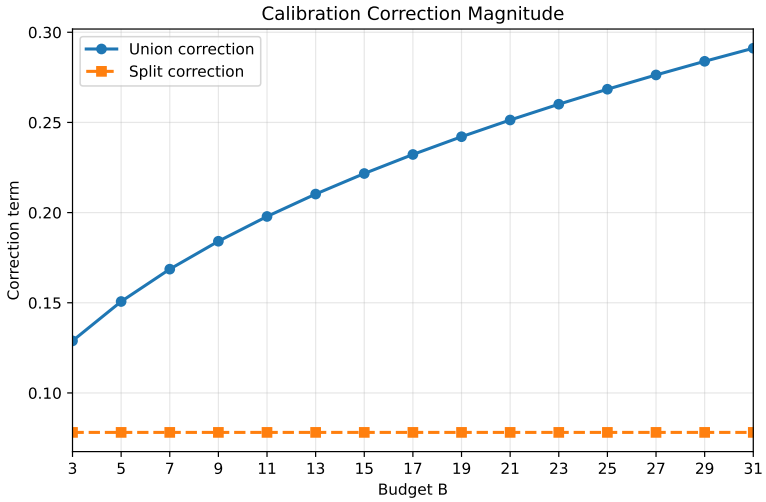

Figure 7: Calibration correction magnitude at the default confidence level. DKW union correction increases with budget; split correction is constant.

Table 9: Calibration settings at fixed $\alpha = 0.05$ for tested budgets and confidence levels.

B	δ	Method	Quantile	Corr.	Radius	Viol.	Gap
9	10^{-5}	DKW union	3.9942	0.1841	4.1783	0.0490	-0.0010
9	10^{-5}	Split	3.9942	0.0781	4.0723	0.0588	0.0088
9	10^{-4}	DKW union	3.9942	0.1810	4.1752	0.0492	-0.0008
9	10^{-4}	Split	3.9942	0.0704	4.0646	0.0600	0.0100
9	10^{-3}	DKW union	3.9942	0.1778	4.1720	0.0492	-0.0008
9	10^{-3}	Split	3.9942	0.0616	4.0559	0.0606	0.0106
19	10^{-5}	DKW union	3.9942	0.2421	4.2363	0.0432	-0.0068
19	10^{-5}	Split	3.9942	0.0781	4.0723	0.0588	0.0088
19	10^{-4}	DKW union	3.9942	0.2397	4.2339	0.0438	-0.0062
19	10^{-4}	Split	3.9942	0.0704	4.0646	0.0600	0.0100
19	10^{-3}	DKW union	3.9942	0.2373	4.2315	0.0438	-0.0062
19	10^{-3}	Split	3.9942	0.0616	4.0559	0.0606	0.0106
29	10^{-5}	DKW union	3.9942	0.2839	4.2781	0.0390	-0.0110
29	10^{-5}	Split	3.9942	0.0781	4.0723	0.0588	0.0088
29	10^{-4}	DKW union	3.9942	0.2819	4.2761	0.0392	-0.0108
29	10^{-4}	Split	3.9942	0.0704	4.0646	0.0600	0.0100
29	10^{-3}	DKW union	3.9942	0.2798	4.2740	0.0396	-0.0104
29	10^{-3}	Split	3.9942	0.0616	4.0559	0.0606	0.0106

For each method and budget, the following normalized metrics are evaluated:

$$F_{\text{ratio}}(S) := \frac{F(S; T_{\text{eval}})}{F([N]; T_{\text{eval}})}, \qquad G_{\text{ratio}}(S) := \frac{G(S; T_{\text{eval}})}{G(\varnothing; T_{\text{eval}})}.$$

The correlation is computed between F_{ratio} and $-G_{\text{ratio}}$, so larger values on both axes correspond to better performance.

The statistics in this subsection use all available budgets exported by the synthetic experiment. Table 10 reports the aggregate statistics. The Spearman correlation indicates a monotone alignment between the surrogate objective and the certificate proxy in this setup. The Greedy Coverage and Greedy MaxGap comparison shows divergence only at budgets $B \in \{1, 3, 4\}$. This localizes the mismatch to the micro budget regime.

Figure 9 plots the normalized surrogate objective against the normalized certificate proxy across methods and budgets, together with a direct Greedy Coverage versus Greedy MaxGap difference diagnostic. The first panel shows a monotone trend. The second panel reports two normalized differences for each budget:

$$\Delta F := F_{\text{ratio}}^{\text{GC}} - F_{\text{ratio}}^{\text{GM}}, \qquad \Delta G_{\text{score}} := G_{\text{ratio}}^{\text{GM}} - G_{\text{ratio}}^{\text{GC}},$$

where positive ΔG_{score} means Greedy Coverage achieves a smaller certificate proxy. A sign disagreement between ΔF and ΔG_{score} indicates a surrogate versus certificate divergence.

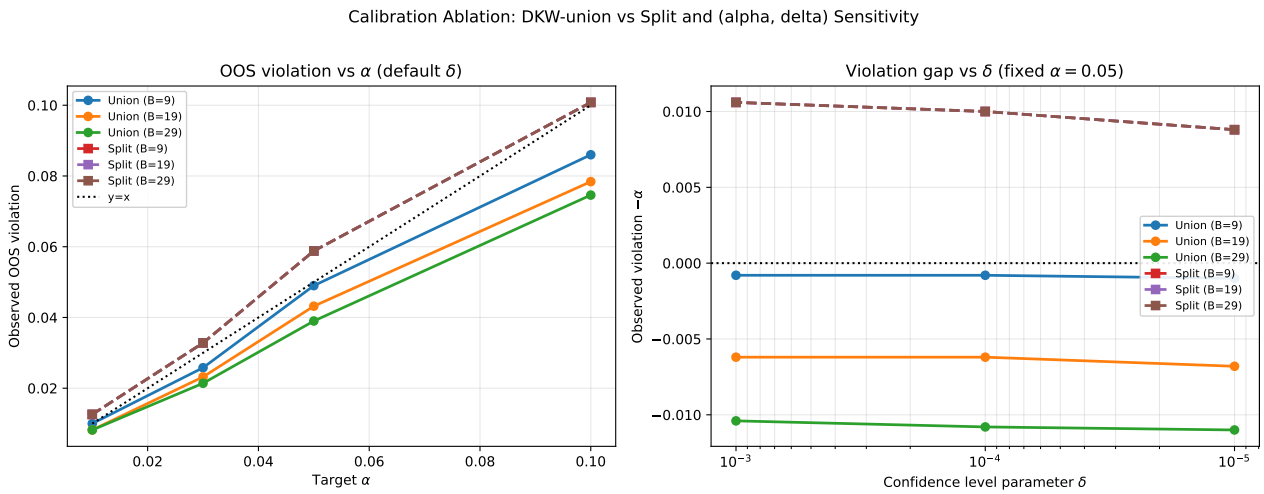

Figure 8: Calibration ablation with (α, δ) sweeps. Left: observed out-of-sample violation versus target α at default confidence level. Right: violation gap $(\widehat{\text{viol}} - \alpha)$ versus δ at fixed $\alpha = 0.05$.

Table 10: Summary of surrogate versus certificate alignment in the synthetic geometric validation.

Quantity	Value
Number of analyzed points	124
Pearson corr$(F_{\text{ratio}}, -G_{\text{ratio}})$	0.8554
Spearman corr$(F_{\text{ratio}}, -G_{\text{ratio}})$	0.9617
Budgets with GC versus GM divergence	3
Divergence budgets	$\{1, 3, 4\}$

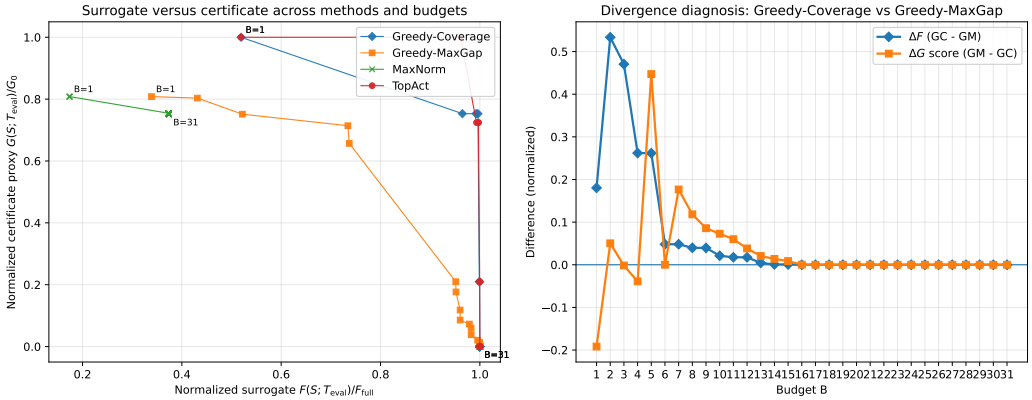

Figure 9: Surrogate versus certificate analysis in the synthetic geometric validation. Left: normalized surrogate objective F_{ratio} versus normalized certificate proxy G_{ratio} across methods and budgets. Lower values of G_{ratio} are better. Right: direct comparison between Greedy Coverage and Greedy MaxGap through ΔF and ΔG_{score} as functions of the budget.

Table 11 reports representative budget values for the Greedy Coverage versus Greedy MaxGap comparison. Divergence is confined to the smallest budgets, while the two criteria align after that regime.

Table 12 summarizes the start and end points of each method trajectory. The normalized trajectories support the same conclusion: the surrogate objective acts as a guide for the certificate proxy beyond the smallest budgets in this synthetic setup.

In the synthetic geometric validation, the surrogate objective F and the certificate proxy G align well overall, with divergence concentrated in the micro budget regime. This supports using the submodular surrogate for selection while preserving the worst deficit metric as the reporting and diagnostic quantity.

Table 11: Representative Greedy Coverage versus Greedy MaxGap comparisons. Positive ΔF favors Greedy Coverage on the surrogate objective. Positive ΔG_{score} favors Greedy Coverage on the certificate proxy.

B	ΔF	ΔG_{score}	GC F_{ratio}	GC G_{ratio}	GM F_{ratio}	GM G_{ratio}
1	0.1803	−0.1918	0.5193	1.0000	0.3390	0.8082
3	0.4705	−0.0017	0.9924	0.7531	0.5219	0.7513
4	0.2620	−0.0390	0.9962	0.7531	0.7342	0.7140
5	0.2618	0.4472	0.9989	0.2097	0.7371	0.6568
9	0.0396	0.0857	1.0000	0.0000	0.9604	0.0857
19	0.0000	0.0000	1.0000	0.0000	1.0000	0.0000
29	0.0000	0.0000	1.0000	0.0000	1.0000	0.0000

Table 12: Method trajectory endpoints in normalized surrogate and certificate coordinates.

Method	F_{ratio} (start → end)	G_{ratio} (start → end)
Greedy Coverage	0.5193 → 1.0000	1.0000 → 0.0000
Greedy MaxGap	0.3390 → 1.0000	0.8082 → 0.0000
MaxNorm	0.1736 → 0.3737	0.8082 → 0.7514
TopAct	0.5193 → 1.0000	1.0000 → 0.0000

H.3 OPERATIONALIZING THE ε TERM IN THEOREM 3.1

This section reports an empirical evaluation of the discretization term in Theorem 3.1, quantifying the approximation gap induced by replacing the direction set $q(X)$ with a finite set T.

Let

$$d_D(s,t) = \max_i |\langle d_i, s-t\rangle|$$

and let T_{probe} denote an independent probe pool of normalized directions. The empirical discretization error is defined as

$$\widehat{\varepsilon}(T) = \max_{s\in T_{\text{probe}}} \min_{t\in T} d_D(s,t).$$

This quantity is the empirical counterpart of the ε net radius in Theorem 3.1.

In the synthetic two dimensional experiment, the same finite set T is used for selection and reporting, with $|T| = 801$, and the independent probe set has $|T_{\text{probe}}| = 4000$. The estimate is:

$$\widehat{\varepsilon}(T) = 1.824282, \qquad 2\widehat{\varepsilon}(T) = 3.648563.$$

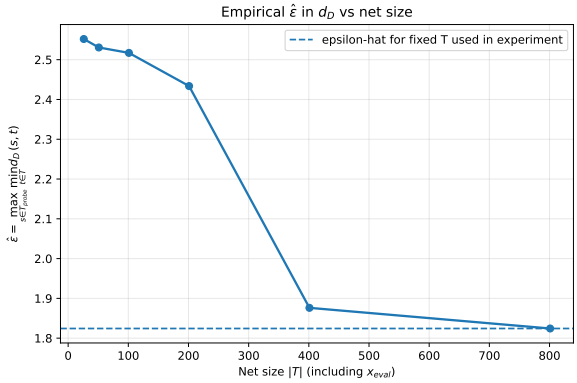

Figure 10: Empirical discretization error $\widehat{\varepsilon}(T)$ in the metric d_D as a function of the net size $|T|$. The curve is computed from nested prefixes of the direction pool and an independent probe set T_{probe}. The estimated curve decreases monotonically over the reported refinement path.

Table 13 summarizes the statistics. The mean and the 95th percentile of the nearest net distance are small relative to the supremum $\widehat{\varepsilon}(T)$, indicating the worst case discretization term is driven by a small subset of probe directions.

Table 13: Summary of the empirical $\widehat{\varepsilon}(T)$ estimate in the synthetic two dimensional experiment.

Quantity	Value
$\lvert T\rvert$ used in selection and reporting	801
$\lvert T_{\mathrm{probe}}\rvert$	4000
Number of metric atoms used for d_D	8
$\widehat{\varepsilon}(T)$	1.824282
$2\widehat{\varepsilon}(T)$	3.648563
Mean nearest net distance on T_{probe}	0.013876
95th percentile nearest net distance on T_{probe}	0.027180

Table 14: Empirical $\widehat{\varepsilon}(T)$ along the net refinement path.

$\lvert T\rvert$	$\widehat{\varepsilon}(T)$
26	2.551961
51	2.530695
101	2.517097
201	2.433706
401	1.876200
801	1.824282

Table 14 reports the refinement path. The empirical estimate decreases from 2.551961 to 1.824282 as $\lvert T\rvert$ increases from 26 to 801, representing a reduction of 28.51%.

Theorem 3.1 bounds the robust value gap through the term

$$\max_{t\in T} \Delta_S(t) + 2\varepsilon.$$

Using the empirical estimate $\widehat{\varepsilon}(T)$, this section reports the practical decomposition

$$\max_{t\in T} \Delta_S(t) + 2\widehat{\varepsilon}(T).$$

Figure 11 shows this decomposition for Greedy Coverage and Greedy MaxGap.

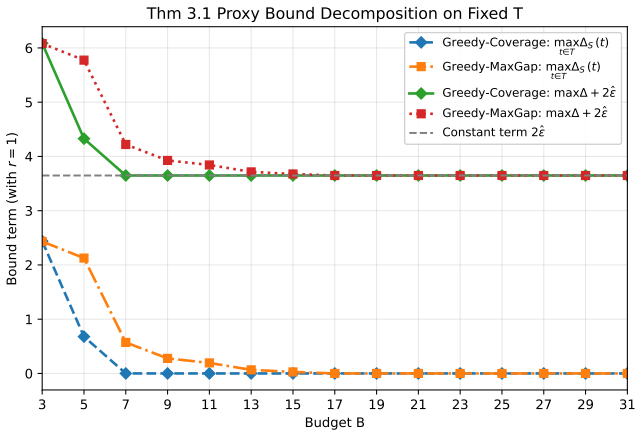

Figure 11: Empirical decomposition of the Theorem 3.1 bound term on the finite set T. The curves show $\max_{t\in T} \Delta_S(t)$ and $\max_{t\in T} \Delta_S(t) + 2\widehat{\varepsilon}(T)$ for Greedy Coverage and Greedy MaxGap. The horizontal line corresponds to the constant discretization contribution $2\widehat{\varepsilon}(T)$.

The decomposition reveals two regimes. For small budgets, the reduction in the bound results from the decrease in $\max_{t\in T} \Delta_S(t)$. For moderate budgets, the alignment deficit becomes small, and the constant term $2\widehat{\varepsilon}(T)$ dominates. In this

range, increasing the atom budget B has minimal effect on the certificate, and the primary method to tighten the bound is the refinement of the direction set T.

For reference, Greedy Coverage reaches $\max_{t\in T} \Delta_S(t) = 0$ at $B = 9$ in this experiment, while Greedy MaxGap reaches $\max_{t\in T} \Delta_S(t) = 0$ at $B = 17$.

Table 15: Representative values of the practical Theorem 3.1 decomposition for Greedy Coverage and Greedy MaxGap.

B	Greedy Coverage		Greedy MaxGap		$2\widehat{\varepsilon}(T)$
	$\max_{t\in T} \Delta_S(t)$	$+ 2\widehat{\varepsilon}(T)$	$\max_{t\in T} \Delta_S(t)$	$+ 2\widehat{\varepsilon}(T)$	
3	2.438163	6.086726	2.432507	6.081070	3.648563
5	0.678860	4.327424	2.126637	5.775201	3.648563
9	0.000000	3.648563	0.277614	3.926178	3.648563
19	0.000000	3.648563	0.000000	3.648563	3.648563
29	0.000000	3.648563	0.000000	3.648563	3.648563

Although Greedy Coverage optimizes an average coverage objective, it matches or improves the worst case deficit $\max_{t\in T} \Delta_S(t)$ at several budgets shown in Table 15. The empirical estimate $\widehat{\varepsilon}(T)$ makes the discretization term in Theorem 3.1 computable. The reported decomposition shows atom selection controls the alignment deficit at small budgets, while net refinement controls the certificate once $\max_{t\in T} \Delta_S(t)$ is small. This provides a diagnostic for deciding whether to increase B or to enrich the direction set T.

H.4 TRANSACTION-COST-AWARE ROBUST PORTFOLIO ALLOCATION

The portfolio experiment tests the same sparse robustification principle in a classical robust optimization setting. We use Fama–French portfolios, a standard empirical asset-pricing dataset introduced by Fama and French [Fama and French, 1993]. At transaction cost $\tau = 0.0025$, Greedy-Coverage retains nearly the same net Sharpe as Mean Variance while reducing turnover and annualized trading cost.

Table 16: Transaction-cost-aware portfolio allocation. Higher annualized net Sharpe and net mean are better; lower turnover and annualized cost are better.

Method	Annualized net Sharpe ↑	Annualized net mean ↑	One-way turnover ↓	Annualized cost ↓
Greedy-Coverage	**0.7359**	**0.1069**	0.0286	0.00086
Mean Variance	0.7391	0.1071	0.0456	0.00137
Equal Weight	0.5161	0.0977	**0.0123**	**0.00037**

Greedy-Coverage retains 99.6% of the Mean Variance net Sharpe while reducing turnover and annualized trading cost by 37.2%. This supports a favorable performance–cost tradeoff for the objective-aligned sparse robust method.

H.5 FROZEN FEATURE CIFAR CLASSIFICATION

A pretrained ResNet 18 backbone is used to extract frozen features. Tasks are binary, defined by two classes from CIFAR 10 or CIFAR 100, with labels mapped to $\{-1, +1\}$. The training features are split into disjoint parts with proportions

$$[0.5,\ 0.2,\ 0.15,\ 0.15],$$

used for classifier training, dictionary mining, selection directions, and reporting directions.

A linear classifier is trained with Adam and fixed hyperparameters. A dictionary is mined in feature space using a projected gradient attack in ℓ_2 with radius ε. Normalized perturbation directions are stored, and certified accuracy $\mathrm{CertAcc}(\varepsilon)$ alongside projected gradient robust accuracy are evaluated.

Certified accuracy. Let $w \in \mathbb{R}^d$ be the weight vector of the linear classifier, and let $\mathcal{D} = \{d_i\}_{i=1}^N \subseteq \mathbb{R}^d$ be the directional dictionary in feature space. Define

$$\mathrm{pen}_{\mathcal{D}}(w) \;:=\; \max_{i\in[N]} \big|\langle d_i, w\rangle\big|.$$

For $\varepsilon > 0$, define

$$\mathrm{CertAcc}_{\mathcal{D}}(\varepsilon) \;:=\; \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\{y_j\langle w, x_j\rangle > \varepsilon\,\mathrm{pen}_{\mathcal{D}}(w)\}\,,$$

where (x_j, y_j) are test examples in feature space and $y_j \in \{-1, +1\}$. For a selected subset $S \subseteq [N]$, the reported metric is $\mathrm{CertAcc}_S(\varepsilon)$, obtained by replacing $\mathrm{pen}_{\mathcal{D}}$ with $\mathrm{pen}_S(w) = \max_{i\in S} |\langle d_i, w\rangle|$.

H.6 FULL INSTANCE CONTEXT FOR BUDGET EFFICIENCY AND RELIABILITY ANALYSIS

Section 5.4 reports onset budget $B^\star$ and budget-wise success rates on the subset of instances that are robust feasible under the full dictionary baseline, with zero violations. This conditioning isolates budget efficiency from baseline infeasibility.

Across all generated seeds ($N = 100$), the full dictionary baseline is robust feasible on 64 seeds (64.0%) and infeasible on 36 seeds (36.0%). Table 17 summarizes policy-level certification counts over all 100 seeds on the same budget grid used in Section 5.4, making the screening step explicit.

Table 17: Policy-level certification summary over all generated instances ($N = 100$) for the experiment in Section 5.4. Full feasible and Full infeasible refer to feasibility under the full dictionary baseline.

Policy	Full feasible	Full infeasible	Total certified	Not certified
greedy-maxgap	58/64 (90.6%)	3/36 (8.3%)	61/100 (61.0%)	39/100 (39.0%)
greedy-coverage	55/64 (85.9%)	3/36 (8.3%)	58/100 (58.0%)	42/100 (42.0%)
max-gamma	53/64 (82.8%)	3/36 (8.3%)	56/100 (56.0%)	44/100 (44.0%)

The $N = 64$ subset in Section 5.4 is the relevant population for onset budget and budget-wise efficiency comparisons because those metrics are meaningful only when the full dictionary baseline is robust feasible. Table 17 complements that analysis by reporting certification outcomes on the full instance pool.

Certification on full-baseline infeasible seeds does not imply feasibility of the full dictionary baseline. It means that a policy achieves zero violations for that seed at some evaluated budget in its policy-specific sweep, while full-baseline feasibility is defined with respect to the full dictionary baseline.