Ranking-Preserving Coresets for Global NAS via Pairwise-Margin Uniform Convergence

Liz Lemma Future Detective

January 20, 2026

Abstract

Efficient global neural architecture search (NAS) is bottlenecked by candidate evaluation: on large datasets and high resolutions, even short training runs dominate search cost. Recent work on efficient global NAS improves ranking fidelity by adapting training budgets to architecture capacity, but still evaluates candidates on the full dataset, making search time scale with data size and image resolution. We propose a complementary approach: select a small search-time coreset that preserves the relative ranking of architectures under short training, rather than maximizing absolute accuracy of any single model. We formalize ranking preservation through pairwise score margins and Kendall-τ distortion, and derive bounds linking uniform convergence of proxy scores on a subset to ranking correctness for all architecture pairs with sufficient margin. Using pseudo-dimension/Rademacher tools, we show that $\tilde{O}(d/\varepsilon^2)$ examples suffice to guarantee margin-preserving rankings for a hypothesis class of dimension d, and we provide matching lower bounds. To make the method constructive and practical, we design a probe-guided greedy selection algorithm that maximizes a monotone submodular surrogate of ranking preservation over a probe pool of architectures, yielding a $(1 - 1/e)$ guarantee for the surrogate. We outline experiments showing that global NAS on ImageNet-scale and CASIA-WebFace-like regimes can be performed on tiny ranking-preserving subsets with controlled distortion, producing comparable discovered architectures at substantially reduced search compute.

Table of Contents

3. 3. Problem Formulation: Ranking-Preserving Coreset Selection (RPCS) and its variants (for a finite probe set P vs uniform-over-$\mathcal{H}$; fixed-fidelity vs multi-fidelity).

4. 4. Ranking Preservation via Margin Stability: define pairwise margins; show how uniform score approximation implies preservation of all pairwise relations above margin; translate to Kendall-τ bounds.

5. 5. Upper Bounds (Sample Complexity): uniform convergence bounds (pseudo-dimension/Rademacher) for $\sup_h |s_f(h;S) - s_f(h;\mathcal{D})|$; explicit m scaling; margin-based ranking corollaries.

6. 6. Lower Bounds: information-theoretic lower bound $\Omega(d/\varepsilon^2)$ for uniform approximation; construct adversarial classes showing necessity; implication for ranking preservation when margins are small.

7. 7. Constructive Algorithms: (a) probe-guided greedy selection optimizing a submodular surrogate, (b) streaming/online variant, (c) discussion of alternative solvers (DPP/leverage scores) and when they match theory.

8. 8. Complexity Analysis: runtime in terms of dataset size n, probe size $|P|$, subset size m, and fidelity cost; trade-offs between evaluation passes and ranking guarantees.

9. 9. Experimental Protocol (Implementation-Strengthening Section): benchmark setup on NAS-Bench-style spaces and global macro-micro spaces; scaling to ImageNet-proxies and CASIA-WebFace-like regimes; measuring ranking distortion and NAS outcome quality vs compute.

10. 10. Discussion and Limitations: dependence on margins; robustness across seeds/augmentations; extension to hardware-aware ranking and multi-objective NAS; open problems.

1 1. Introduction and Motivation: global NAS evaluation cost scaling with dataset size/resolution; why accuracy-preserving coresets are insufficient for NAS; relation to architecture-aware ranking from the source material.

Global neural architecture search (NAS) is, at its core, an exercise in repeated comparison. For each candidate architecture we must obtain a numerical estimate of its performance under a prescribed training protocol, and the search logic (be it evolutionary, Bayesian, reinforcement-learning, or gradient-based) is driven primarily by relative, not absolute, assessments. In contemporary regimes the dominant cost in this process is not the combinatorics of traversing the architecture space but the expense of producing sufficiently reliable evaluations: even when we restrict to a fixed low-fidelity training recipe, each candidate still incurs a nontrivial training cost and an evaluation cost that scales linearly with the number of examples and with the per-example computational footprint (e.g., input resolution and augmentation). If we denote by n the number of evaluation examples and by c_{eval} the cost of one forward pass, then the evaluation component alone scales as $\Theta(n\, c_{\mathrm{eval}})$ per candidate; multiplying by the number of candidates explored yields a budget that quickly becomes prohibitive as either dataset size or resolution increases.

This scaling pressure motivates the use of subsets of the evaluation split. The naive hope is that, by replacing the full evaluation set with a substantially smaller subset of size $m \ll n$, we might reduce the wall-time of search nearly by the factor n/m while retaining the quality of the final selected architecture. However, the criterion by which such a subset should be judged is subtle. Many existing coreset constructions aim to preserve the *accuracy* (or loss) of a single predictor, or of predictors trained by a fixed learning algorithm, on the full distribution. Such objectives are well matched to classical model selection or to efficient training, where one cares about estimating a single risk value accurately. NAS differs in a crucial respect: we do not seek an accurate estimate for one model, but a stable *ordering* over many competing architectures, often with small score gaps.

The distinction between accurate risk estimation and accurate ranking is not merely semantic. Suppose two architectures h and h' have nearly identical full-data performance; then any subset that is adequate for predicting the loss of a single fixed model may still induce a spurious swap between h and h' due to sampling noise or selection bias. A small number of such swaps is not necessarily problematic if they occur among architectures that are truly indistinguishable. Yet NAS procedures frequently operate near the frontier of attainable performance, where the search must discriminate among can-

didates whose differences are within a narrow band. In that regime, even modest ranking distortion can redirect the search trajectory, leading to a different region of the architecture space and ultimately a different final selection. Thus, the operational requirement is that the subset preserve the pairwise comparisons that actually drive the search, especially among nearties.

This observation also clarifies why a subset that preserves *average* accuracy may be insufficient. A subset can provide an unbiased estimate of the mean score while still exhibiting large variance on the *differences* between models, and it is precisely these differences that determine the ranking. Concretely, the quantity of interest for ranking is the gap $s(h') - s(h)$ rather than either score in isolation. If the subset induces correlated errors across architectures, the ranking may remain stable even when absolute scores drift; conversely, if the subset induces differential errors that vary unpredictably with architecture, the ranking becomes unreliable. Hence, an appropriate coreset notion for NAS must be architecture-aware in the sense that it controls, either directly or indirectly, the deviation of such pairwise gaps between the subset and the full evaluation set.

We emphasize that architecture-awareness is demanded not because architectures are trained differently across candidates (we fix the training protocol at the chosen fidelity), but because the per-example loss landscape depends strongly on the inductive biases encoded by the architecture. Two architectures may agree on most examples and disagree on a small, structurally coherent subset (e.g., images with certain textures, rare classes, or long-range dependencies). Those disagreement sets can dominate the relative ordering even when they constitute a small fraction of the data. Therefore, a subset selected solely to match the marginal label distribution, or to cover the input space in a geometric sense, can omit precisely the examples that separate strong candidates from merely adequate ones. For NAS we must preferentially retain *ranking-sensitive* examples: those for which candidate architectures exhibit heterogeneous losses and hence contribute substantially to pairwise score differences.

A second complication is that NAS is conducted under explicit computational constraints that are naturally expressed through a *fidelity* parameter. In practice, the fidelity may correspond to fewer optimization steps, reduced input resolution, smaller batch sizes, weaker regularization, or reduced data fractions. Coreset selection must therefore be aligned with the fidelity actually used during the search. A subset that preserves ranking at one fidelity may fail at another, because the relative behavior of architectures can change with training time or resolution. For instance, an architecture that learns rapidly might appear superior at low fidelity yet be overtaken at higher fidelity. Our goal is not to eliminate this intrinsic fidelity-induced bias; rather, we aim to ensure that, *conditional on the chosen fidelity*, the ranking induced by evaluating on the subset matches as closely as possible the ranking in-

duced by evaluating on the full evaluation split at the same fidelity.

These considerations suggest a reorientation of coreset design from pointwise risk approximation to ranking preservation. A direct objective would be to minimize a distance between rankings induced by the subset and by the full set, such as Kendall–τ or Spearman metrics, over a relevant collection of architectures. Yet this direct objective is combinatorial and, in general, computationally intractable: it depends on the signs of many pairwise differences and leads to a non-smooth, non-submodular set function. Accordingly, we seek surrogates that (i) are sensitive to pairwise orderings, (ii) can be optimized efficiently under a cardinality constraint, and (iii) admit theoretical connections back to ranking stability. The surrogate viewpoint also matches the practical access model: we can afford to train and evaluate only a modest number of probe architectures at the chosen fidelity, and we must select a subset using limited passes over the dataset.

The resulting picture is as follows. We treat the dataset as an evaluation resource whose elements contribute to pairwise discrimination among architectures. We then design a selection rule that aggregates, across many probe pairs, the evidence provided by each example toward the correct ordering. Intuitively, examples on which all probes behave similarly are redundant for ranking, whereas examples that induce diverse probe losses are informative. By formalizing and optimizing this intuition we obtain subsets that are small yet tailored to the comparative structure of the architecture class at the target fidelity. The remainder of our development makes these notions precise by introducing (a) the score and ranking objects, (b) a uniform approximation condition that suffices for margin-stable ranking, and (c) a constructive algorithm that uses probe architectures to greedily optimize a tractable surrogate of ranking preservation.

2 Preliminaries

Data and evaluation protocol. We fix an evaluation split (or evaluation multiset) denoted by $\mathcal{D} = \{z_1, \ldots, z_n\}$ with $|\mathcal{D}| = n$, where each z is a labeled example (e.g., an image–label pair). A *subset* (or coreset) is a set $S \subseteq \mathcal{D}$ of size $|S| = m \ll n$; unless stated otherwise we view S as unweighted and we evaluate by uniform averaging over its elements.[1] Throughout, we consider a fixed training split (disjoint from $\mathcal{D}$) and a fixed training recipe; the subset selection problem concerns only how we *evaluate* trained candidates during NAS.

[1] Weighted variants are obtained by replacing uniform averages with $\sum_{z \in S} w_z \ell_f(h; z)$, $\sum_{z \in S} w_z = 1$, and do not change the ranking formalism; we restrict to unweighted subsets to keep the cardinality constraint explicit.

Architectures, hypotheses, and fidelity. Let $\mathcal{H}$ denote the class of architectures under consideration. We write $h \in \mathcal{H}$ for a single architecture together with the fixed training recipe (optimizer, augmentation, regularization, etc.), except for a tunable *fidelity* parameter f. The fidelity f abstracts the computational budget used to obtain a proxy evaluation for NAS; examples include the number of optimization steps or epochs, input resolution, batch size, early-stopping time, or a data-fraction schedule. For a chosen f, we train h on the fixed training split under fidelity f, and we then evaluate on examples $z \in \mathcal{D}$. If the training procedure is randomized (initialization, data order), we may regard all quantities below as conditional on the realized randomness, or else as expectations over it; for notational economy we write deterministic expressions and treat concentration over data selection separately.

Per-example loss and proxy scores. For each $h \in \mathcal{H}$ and example $z \in \mathcal{D}$, let $\ell_f(h; z)$ denote the per-example evaluation loss after training h under fidelity f. We assume $\ell_f(h; z) \in [0, 1]$ (or, more generally, sub-Gaussian); this boundedness is used only for uniform convergence arguments and can be enforced by rescaling standard losses. Given any evaluation set $A \subseteq \mathcal{D}$, we define the *proxy score*

$$s_f(h; A) \;=\; \frac{1}{|A|} \sum_{z \in A} \ell_f(h; z), \tag{1}$$

where lower values indicate better performance. One may replace loss by negative accuracy or error rate without changing the subsequent ranking-based definitions, since any strictly monotone transformation of $s_f(h; A)$ induces the same total order when ties are broken deterministically. We will frequently compare the full-data score $s_f(h; \mathcal{D})$ with the subset score $s_f(h; S)$, emphasizing that both are evaluated *at the same fidelity* f.

Pairwise margins and sign stability. Because NAS dynamics are driven by comparisons, it is convenient to express relative performance via margins. For $h, h' \in \mathcal{H}$ we define the (full-data) pairwise margin

$$\Delta_f(h, h') \;=\; s_f(h'; \mathcal{D}) - s_f(h; \mathcal{D}), \tag{2}$$

and the subset-induced estimate

$$\widehat{\Delta}_f(h, h') \;=\; s_f(h'; S) - s_f(h; S). \tag{3}$$

Thus, $\Delta_f(h, h') < 0$ means that h is better than h' on the full evaluation split at fidelity f, whereas $\widehat{\Delta}_f(h, h')$ is the comparison available to NAS when only S is evaluated. The central event we wish to control is *sign agreement*, namely $\operatorname{sign}(\widehat{\Delta}_f(h, h')) = \operatorname{sign}(\Delta_f(h, h'))$, for as many relevant pairs (h, h') as possible; the magnitude $|\Delta_f(h, h')|$ captures how vulnerable the ordering is to perturbations induced by using S.

Rankings as total orders on finite sets. In order to define ranking distortion we work over a finite set of architectures $A = \{a_1, \ldots, a_L\} \subseteq \mathcal{H}$, with $L \geq 2$; later, A will be either a probe pool P used for selection or a candidate set encountered during NAS. The score vector $\big(s_f(a_\ell; \mathcal{D})\big)_{\ell=1}^{L}$ induces a total order once we specify a tie-breaking rule. Concretely, we define a permutation $\pi_{\mathcal{D}} \in S_L$ such that

$$s_f(a_{\pi_{\mathcal{D}}(1)}; \mathcal{D}) \leq s_f(a_{\pi_{\mathcal{D}}(2)}; \mathcal{D}) \leq \cdots \leq s_f(a_{\pi_{\mathcal{D}}(L)}; \mathcal{D}),$$

with ties broken deterministically (e.g., by architecture index). Analogously, π_S is the permutation induced by $s_f(\cdot; S)$. We write $r_{\mathcal{D}}(a_\ell) \in \{1, \ldots, L\}$ for the rank position of a_ℓ under $\pi_{\mathcal{D}}$, and similarly $r_S(a_\ell)$ under π_S. The pairwise order between two architectures a_i, a_j is then encoded by the sign of $\Delta_f(a_i, a_j)$ (or equivalently by whether $r_{\mathcal{D}}(a_i) < r_{\mathcal{D}}(a_j)$).

Ranking distortion metrics. We quantify disagreement between π_S and $\pi_{\mathcal{D}}$ using standard permutation distances. The *Kendall–τ* distance counts discordant pairs:

$$\tau(\pi_S, \pi_{\mathcal{D}}) \;=\; \sum_{1 \leq i < j \leq L} \mathbb{1}\Big[\big(r_S(a_i) - r_S(a_j)\big)\big(r_{\mathcal{D}}(a_i) - r_{\mathcal{D}}(a_j)\big) < 0\Big], \qquad (4)$$

optionally normalized by $\binom{L}{2}$. The metric τ is directly aligned with pairwise sign agreement: each inversion corresponds to a pair (a_i, a_j) for which the subset flips the ordering relative to the full evaluation. As a complementary notion, the *Spearman* rank correlation measures squared deviations in rank positions,

$$\rho(\pi_S, \pi_{\mathcal{D}}) \;=\; 1 - \frac{6 \sum_{\ell=1}^{L} \big(r_S(a_\ell) - r_{\mathcal{D}}(a_\ell)\big)^2}{L(L^2 - 1)}, \qquad (5)$$

with associated distance $1 - \rho$ (or simply the unnormalized squared rank error). Whereas Kendall–τ is sensitive to pairwise inversions, Spearman emphasizes larger displacements in rank and is sometimes more stable under near-ties. In either case, the subset selection goal is to make π_S a high-fidelity proxy for $\pi_{\mathcal{D}}$ at the fixed fidelity f.

Preview of the selection objective. With these definitions in place, the ranking-preserving coreset problem amounts to choosing $S \subseteq \mathcal{D}$, $|S| = m$, so that either (i) the induced ranking π_S is close to $\pi_{\mathcal{D}}$ on a specified finite architecture set (e.g., a probe set), or (ii) the score function $s_f(h; S)$ uniformly approximates $s_f(h; \mathcal{D})$ over a broader class $\mathcal{H}$, which in turn implies margin-stable rankings for all pairs with nontrivial full-data gaps. This dichotomy (finite-set ranking preservation versus uniform approximation) underlies the variants formalized next.

3 Problem Formulation: Ranking-Preserving Coreset Selection

We formalize *ranking-preserving coreset selection* as a constrained subset selection problem in which the downstream object of interest is not an absolute estimate of performance, but the relative ordering of architectures induced by proxy scores at a fixed fidelity. Throughout this section we take the evaluation protocol, proxy scores $s_f(\cdot;\cdot)$, margins $\Delta_f(\cdot,\cdot)$, and ranking distances (notably Kendall–τ) as defined in §2.

The RPCS objective on a finite target set. Let $A \subseteq \mathcal{H}$ be a finite set of architectures on which we desire ranking fidelity; in practice A is either (i) a *probe pool* P used to guide subset construction, or (ii) a set of candidates encountered during a NAS run. For a fixed fidelity f, the full evaluation split $\mathcal{D}$ induces a ranking $\pi_{\mathcal{D}}$ over A and any subset $S \subseteq \mathcal{D}$ induces π_S. The most direct formulation is

$$\min_{S \subseteq \mathcal{D}} \ \tau(\pi_S, \pi_{\mathcal{D}}) \qquad \text{subject to} \qquad |S| = m, \tag{6}$$

where τ counts pairwise inversions on A as in (4). Equivalently, since each inversion corresponds to a pair whose induced ordering flips, (6) can be written as maximizing pairwise sign agreement:

$$\max_{S \subseteq \mathcal{D}} \sum_{\substack{h,h' \in A \\ h \neq h'}} \mathbb{1}\Big[\operatorname{sign}\big(\widehat{\Delta}_f(h,h')\big) = \operatorname{sign}\big(\Delta_f(h,h')\big)\Big] \qquad \text{s.t.} \qquad |S| = m. \tag{7}$$

Both (6)–(7) express the same goal: the subset should preserve as many pairwise comparisons as possible among architectures of interest at the evaluation fidelity.

Probe-set RPCS versus NAS-time target sets. In general, the relevant A is not known in advance: a NAS algorithm adaptively proposes architectures based on past evaluations, so the set of visited candidates depends on the subset itself. To obtain a tractable and *offline* selection objective, we distinguish:

- **Probe-set RPCS.** Fix a probe pool $P = \{h_1, \ldots, h_M\} \subseteq \mathcal{H}$ and choose S to minimize $\tau(\pi_S, \pi_{\mathcal{D}})$ over P. This yields a well-posed finite objective and provides an explicit interface between subset selection and computation (we only need losses for architectures in P).
- **Candidate-set RPCS.** Let A denote the (random, adaptively generated) set of architectures evaluated during NAS. One may view the selection objective as minimizing $\mathbb{E}[\tau(\pi_S, \pi_{\mathcal{D}}) \mid A]$ over this random A, or controlling worst-case distortion over all size-L sets A. This motivates guarantees that hold *uniformly* over $\mathcal{H}$, discussed below.

Margin-aware surrogates for finite-set ranking fidelity. The indicator objectives in (7) are discontinuous and typically intractable for combinatorial search. We therefore consider margin-aware relaxations that emphasize preservation of *near-ties* (pairs with small $|\Delta_f|$), since large-margin pairs are intrinsically stable. A generic surrogate is a weighted, margin-robust hinge:

$$\min_{S\subseteq\mathcal{D}} \sum_{\substack{h,h'\in A\\ h\neq h'}} w(h,h')\cdot\max\Big\{0,\ \gamma-\operatorname{sign}\big(\Delta_f(h,h')\big)\,\widehat{\Delta}_f(h,h')\Big\} \qquad \text{s.t.} \quad |S|=m, \tag{8}$$

where $\gamma>0$ is a target margin and $w(h,h')\geq 0$ can be chosen to upweight pairs estimated to be ambiguous on $\mathcal{D}$ (for example, by using a cheap preliminary estimate of Δ_f on a small pilot sample). While (8) is still combinatorial in S, it admits useful relaxations and, more importantly for our purposes, it suggests which examples z are valuable: those for which per-example loss differences $\ell_f(h';z)-\ell_f(h;z)$ contribute substantially to stabilizing uncertain margins.

Uniform-over-$\mathcal{H}$ formulation (score approximation). A conceptually distinct variant is to require that the subset approximates full-data scores uniformly over the entire architecture class:

$$\min_{S\subseteq\mathcal{D}} \sup_{h\in\mathcal{H}} \big|s_f(h;S)-s_f(h;\mathcal{D})\big| \qquad \text{s.t.} \qquad |S|=m. \tag{9}$$

This formulation does not reference a particular finite set A and therefore aligns with the adaptive nature of NAS. Its utility is that a small value of (9) implies, deterministically, that *all* pairwise orderings with full-data gap exceeding 2ε are preserved (cf. the margin-stability principle developed in the next section). In this sense, (9) is a sufficient condition for small Kendall–τ distortion on any finite candidate set, with the distortion dominated by the number of near-tie pairs.

Fixed-fidelity versus multi-fidelity RPCS. The preceding objectives treat f as fixed. Many NAS procedures are, however, inherently multi-fidelity: they compare candidates at several fidelities (e.g., short training followed by longer training for finalists), or they use a fidelity schedule over time. To model this, let $\mathcal{F}=\{f_1,\dots,f_K\}$ be a set of fidelities that may be invoked during search. A *shared coreset* S is then required to preserve rankings across all fidelities in $\mathcal{F}$. Two natural formulations are:

$$\min_{S\subseteq\mathcal{D}} \max_{f\in\mathcal{F}} \tau\big(\pi_S^{(f)},\pi_{\mathcal{D}}^{(f)}\big) \qquad \text{s.t.} \quad |S|=m, \tag{10}$$

$$\min_{S\subseteq\mathcal{D}} \sum_{f\in\mathcal{F}} \alpha_f\cdot\sup_{h\in\mathcal{H}} \big|s_f(h;S)-s_f(h;\mathcal{D})\big| \qquad \text{s.t.} \quad |S|=m, \tag{11}$$

where $\pi_S^{(f)}$ denotes the ranking induced by $s_f(\cdot; S)$ and $\alpha_f \geq 0$ are weights reflecting how frequently each fidelity is used (or how strongly it influences the final selection). The worst-case objective (10) directly targets ranking distortion, while (11) extends the uniform approximation criterion across fidelities, enabling margin-based stability arguments fidelity-by-fidelity. One may also allow *fidelity-specific* subsets S_f with a total budget constraint $\sum_f |S_f| \leq m_{\text{tot}}$, but we focus on shared S since it yields a single evaluation set usable throughout NAS without bookkeeping.

Access model and the role of probe architectures. The formulations above are information-theoretic in the sense that they reference $s_f(h; \mathcal{D})$, which may be too expensive to compute for many h. In the selection stage we therefore restrict attention to a probe set P and to quantities computable from per-example probe losses $\ell_f(h; z)$ for $h \in P$, $z \in \mathcal{D}$. The algorithmic question becomes: using only this restricted access, can we choose S that approximately optimizes a probe-based analogue of (6) or (9), and which consequently yields low ranking distortion for architectures beyond P? The subsequent sections make this precise via margin stability (linking score deviation to ranking preservation) and via constructive objectives that are amenable to greedy optimization.

4 Ranking Preservation via Margin Stability

We now record a deterministic principle that links *score approximation* to *ranking preservation*. The role of this section is purely structural: it isolates a sufficient condition under which a subset S induces (almost) the same ordering as the full evaluation split $\mathcal{D}$. Probabilistic statements (i.e., when such a condition holds for a random or constructed S) are deferred to §5.

Pairwise margins and their subset estimates. Fix a fidelity f. For two architectures $h, h' \in \mathcal{H}$, we define the (full-data) pairwise margin

$$\Delta_f(h, h') \;:=\; s_f(h'; \mathcal{D}) - s_f(h; \mathcal{D}), \tag{12}$$

and, for a subset $S \subseteq \mathcal{D}$, the corresponding subset-based margin estimate

$$\widehat{\Delta}_f(h, h') \;:=\; s_f(h'; S) - s_f(h; S). \tag{13}$$

Since lower proxy score is better, the sign of $\Delta_f(h, h')$ encodes the ordering between h and h' induced by $\mathcal{D}$: specifically, $\Delta_f(h, h') > 0$ means that h is preferred to h' on $\mathcal{D}$. Ranking distortion therefore arises precisely when $\operatorname{sign}(\widehat{\Delta}_f(h, h')) \neq \operatorname{sign}(\Delta_f(h, h'))$ for some pair.

Uniform score approximation. The condition we analyze is uniform deviation of subset scores from full-data scores:

$$\sup_{h\in\mathcal{H}}\big|s_f(h;S)-s_f(h;\mathcal{D})\big| \;\le\; \varepsilon. \tag{14}$$

We emphasize that (14) is a property of the realized subset S; no randomness is assumed in the statement below. Moreover, although (14) is stated over $\mathcal{H}$, the same conclusions hold verbatim if we replace $\mathcal{H}$ by any finite target set $A \subseteq \mathcal{H}$ (e.g., a probe pool), with $\sup_{h\in A}$ in place of $\sup_{h\in\mathcal{H}}$.

Theorem 4.1 (Margin-stable ranking from uniform score approximation)**.** *Fix f and let $S \subseteq \mathcal{D}$ satisfy* (14)*. Then for any $h, h' \in \mathcal{H}$ with $|\Delta_f(h,h')| > 2\varepsilon$, we have*

$$\operatorname{sign}\big(\widehat{\Delta}_f(h,h')\big) \;=\; \operatorname{sign}\big(\Delta_f(h,h')\big). \tag{15}$$

Consequently, for any finite $A \subseteq \mathcal{H}$, every inversion between the rankings induced by S and by $\mathcal{D}$ over A must occur on a pair whose full-data margin magnitude is at most 2ε. In particular,

$$\tau(\pi_S,\pi_{\mathcal{D}}) \;\le\; \Big|\big\{\{h,h'\}\subseteq A:\; |\Delta_f(h,h')|\le 2\varepsilon\big\}\Big|, \tag{16}$$

where τ counts inversions on A (with any fixed tie-breaking rule if needed).

Proof. We first bound the margin estimation error by a direct triangle inequality. For any $h, h' \in \mathcal{H}$,

$$\begin{aligned}\big|\widehat{\Delta}_f(h,h')-\Delta_f(h,h')\big| &= \big|\big(s_f(h';S)-s_f(h;S)\big)-\big(s_f(h';\mathcal{D})-s_f(h;\mathcal{D})\big)\big| \\ &\le |s_f(h';S)-s_f(h';\mathcal{D})| + |s_f(h;S)-s_f(h;\mathcal{D})| \\ &\le 2\varepsilon, \end{aligned}\tag{17}$$

where the last step uses (14). Now suppose $|\Delta_f(h,h')| > 2\varepsilon$. Then (17) implies $\widehat{\Delta}_f(h,h')$ lies within a radius-2ε interval centered at $\Delta_f(h,h')$, which cannot cross zero. Hence $\widehat{\Delta}_f(h,h')$ and $\Delta_f(h,h')$ have the same sign, proving (15).

For the Kendall–τ statement, consider any unordered pair $\{h,h'\} \subseteq A$. An inversion between π_S and $\pi_{\mathcal{D}}$ on this pair can only happen if the induced ordering flips, which requires $\operatorname{sign}(\widehat{\Delta}_f(h,h')) \neq \operatorname{sign}(\Delta_f(h,h'))$ (modulo tie-breaking on exact zeros). By the first part, such a flip is impossible whenever $|\Delta_f(h,h')| > 2\varepsilon$. Therefore, only pairs with $|\Delta_f(h,h')| \le 2\varepsilon$ can contribute to $\tau(\pi_S,\pi_{\mathcal{D}})$, yielding (16). □

Interpretation: near-ties control ranking distortion. Theorem 4.1 formalizes an intuition used repeatedly in what follows. If S approximates scores uniformly within ε, then the only potentially unstable comparisons are

those whose true gaps are already small, i.e., near-ties under $\mathcal{D}$. Conversely, large-margin pairs are *automatically* robust to subset-induced noise. Thus, ranking preservation is naturally a margin-sensitive objective: the combinatorial difficulty of exactly preserving $\pi_{\mathcal{D}}$ is concentrated on architectures that are essentially indistinguishable at the chosen fidelity.

Finite-set specialization and approximate ordering. When our goal is ranking fidelity only on a finite A (e.g., a probe pool), it suffices to ensure

$$\max_{h \in A} \big| s_f(h; S) - s_f(h; \mathcal{D}) \big| \leq \varepsilon, \tag{18}$$

which implies the same pairwise and Kendall–τ conclusions over A. This specialization is relevant algorithmically because we may be able to control (18) using only losses of architectures in A, whereas controlling (14) over all $\mathcal{H}$ is inherently more demanding.

Bridge to sample complexity. The preceding arguments are deterministic: any mechanism (random sampling, leverage-style sampling, greedy submodular selection driven by probes) that yields small uniform deviation immediately yields margin-stable rankings. The remaining question is quantitative: how large must $m = |S|$ be to make (14) (or (18)) hold with high probability under a concrete selection procedure? In §5 we answer this via uniform convergence bounds in terms of the pseudo-dimension (or related capacity measures), and we combine them with Theorem 4.1 to obtain explicit ranking-preservation guarantees.

5 Upper Bounds: Uniform Convergence and Sample Complexity

We now give sufficient conditions on the subset size $m = |S|$ under which the uniform deviation

$$\sup_{h \in \mathcal{H}} \big| s_f(h; S) - s_f(h; \mathcal{D}) \big|$$

is small with high probability. Combined with the margin-stability principle of §4, these bounds yield explicit, margin-sensitive ranking guarantees.

A probabilistic model for subset formation. For the purposes of upper bounds we consider the simplest baseline: S is obtained by sampling m points i.i.d. uniformly from $\mathcal{D}$ (with replacement). This induces a standard empirical-process setting on the finite population $\mathcal{D}$: we may regard $\mathcal{D}$ as defining the uniform distribution over its elements, and $s_f(h; \mathcal{D})$ as the population mean of the bounded function $z \mapsto \ell_f(h; z)$, while $s_f(h; S)$

is the empirical mean over m draws. Sampling *without* replacement admits comparable bounds (often slightly sharper via finite-population corrections), and we omit these refinements since they do not change the scaling in d, ε, δ.

Capacity control via pseudo-dimension. Let $\mathcal{F} = \{z \mapsto \ell_f(h; z) : h \in \mathcal{H}\}$ denote the loss function class at fidelity f, and let d be its pseudo-dimension. The following theorem is a direct specialization of standard uniform convergence results for bounded real-valued function classes; we record it in our notation.

Theorem 5.1 (Uniform convergence upper bound)**.** *Assume $\ell_f(h; z) \in [0, 1]$ for all $h \in \mathcal{H}$ and $z \in \mathcal{D}$, and that $\mathcal{F}$ has pseudo-dimension d. Let S be formed by drawing m points i.i.d. uniformly from $\mathcal{D}$ (with replacement). There exists a universal constant $c > 0$ such that, for any $\varepsilon \in (0, 1)$ and $\delta \in (0, 1)$, if*

$$m \;\geq\; \frac{c}{\varepsilon^2}\Big(d \log \frac{1}{\varepsilon} + \log \frac{1}{\delta}\Big), \tag{19}$$

then with probability at least $1 - \delta$,

$$\sup_{h \in \mathcal{H}} \big| s_f(h; S) - s_f(h; \mathcal{D}) \big| \;\leq\; \varepsilon. \tag{20}$$

Proof approach (sketch). We interpret $s_f(h; S)$ as the empirical average of $f_h(z) := \ell_f(h; z)$ over m independent samples from the uniform distribution on $\mathcal{D}$. The bound (20) follows from symmetrization and a control of the uniform empirical process $\sup_h |\frac{1}{m} \sum_{i=1}^{m} (f_h(Z_i) - \mathbb{E} f_h)|$ via either (i) Rademacher complexity upper bounds combined with Dudley-type entropy integrals, or (ii) covering-number bounds for real-valued classes in terms of pseudo-dimension. The logarithmic factor $\log(1/\varepsilon)$ arises from metric-entropy control of $\mathcal{F}$ at scale ε; we do not attempt to optimize constants. Extensions to sub-Gaussian losses replace Hoeffding-type steps by Bernstein-type steps, yielding the same ε^{-2} scaling up to variance factors.

Finite target sets. When we only require uniform deviation over a finite target set $A \subseteq \mathcal{H}$ (e.g., a probe pool), a simpler argument suffices. For each fixed $h \in A$, Hoeffding's inequality gives

$$\Pr\Big(\big| s_f(h; S) - s_f(h; \mathcal{D}) \big| > \varepsilon \Big) \;\leq\; 2e^{-2m\varepsilon^2}.$$

A union bound over $h \in A$ yields

$$\Pr\Big(\max_{h \in A} \big| s_f(h; S) - s_f(h; \mathcal{D}) \big| > \varepsilon \Big) \;\leq\; 2|A| e^{-2m\varepsilon^2}, \tag{21}$$

so it suffices to take $m \gtrsim (\log(|A|/\delta))/\varepsilon^2$. Theorem 5.1 may be viewed as the infinite-class analogue where $|A|$ is replaced by an effective cardinality controlled by d.

Consequences for ranking preservation. Combining uniform deviation bounds with the deterministic margin-stability statement from §4 immediately yields a high-probability ranking guarantee. We state the implication in a form convenient for NAS, where one typically evaluates only a finite set of candidate architectures.

Corollary 5.2 (Ranking preservation sample complexity)**.** *Under the assumptions of Theorem 5.1, fix any finite $A \subseteq \mathcal{H}$. If m satisfies* (19)*, then with probability at least $1-\delta$ all pairwise orderings within A whose full-data margins exceed 2ε are preserved by the subset ranking induced by S. Moreover, the Kendall–τ distance between the rankings on A induced by S and by $\mathcal{D}$ is at most the number of unordered pairs $\{h, h'\} \subseteq A$ whose full-data margin magnitude is at most 2ε.*

Discussion: what the bound does and does not say. First, the rate $m = \tilde{O}(d/\varepsilon^2)$ is agnostic to the NAS procedure: it guarantees that *all* architectures in $\mathcal{H}$ have their proxy scores approximated simultaneously, hence it is deliberately worst-case. This is appropriate when the search algorithm adaptively explores architectures, since adaptivity can enlarge the effective set of queried models. Second, the corollary is margin-sensitive: if the full-data ranking on A has many pairs with tiny score gaps, then the Kendall–τ bound can be loose, reflecting genuine instability. Conversely, when the ranking has few near-ties at fidelity f, even moderate ε can suffice to preserve almost all pairwise comparisons.

Finally, we emphasize that (19) is an *existential* sufficient condition for random sampling. Constructive subset-selection procedures may exploit additional structure (e.g., redundancy in $\mathcal{D}$ relative to probe losses) to obtain smaller m in practice, but any such improvement must be understood relative to the information-theoretic limitations discussed next in §6.

6 Lower Bounds: Information-Theoretic Limitations

We complement the sufficient conditions of §5 by recalling that the scaling $m = \tilde{O}(d/\varepsilon^2)$ is, in general, unavoidable: without additional structure beyond boundedness and pseudo-dimension control, no algorithm can guarantee uniform approximation from substantially fewer than $\Theta(d/\varepsilon^2)$ examples. This justifies treating (19) as the correct worst-case baseline, and clarifies which parts of our ranking guarantees can and cannot be improved without extra assumptions (e.g., margin conditions, benign loss geometry, or restrictions to a finite target set).

A lower bound for uniform score approximation. Let $\mathcal{F} = \{z \mapsto \ell_f(h; z) : h \in \mathcal{H}\}$ be a bounded function class with pseudo-dimension d. Consider the task of producing, from m sampled examples, an empirical

mean that uniformly approximates the population mean over *all* $f \in \mathcal{F}$. The following theorem is a standard information-theoretic lower bound for uniform convergence in terms of pseudo-dimension; we state it in a form aligned with our notation.

Theorem 6.1 (Information-theoretic lower bound)**.** *Fix any $d \geq 1$ and $\varepsilon \in (0, 1/4)$. There exist a distribution P over examples z and a bounded function class $\mathcal{F} \subseteq [0,1]^{\mathcal{Z}}$ of pseudo-dimension d such that the following holds. Let $Z_1, \ldots, Z_m \overset{\text{i.i.d.}}{\sim} \mathsf{P}$, and let $S = \{Z_1, \ldots, Z_m\}$. Then for any (possibly randomized) procedure that outputs S of size m (equivalently, any estimator based on m samples), if*

$$m \;<\; c_0 \, \frac{d}{\varepsilon^2}$$

for a universal constant $c_0 > 0$, we have

$$\Pr\Big(\sup_{f \in \mathcal{F}} \Big| \mathbb{E}_{Z \sim \mathsf{P}}[f(Z)] - \frac{1}{m} \sum_{i=1}^{m} f(Z_i) \Big| > \varepsilon \Big) \;\geq\; \frac{1}{3}.$$

In particular, achieving $\sup_{f \in \mathcal{F}} |\mathbb{E} f - \widehat{\mathbb{E}} f| \leq \varepsilon$ with success probability at least $2/3$ requires $m = \Omega(d/\varepsilon^2)$.

Proof approach (sketch). We proceed by reduction to a multi-parameter mean-estimation problem on a shattered set. By pseudo-dimension d, there exist points $z_1, \ldots, z_d$ and thresholds $t_1, \ldots, t_d$ such that for every $\sigma \in \{0,1\}^d$ there exists $f_\sigma \in \mathcal{F}$ with $f_\sigma(z_i) \leq t_i$ when $\sigma_i = 0$ and $f_\sigma(z_i) > t_i$ when $\sigma_i = 1$. One then defines a family of nearby distributions $\{\mathsf{P}_\theta : \theta \in \{\pm 1\}^d\}$ supported on $\{z_1, \ldots, z_d\}$ whose means differ in d independent directions by magnitude on the order of ε. The key step is that, for each coordinate i, distinguishing whether the mean in direction i is $+\varepsilon$ or $-\varepsilon$ requires $\Omega(1/\varepsilon^2)$ samples by classical two-point (Le Cam) or bounded-variance testing bounds. Aggregating across d coordinates via Assouad's lemma or Fano-type arguments yields an $\Omega(d/\varepsilon^2)$ sample requirement for uniformly controlling deviations over the whole family $\{f_\sigma\}$ simultaneously. We omit the constant-optimization details, as the argument is now classical in empirical process theory and minimax lower bounds for VC/pseudo-dimension classes.

Interpretation for subset selection. Theorem 6.1 is best read as a statement about *information*: if the only guarantee we demand is worst-case uniform approximation over a class of complexity d, then a budget of $m = o(d/\varepsilon^2)$ examples cannot, in general, resolve which of exponentially many functions (or architectures) is being evaluated closely enough to ensure small uniform error. In our setting, this means that any method claiming

uniform approximation of $s_f(h; \mathcal{D})$ for all $h \in \mathcal{H}$ from a subset S must either (i) use m on the order of d/ε^2, or (ii) exploit additional structure not captured by pseudo-dimension alone (e.g., restrictions to a small target set, strong margins, or low effective dimension of the realized loss vectors on $\mathcal{D}$).

Consequences for ranking preservation and near-ties. Recall from §4 that preserving a pairwise ordering between h and h' is ensured whenever the full-data margin satisfies $|\Delta_f(h, h')| > 2\varepsilon$ and we have uniform score deviation at most ε. The lower bound therefore has an immediate (and unavoidable) implication: without at least $\Omega(d/\varepsilon^2)$ examples, there exist problems for which *some* architecture will have its score mis-estimated by more than ε, and hence any ranking guarantee that relies on controlling such deviations must fail for sufficiently small margins.

More concretely, one may construct adversarial collections of architectures $\{h_1, \ldots, h_N\}$ whose losses correspond to a shattered set, so that many pairwise score gaps on the full distribution are only $\Theta(\varepsilon)$. In such a regime, even an oracle subset cannot robustly identify the true ordering: the true ranking contains $\Theta(N^2)$ comparisons with margin comparable to the unavoidable estimation noise, and Theorem 6.1 implies that with $m = o(d/\varepsilon^2)$ there is nontrivial probability of flipping at least one such comparison. This phenomenon is not an artifact of our analysis: when margins are of the same scale as estimation error, the ranking problem is statistically ill-posed, and Kendall–τ distortion cannot be uniformly controlled.

What can still be improved. The lower bound is a worst-case statement over $\mathcal{F}$ and P. It does *not* preclude smaller subsets in favorable instances relevant to NAS. In particular, if we restrict attention to a finite set $A \subseteq \mathcal{H}$ (e.g., a probe pool) then the effective complexity becomes $\log |A|$ rather than d, and correspondingly one can hope for $m = O((\log |A|)/\varepsilon^2)$-type behavior (cf. (21)). Similarly, if the full-data ranking exhibits large margins for most pairs at fidelity f, then the required ε for low Kendall–τ distortion may be relatively coarse. These are precisely the regimes in which constructive, probe-guided subset selection can outperform naive random sampling in practice, even though the information-theoretic worst-case barrier remains in force.

7 Constructive Algorithms: Probe-Guided Subset Selection

We now describe practical procedures for constructing a subset $S \subseteq \mathcal{D}$ intended to preserve architecture rankings at a fixed fidelity f while reducing the number of evaluated examples during NAS. The methods below share a common template: we (i) instantiate a small probe pool $P = \{h_1, \ldots, h_M\} \subset$

$\mathcal{H}$, (ii) measure per-example probe losses, and (iii) select m examples so that the subset score statistics match (or at least stabilize) the pairwise comparisons that drive Kendall–τ distortion.

Probe loss embeddings and pairwise features. After training each $h_i \in P$ under fidelity f on a fixed training split, we evaluate its loss on each example $z \in \mathcal{D}$ and form the probe loss vector

$$v(z) \;=\; \big(\ell_f(h_1; z), \ldots, \ell_f(h_M; z)\big) \in [0,1]^M.$$

The ranking over P is determined by the probe scores $s_f(h_i; \mathcal{D}) = \frac{1}{n}\sum_{z\in\mathcal{D}} v_i(z)$, hence by the aggregated pairwise margins

$$\Delta_f(h_i, h_j) \;=\; \frac{1}{n}\sum_{z\in\mathcal{D}} \big(v_j(z) - v_i(z)\big).$$

It is therefore natural to treat each example as contributing a signed "vote" to each ordered pair (i, j) via the pairwise-difference feature

$$d_{ij}(z) \;=\; v_j(z) - v_i(z) \in [-1, 1].$$

Selecting S corresponds to approximating these aggregated statistics by $\frac{1}{m}\sum_{z\in S} d_{ij}(z)$ (possibly after reweighting). In particular, if we can ensure $|\widehat{\Delta}_f(h_i, h_j) - \Delta_f(h_i, h_j)|$ is small for most relevant pairs, then the ranking over P becomes margin-stable.

A monotone submodular surrogate and greedy selection. Directly optimizing Kendall–τ or pairwise sign agreement over P is combinatorial and, in general, computationally intractable. We therefore optimize a surrogate that (a) is sensitive to pairwise margins, (b) is monotone and admits submodularity under mild design choices, and (c) can be maximized by greedy selection with approximation guarantees.

Let $\mathcal{Q} \subseteq [M] \times [M]$ be a chosen set of ordered pairs (often a sparsified set, e.g., "near ties" estimated from a small pilot). For weights $w_{ij} \geq 0$ and a concave nondecreasing map $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ (e.g., $\phi(u) = \min\{u, T\}$ or $\phi(u) = \sqrt{u + \eta}$), define the set function

$$F(S) \;=\; \sum_{(i,j)\in\mathcal{Q}} w_{ij}\, \phi\left(\left|\sum_{z\in S} d_{ij}(z)\right|\right). \tag{22}$$

Intuitively, the inner sum aggregates evidence for the direction of the (i, j) comparison, while concavity enforces diminishing returns: once a comparison is "decided" with large magnitude, additional examples contribute less. When ϕ is concave and the statistics are additive, F is a standard instance of a monotone submodular objective (a concave-over-modular construction

after splitting absolute values into positive/negative parts), and the greedy algorithm yields a $(1 - 1/e)$ approximation to $\max_{|S|=m} F(S)$.

We implement selection by iterating $t = 1, \ldots, m$ and adding the example with the largest marginal gain

$$\Delta_F(z \mid S) \;=\; F(S \cup \{z\}) - F(S).$$

The marginal gain can be computed from maintained pairwise accumulators $A_{ij}(S) = \sum_{z \in S} d_{ij}(z)$, since

$$\Delta_F(z \mid S) \;=\; \sum_{(i,j) \in \mathcal{Q}} w_{ij} \Big[\phi\big(|A_{ij}(S) + d_{ij}(z)|\big) - \phi\big(|A_{ij}(S)|\big) \Big].$$

Two practical refinements are routinely beneficial: (i) *pair sparsification*, restricting $\mathcal{Q}$ to pairs whose full-data margins appear small under a cheap estimate (to concentrate capacity on potentially flippable comparisons), and (ii) *near-tie weighting*, setting w_{ij} larger when the estimated $|\Delta_f(h_i, h_j)|$ is small, since such pairs contribute most to Kendall–τ distortion under bounded estimation error.

A streaming/online variant. When n is large, storing all $v(z)$ or repeatedly scanning $\mathcal{D}$ during greedy selection may be undesirable. We can instead select S in a single pass using streaming submodular maximization. The key observation is that the sufficient statistics for (22) are the accumulators $A_{ij}(S)$, which require only $O(|\mathcal{Q}|)$ memory, whereas examples can be processed sequentially.

One option is to use a thresholded acceptance rule: maintain a current set S and, upon seeing z, compute $\Delta_F(z \mid S)$; if $|S| < m$ and $\Delta_F(z \mid S)$ exceeds a time-varying threshold, insert z, otherwise discard it. More robustly, one may apply standard streaming algorithms for monotone submodular maximization under a cardinality constraint (e.g., sieve-style methods) that maintain multiple candidate thresholds and guarantee a constant-factor approximation to the optimal value of F using $O(m \log U)$ space for an appropriate value range U. In our setting, the per-item update cost is dominated by computing $d_{ij}(z)$ for $(i, j) \in \mathcal{Q}$, which is feasible provided $|\mathcal{Q}|$ is not too large (or can be factored, as discussed below). The streaming construction yields an "anytime" subset that can be stopped early when the NAS budget dictates.

Alternative solvers: DPP and leverage-style sampling. The preceding greedy methods are tailored to pairwise ranking preservation. Nevertheless, it is useful to relate them to more classical coreset selection tools, since these can be competitive under additional structure.

DPP-style diversity sampling. If we represent each example by a feature vector $\psi(z) \in \mathbb{R}^p$ (for instance, $\psi(z) = v(z)$, or a concatenation of

selected $d_{ij}(z)$), a determinantal point process (DPP) with kernel $K_{zz'} = \langle \psi(z), \psi(z') \rangle$ favors diverse subsets that span the feature space. In regimes where ranking errors are driven by missing directions in the probe-loss geometry (e.g., clustered, redundant examples), DPPs can approximate the effect of greedy coverage. However, DPPs optimize a diversity likelihood rather than a margin-sensitive objective; they align best with our theory when the dominant error mode is *linear* and can be controlled by capturing the principal subspace of the probe features.

Leverage-score and subspace-embedding sampling. Let $\Psi \in \mathbb{R}^{n \times p}$ stack rows $\psi(z)^\top$. If the relevant loss evaluations can be modeled as approximately linear functionals over $\psi(z)$ (e.g., $z \mapsto \ell_f(h; z) \approx a(h)^\top \psi(z)$ for architectures of interest), then selecting rows by leverage scores and reweighting yields a subspace embedding: empirical averages over the coreset approximate full averages uniformly over the span. In such low effective-rank regimes (rank $r \ll p$), one can obtain guarantees with $m = \tilde{O}(r/\varepsilon^2)$ rather than depending on the ambient pseudo-dimension, thereby matching the intuition that NAS losses may concentrate on a low-dimensional manifold at fixed fidelity. This approach is most principled when probe features are expressive and the architecture set being ranked remains close to the probe-induced span; otherwise, margin-sensitive surrogates such as (22) are preferable.

Finally, all methods above admit an iterative refinement: once NAS begins, architectures encountered during search can be appended to P, updating $\mathcal{Q}$ and reselecting (or augmenting) S to focus the subset on the comparisons that are actually queried. This shifts effort from worst-case uniformity toward the realized region of $\mathcal{H}$ explored by the search procedure.

Complexity analysis. We decompose the cost of probe-guided subset selection into (i) training the probe pool P, (ii) evaluating probe losses on $\mathcal{D}$, and (iii) executing the discrete selection rule (greedy or streaming). We express complexity in terms of dataset size $n = |\mathcal{D}|$, probe size $M = |P|$, subset size $m = |S|$, and a fidelity-dependent cost model.

Fidelity-dependent primitives. Let $C_{\text{train}}(f; h)$ denote the cost of training an architecture h under fidelity f on the fixed training split (e.g., t steps at fixed resolution), and let $c_{\text{eval}}(f; h)$ denote the per-example cost of computing $\ell_f(h; z)$ (typically a single forward pass at the evaluation resolution). For coarse accounting we write

$$C_{\text{train}}(f) \;=\; \max_{h \in P} C_{\text{train}}(f; h), \qquad c_{\text{eval}}(f) \;=\; \max_{h \in P} c_{\text{eval}}(f; h),$$

noting that in practice the dependence on h can be reduced by constraining P to a homogeneous family (e.g., similar widths) or by normalizing evaluation costs.

Cost of probe loss acquisition. To form the probe loss embedding $v(z) = (\ell_f(h_1; z), \ldots, \ell_f(h_M; z))$ for all $z \in \mathcal{D}$, we first train each $h_i \in P$ once and then evaluate it on the n examples. The leading-order cost is therefore

$$\underbrace{\sum_{i=1}^{M} C_{\text{train}}(f; h_i)}_{\text{probe training}} + \underbrace{\sum_{i=1}^{M} \sum_{z \in \mathcal{D}} c_{\text{eval}}(f; h_i)}_{\text{probe evaluation}} \approx M\, C_{\text{train}}(f) + M\, n\, c_{\text{eval}}(f).$$

This stage determines the number of passes over $\mathcal{D}$: if we evaluate probes sequentially, we make M passes but with a simple access pattern; if we evaluate all probes in one pass via batched inference, we effectively make one pass while paying similar arithmetic cost (the difference is I/O and caching). When f is low (few steps, reduced resolution), $C_{\text{train}}(f)$ and $c_{\text{eval}}(f)$ are small and this stage is typically amortized across many downstream NAS evaluations.

Naive greedy selection and its reduction. Assume we optimize (22) over a pair set $\mathcal{Q} \subseteq [M] \times [M]$ with $|\mathcal{Q}| = Q$. Given maintained accumulators $A_{ij}(S) = \sum_{z \in S} d_{ij}(z)$, the marginal gain for a candidate z requires $O(Q)$ evaluations of $\phi(\cdot)$ and arithmetic on $A_{ij}(S)$. A naive greedy implementation that, at each of m rounds, scans all remaining elements and computes $\Delta_F(z \mid S)$ thus costs

$$O\big(m\, n\, Q\big) \quad \text{after probe losses are available.}$$

In the fully dense case $Q = M(M-1)$ this becomes $O(m\, n\, M^2)$ and is unacceptable unless M is very small. Two reductions are standard. First, *pair sparsification* chooses $\mathcal{Q}$ to contain only potentially flippable comparisons, for example those with small estimated full-data margins $|\Delta_f(h_i, h_j)|$, so that $Q \ll M^2$ and the dominant work becomes $O(m\, n\, Q)$. Second, one may enforce a *factorized* surrogate: if we replace explicit pair features by a low-rank proxy (e.g., using centered probe vectors and a small number of principal directions), then marginal gains can be computed in $O(M)$ or $O(r)$ time per element (for rank r), giving $O(m\, n\, M)$ or $O(m\, n\, r)$ selection time. These reductions trade exact pairwise control for computational feasibility, and are justified when the ranking-relevant geometry of $v(z)$ is low-dimensional or when only near-ties materially affect Kendall–τ.

Streaming and pass complexity. Greedy selection as stated is multi-pass if we cannot store all $v(z)$: each of the m rounds requires scanning $\mathcal{D}$. In contrast, streaming monotone submodular maximization can be implemented in a single pass after (or during) probe evaluation. In sieve-style variants, we maintain $O(\log U)$ candidate solutions (for a value range U)

and update all candidates upon seeing z. The per-item update cost remains $O(Q \log U)$, yielding overall time $O(n\,Q \log U)$ and memory $O(m \log U + Q)$. The constant-factor approximation of the streaming solver is weaker than the $(1 - 1/e)$ greedy guarantee for F, but the reduction in passes over $\mathcal{D}$ is decisive when dataset I/O dominates.

Space requirements. If we cache probe loss vectors, the memory is $O(nM)$, which is often too large for n in the millions. A streaming computation of $d_{ij}(z)$ avoids this: we store only the current accumulators $A_{ij}(S)$ (cost $O(Q)$) plus the selected subset indices (cost $O(m)$). If we also wish to compute or update weights w_{ij} based on pilot estimates, we may store an additional $O(Q)$ table. Hence the working memory can be reduced to $O(Q + m)$, at the expense of either multiple passes (for greedy) or a weaker streaming approximation.

End-to-end compute trade-offs and break-even. The purpose of selecting S is to reduce the total evaluation cost of NAS. Suppose the search procedure evaluates L candidate architectures at fidelity f and uses the proxy score $s_f(\cdot;\cdot)$ as its feedback. Full-data evaluation costs approximately $L\,n\,c_{\text{eval}}^{\text{cand}}(f)$, whereas subset evaluation costs $L\,m\,c_{\text{eval}}^{\text{cand}}(f)$, where $c_{\text{eval}}^{\text{cand}}(f)$ is the per-example evaluation cost for candidates (often comparable to $c_{\text{eval}}(f)$). The net savings therefore scale like $(n-m)/n$, but we must subtract the one-time subset-selection overhead. A crude break-even condition is

$$L\,(n-m)\,c_{\text{eval}}^{\text{cand}}(f) \;\gtrsim\; M\,n\,c_{\text{eval}}(f) \;+\; M\,C_{\text{train}}(f) \;+\; \text{SelCost}(n, m, M, Q),$$

where $\text{SelCost}(n, m, M, Q)$ is $O(mnQ)$ for naive greedy or $O(nQ \log U)$ for streaming. Thus, for large L (typical in global NAS) the overhead is amortized, whereas for small L (small-batch ablations) uniform sampling or very small M, Q are preferable.

Ranking guarantees versus compute. The theoretical sufficient condition for margin-stable rankings is a uniform deviation bound $\sup_{h \in \mathcal{H}} |s_f(h; S) - s_f(h; \mathcal{D})| \le \varepsilon$, which, even under i.i.d. sampling, typically requires $m = \tilde{O}((d + \log(1/\delta))/\varepsilon^2)$. Increasing m tightens the bound and reduces the count of potentially flippable pairs (those with margin $\le 2\varepsilon$), but the downstream NAS cost scales linearly in m. Probe-guided selection attempts to move along this frontier: for fixed m, we spend additional one-time cost proportional to M and Q to reduce ranking distortion on a relevant set (at minimum, on P), thereby improving NAS outcomes without increasing per-candidate evaluation cost. The appropriate operating point is therefore problem-dependent: we choose m to satisfy the NAS budget, and then select the largest M and richest $\mathcal{Q}$ whose overhead remains negligible compared to the total number of candidate evaluations L expected in the search.

8 Experimental Protocol

We evaluate subset selection for NAS-time ranking preservation along two axes: (i) controlled benchmarks in which the "full-data" ranking is either known exactly or can be exhaustively computed, and (ii) large-scale regimes in which full evaluation is expensive and the practical criterion is NAS outcome quality per unit compute. Throughout, we distinguish the *selection fidelity* (used to compute probe losses and construct S) from the *assessment fidelity* (used to approximate the ground-truth ranking over architectures). When the benchmark admits exhaustive evaluation, the assessment fidelity is simply "full benchmark evaluation"; otherwise we approximate it by a substantially higher fidelity than used during selection.

Search spaces and datasets. We consider two families of spaces.

- *Tabular NAS-Bench-style micro spaces.* We instantiate experiments on NAS-Bench-101 and NAS-Bench-201–type cell spaces, where architectures can be enumerated and their accuracies at several training budgets are available. In these settings, the dataset $\mathcal{D}$ is the standard benchmark dataset (e.g., CIFAR-10/100, ImageNet-16-120), and the "architecture evaluation" is a lookup at the benchmark-reported fidelity. This provides an oracle for $\pi_{\mathcal{D}}$ over a large finite set, enabling direct measurement of ranking distortion.

- *Global macro/micro spaces.* We additionally consider a larger, non-tabular regime in which architectures are instantiated and trained (e.g., a MobileNet/ResNet-like macro space with width/depth/kernel choices, and optionally a micro-cell component). Here, $\mathcal{H}$ is implicit and the full-data proxy score $s_f(h; \mathcal{D})$ is estimated by actual training at a fixed recipe. This regime is used both on ImageNet proxies and on a face-recognition dataset in the scale range of CASIA-WebFace.

For ImageNet-scale experiments we report results on (a) ImageNet-100 (a class subset), and (b) a resolution-reduced or sample-reduced proxy (e.g., $224 \to 160$ or a fixed fraction of images) in which full-data evaluation remains feasible for assessment. For the face-recognition regime we follow the standard verification protocol (e.g., LFW-style evaluation) but use the training loss as ℓ_f and report downstream verification/identification metrics for the final selected architecture; class imbalance is handled by stratified sampling in baselines and by reporting subset class histograms for all methods.

Fidelity specification and "ground truth" rankings. In all non-tabular experiments we define two fidelities f_{low} and f_{high}. The low fidelity is used for probe training and candidate scoring during NAS on S (e.g., few epochs/steps, reduced resolution, light augmentation), while the high fidelity

is used only to estimate $\pi_{\mathcal{D}}$ on a reference set of architectures (e.g., longer training, standard resolution, full augmentation). When the benchmark is tabular, f_{low} and f_{high} correspond to two budget entries. We explicitly verify that f_{high} induces a more stable ranking than f_{low} by measuring rank correlations between intermediate budgets when available.

Probe pool construction and loss acquisition. We choose a probe pool $P = \{h_1, \ldots, h_M\}$ to reflect the local geometry of the search space while remaining inexpensive. In micro spaces we sample probes uniformly from the tabular space; in global spaces we sample from the same generator used by the NAS method (e.g., the initial population for evolutionary search or random architectures for one-shot-free baselines). To reduce variance, we either (i) fix a single training seed for all probes and treat randomness as part of the access model, or (ii) average $\ell_f(h_i; z)$ over a small number of seeds for the probes only; we report which choice is used and keep it consistent across methods. Probe evaluation produces the vectors $v(z) \in \mathbb{R}^M$ needed by the selection rule.

Methods and baselines. We compare our probe-guided selection (RPCS-Greedy and, when I/O constrained, a streaming variant) to the following baselines at matched subset size m:

- *Uniform* sampling of m examples (optionally stratified by class).
- *Loss-only* heuristics that select examples with large mean probe loss or large variance across probes, thereby ignoring pairwise ranking structure.
- *Training-coreset baselines* (e.g., herding/gradient matching style) applied to probe gradients where feasible; these optimize training fidelity rather than ranking preservation and serve to separate objectives.

We also include ablations of the surrogate: choice of ϕ (clipped absolute value versus smooth concave alternatives), weight definitions w_{ij} (uniform versus near-tie emphasis), and pair sparsification level Q (dense, top-Q near-ties, or low-rank proxies).

Ranking-distortion metrics. On a finite evaluation set $A \subset \mathcal{H}$ (tabular: typically the full benchmark; non-tabular: a large held-out set of architectures not used in P), we compute:

- Kendall–τ distance $\tau(\pi_S, \pi_{\mathcal{D}})$ induced by $s_f(\cdot; S)$ versus $s_f(\cdot; \mathcal{D})$, and its normalized variant.
- *Pairwise sign accuracy* $\frac{1}{|A|(|A|-1)} \sum_{h \neq h'} \mathbb{1}\Big[\operatorname{sign}(\widehat{\Delta}_f(h, h')) = \operatorname{sign}(\Delta_f(h, h'))\Big]$.

- *Top-k overlap* between the best k architectures under π_S and $\pi_{\mathcal{D}}$, for multiple k (e.g., $k \in \{1, 5, 10, 50\}$), which is more directly tied to NAS outcomes.

When f_{high} is used as assessment, we compute these metrics with $\pi_{\mathcal{D}}$ replaced by the ranking induced by $s_{f_{\text{high}}}(\cdot; \mathcal{D})$, and we report the remaining "fidelity gap" by also reporting the correlation between f_{low} and f_{high} on full data.

NAS outcome metrics and compute accounting. To connect ranking preservation to practical search, we run a fixed NAS algorithm (e.g., evolutionary search, random search with successive halving, or a standard global search loop) using proxy evaluations on S only. We report: (i) the final architecture's performance when retrained at f_{high} (or full training), (ii) the *simple regret* relative to the best found under full-data evaluation at matched search budget (when feasible), and (iii) the stability across seeds. Compute is reported in two forms: (a) the number of example-evaluations $L \cdot m$ versus $L \cdot n$, and (b) measured wall-clock or GPU-hours including one-time selection overhead. We ensure fair comparisons by fixing L (number of candidate evaluations) and varying only m and the selection method; in addition, we include a compute-matched setting in which methods are given equal total budget (selection + NAS), which may favor simpler selectors when L is small.

Scaling protocol and reporting. For each dataset/space we sweep m/n over a logarithmic grid (e.g., $0.1\%, 0.3\%, 1\%, 3\%, 10\%$) and probe sizes M over a modest range (e.g., $M \in \{16, 32, 64, 128\}$), and we repeat each configuration over multiple random seeds affecting P, subset sampling (where applicable), and NAS stochasticity. We report mean $\pm$ standard error for ranking metrics and NAS outcomes, and we include diagnostic plots of margin distributions $|\Delta_f(h, h')|$ (estimated on a large architecture sample) to contextualize the observed Kendall–τ in light of the margin-dependent guarantees. This protocol isolates three effects: the statistical effect of increasing m, the algorithmic effect of probe-guided selection at fixed m, and the systems effect of selection overhead amortization as L grows.

9 Discussion and Limitations

Margin dependence and what our bounds do (and do not) guarantee. Our ranking-preservation statements are intrinsically *margin dependent*. Thm. 1 shows that, under the uniform score approximation condition

$$\sup_{h \in \mathcal{H}} \left| s_f(h; S) - s_f(h; \mathcal{D}) \right| \leq \varepsilon,$$

only those pairs (h, h') with full-data gap $|\Delta_f(h, h')| \leq 2\varepsilon$ can have their relative order flipped by replacing $\mathcal{D}$ with S. Consequently, if the search space exhibits many near-ties at the selection fidelity (i.e., the empirical distribution of $|\Delta_f(h, h')|$ is concentrated near 0 on the set of architectures under consideration), then no method can promise small Kendall–τ distortion without correspondingly small ε, hence without a larger subset size m (cf. Thm. 2 and the lower bound Thm. 3). In this sense, our guarantees should be interpreted as *conditional stability*: when the ranking is well-separated, modest m suffices; when it is ill-separated, the ranking is statistically fragile and the natural target should be weakened (e.g., top-k identification, or preservation of a coarsened ordering with indifference bands).

A practical corollary of the same phenomenon is that the relevant margins are not uniform across the space. We may only care about preserving the ordering among a set A of architectures visited by the NAS procedure, or among the top portion of the ranking. This suggests adapting both the theoretical lens and the algorithmic surrogate: rather than attempting to minimize $\tau(\pi_S, \pi_{\mathcal{D}})$ over a large A, one may emphasize pairs estimated to be near the decision boundary (via weights w_{ij}) or restrict attention to a dynamically maintained candidate pool. Formally, one can replace uniform convergence over $\mathcal{H}$ with guarantees over a data-dependent finite set A (via union bounds) or via localized complexity measures; we do not develop these refinements here.

Stochastic training, random seeds, and augmentations. Our notation treats $\ell_f(h; z)$ as a deterministic per-example quantity obtained after training h under fidelity f. In non-tabular regimes this is an idealization: SGD noise, data-order randomness, and stochastic augmentations can induce variability comparable to the margins we seek to preserve. If we denote by ω all training-time randomness, the more faithful object is $\ell_f(h; z, \omega)$ and the score is either a conditional realization $s_f(h; S, \omega)$ or the expectation $\mathbb{E}_\omega[s_f(h; S, \omega)]$. The uniform deviation requirement in Thm. 1 can fail if the randomness-induced variance dominates $|\Delta_f(h, h')|$, even when m is large. Two mitigations are immediate: (i) fix ω for all evaluations in both selection and NAS-time scoring (which makes the access model deterministic but ties the ranking to a particular stochastic instance), or (ii) average over a small number of seeds for probes and/or candidate evaluations, effectively replacing ℓ_f by a lower-variance estimator. The latter increases compute but can be targeted: one may average only for near-tie pairs, consistent with a margin-aware strategy.

Stochastic augmentations introduce a second subtlety: if augmentations depend on the example z in a structured way, the induced losses can change the *relative* importance of examples for ranking preservation. Conceptually, this suggests defining the selection objective with respect to the augmen-

tation distribution used during NAS-time evaluation, and treating $\ell_f(h; z)$ as an expectation over augmentation draws. Algorithmically, it motivates storing or sampling augmentation seeds consistently during probe-loss acquisition so that the vectors $v(z)$ capture the same effective evaluation distribution as the subsequent NAS loop.

Probe-set bias and generalization beyond P. Our constructive method relies on a finite probe pool P to define surrogate statistics (e.g., pairwise differences $d_{ij}(z)$) and to drive the greedy selection. This introduces an unavoidable bias: the subset S is optimized to preserve rankings *as seen through* P, and may fail to preserve rankings for architectures outside P, particularly if P does not cover the modes of the search distribution. Theorems such as Thm. 5 provide guarantees only over P (or over sets A for which one can verify margin conditions and approximation quality). In practice one may reduce this gap by (i) sampling P from the same generator used by the NAS algorithm, (ii) updating P online with architectures encountered during search, and (iii) regularizing the surrogate toward more uniform coverage (e.g., via diversity constraints on P). A principled analysis of adaptive probe updates, where P depends on previously selected subsets and observed losses, remains open; it would likely require tools from adaptive data analysis rather than classical uniform convergence.

Hardware-aware ranking and multi-objective NAS. Many NAS settings rank architectures by objectives that combine accuracy with hardware metrics (latency, energy, memory footprint) or by explicit constraints. Our framework can accommodate such extensions insofar as the objective induces a total preorder. For example, for a scalarized objective

$$\widetilde{s}(h; A) \;=\; s_f(h; A) \;+\; \lambda \cdot c(h),$$

where $c(h)$ is a deterministic hardware cost independent of z, subset selection affects only the data-dependent term $s_f(h; A)$, and the same deviation bounds apply verbatim to $\widetilde{s}$. More delicate is the constrained case (e.g., minimize loss subject to $c(h) \le C$), where the induced ranking depends on feasibility. Here, small perturbations in s_f can change which feasible architecture is optimal, even if pairwise margins among feasible models are large, because the active set can shift. One approach is to incorporate feasibility into the probe pool by restricting P to cost-feasible architectures and evaluating ranking preservation only within that slice. Another is to treat constraint violation as an additional loss term that is example-independent, again reducing to scalarization.

Truly multi-objective NAS (Pareto ranking) requires a partial order. One can still define a pairwise relation, e.g. $h \prec h'$ if h' dominates h in all objectives, and aim to preserve dominance relations under subsampling. The

pairwise-margin machinery then applies to each objective separately, but the interaction between objectives complicates both the surrogate and the notion of Kendall–τ; developing a submodular surrogate that targets Pareto-front stability is an open direction.

Systems limitations and open problems. Finally, our method shifts compute from NAS-time evaluation to one-time selection overhead. When L (the number of candidates evaluated during NAS) is small, amortization can fail and uniform sampling may be competitive. Moreover, storing probe-loss vectors $v(z) \in \mathbb{R}^M$ can be I/O bound for large n and moderate M, and the naive greedy update can be expensive without sparsifying pairs or using streaming approximations. A deeper open problem is to obtain *end-to-end* guarantees that couple (a) the selection procedure, (b) the NAS algorithm's adaptivity, and (c) fidelity mismatch between f_{low} and f_{high}, yielding a bound on final regret under realistic stochastic training. At present, our theoretical results isolate the ranking-preservation component; closing this loop would require a joint analysis of optimization dynamics and statistical approximation, and it is unclear whether worst-case bounds of useful magnitude are attainable without additional structural assumptions on the search space and training recipe.