Near-Optimal Label Complexity for Active Huber Regression

Abstract

We resolve, up to logarithmic factors, the Huber dimension-dependence question posed by Musco, Musco, Woodruff, and Yasuda [14], reducing the rank exponent from $4 - 2\sqrt{2}$ to one. We study fixed-design active regression in the exact-real label-query model: the feature matrix $A \in \mathbb{R}^{n \times d}$ is known, while an arbitrary response vector is accessible only through coordinate queries. For $r = \operatorname{rank}(A)$, our randomized nonadaptive algorithm queries $\min\{n, \widetilde{O}(r/\varepsilon^2)\}$ labels and, with probability at least 0.99, returns a $(1+\varepsilon)$-approximation to the square root of the empirical Huber loss. All query locations are fixed before any label is read. For every admissible pool size and rank, and $0 < \varepsilon \le 1/800$, we construct a fixed rank-r design on which every adaptive randomized algorithm with the same worst-case guarantee needs $\Omega(\min\{n, r/\varepsilon^2\})$ labels in worst-case expectation. Thus label complexity is jointly optimal in pool size, rank, and accuracy up to logarithmic factors in this range. The guarantee requires no distributional or noise assumptions and includes zero-optimum instances. The key ingredient is a centered affine transfer lemma: design-only multiscale scores preserve the objective increments needed to locate the minimizer after an independent constant-factor pilot. This overcomes the hidden-label dependence of standard affine sparsification and turns near-linear design sparsification into a near-optimal active algorithm.

Note. This paper was generated entirely by AI, including MiMo, using an automated research pipeline developed by Chenghua Liu and Hanyu Li.

1 Introduction

The Huber loss is quadratic for small residuals and linear for large ones, combining the local efficiency of least squares with bounded influence in the tails [9]. It is consequently a standard robust surrogate in statistics and machine learning. In large pools of unlabeled examples, however, observing a response may itself be the dominant cost: a label can require a human annotation, a physical experiment, or an expensive simulation. This motivates *active regression*: the feature matrix is available in full, but the algorithm should fit the responses after inspecting as few of them as possible.

We study the strongest transductive version of this question. The design $A \in \mathbb{R}^{n \times d}$ is arbitrary and known, the response $b \in \mathbb{R}^n$ is arbitrary and hidden, and one query reveals one coordinate of b. There is no distributional model, no stochastic-noise assumption, and no promise that the optimum is bounded away from zero. The target is a relative approximation to the empirical Huber objective on the entire fixed pool. This worst-case formulation sharply separates label complexity from the statistical literature on Huber regression, where samples are drawn from a population and robustness is measured through estimation or prediction risk [18].

For squared loss, geometric sampling gives label complexity linear in the effective dimension for constant accuracy [4]. Lewis-weight methods likewise give nearly tight guarantees for least absolute deviations [2, 15]. Huber regression is harder because it is not homogeneous: the same direction can behave quadratically at one radius and linearly at another. Musco, Musco, Woodruff, and Yasuda developed a general active-regression framework. After the standard reduction to the column space, their Huber specialization uses $\widetilde{O}(r^{4-2\sqrt{2}}\,\mathrm{poly}(1/\varepsilon))$ queries, where $r = \operatorname{rank}(A)$ [14]. Their

dimension-reduction discussion motivates asking whether the exponent of this effective dimension can be reduced to one.

At first sight, the near-linear unshifted Huber sparsifier of Jambulapati et al. [12] appears to settle the question. It does not. A standard affine reduction replaces each row by $(a_i, -b_i)$, so the importance scores of the resulting sparsifier inspect precisely the labels that an active algorithm is trying not to query. Nor can one hope for a label-independent strong affine coreset for every response vector: an unqueried coordinate can hide an arbitrarily large constant term. The algorithm must preserve less than the whole objective, but enough to locate its minimizer.

Our results. We give a nonadaptive, label-oblivious algorithm with query complexity

$$\min\{n, \widetilde{O}(r/\varepsilon^2)\}, \qquad r = \operatorname{rank}(A), \tag{1.1}$$

and constant success probability. After the same column-space reduction, the advance is the linear, rather than $r^{4-2\sqrt{2}}$, rank dependence together with the joint dependence on n and ε. The sampling locations are fixed before any response is read. Our lower bound holds for the same approximation criterion even against adaptive randomized algorithms: for every admissible triple (n, d, r) and $0 < \varepsilon \le 1/800$, some fixed rank-$r$ design requires

$$\Omega\Big(\min\Big\{n, \frac{r}{\varepsilon^2}\Big\}\Big) \tag{1.2}$$

queries in worst-case expectation. Hence, in this small-error range, (1.1) is optimal simultaneously in sample size, rank, and accuracy up to logarithmic factors. In particular, the full-query cap is not merely an artifact of the algorithm: when the pool is too small to support r/ε^2 informative responses, a fixed positive fraction of the entire pool can be necessary.

The technical contribution behind the upper bound is a centered affine transfer principle. From A alone we construct one multiscale score vector of total mass $O(r \log n)$. For any fixed residual z, sampling by these scores uniformly preserves the increments of the translated objective $F_z(h) := \sum_i H(z_i + a_i^{\mathsf{T}} h)$ over the relevant unshifted Huber ball. The baseline $F_z(0)$, which may contain unseen and arbitrarily large labels, cancels. An independent pilot supplies a constant-factor center, and the centered estimate then localizes its own unconstrained minimizer. The complete query set is chosen before any label is read.

The pilot, residualization, and localization steps originate in Musco et al. [14]; the multiscale unshifted sparsifier comes from Jambulapati et al. [12]; and the common-basepoint chaining principle is already present in Jambulapati et al. [11]. Our use of these ingredients is specific: the design-only multiscale scores control centered affine increments after an independent pilot, avoiding label-dependent augmented rows. We do not claim that centering, localization, or Huber sparsification is new in isolation.

The lower bound has two regimes. When n/r is large, independent biased-Bernoulli blocks of size N and bias $\beta = \max\{200\varepsilon, \sqrt{10/N}\}$ encode r hidden signs. A relative Huber solution decodes almost all signs, while each adaptively selected response carries only $O(\beta^2) = O(\varepsilon^2 + 1/N)$ information. When n/r is small, a realizable instance has zero optimum, so a multiplicative guarantee forces exact recovery of $r = \Omega(n)$ independent labels. This second regime is what turns the earlier fixed-row-count construction into the all-n minimax statement.

Organization. Section 2 formalizes the model and the minimax result. Section 3 develops the design-only scores, centered transfer, and two-batch upper bound. Section 4 proves the adaptive lower bound. Section 5 places the result in the active-regression and sparsification literatures and records its scope. Technical constructions and auxiliary calculations appear after the references.

2 Problem, model, and main results

This section fixes the loss, query model, and approximation convention before stating the two minimax bounds. It also records the width scaling used to reduce the analysis to the width-one loss.

For the width-one Huber loss, write

$$H(t) := \begin{cases} t^2/2, & |t| \le 1, \\ |t| - 1/2, & |t| > 1, \end{cases} \qquad F_b(x) := \sum_{i=1}^{n} H(a_i^\mathsf{T} x - b_i), \tag{2.1}$$

where a_i^T is row i of A. Following the active regression literature, we call $F_b(x)^{1/2}$ the Huber norm objective, although this terminology is only shorthand for the square root of the loss. The algorithm knows A in advance and learns b_i only by querying coordinate i. It may perform unrestricted computation on A and on queried labels; in particular, the results below concern label complexity, not total running time, and permit exact minimization of finite convex Huber objectives.

A query count always means the number of distinct response coordinates inspected: repeated requests are answered from a cache, and a run that never returns has query count $Q = \infty$. Absolute constants below are universal unless a dependence is displayed, and $[k] := \{1, \dots, k\}$ for a positive integer k. A $(1+\varepsilon)$-approximation to $F_b^{1/2}$ is equivalently a $(1+\varepsilon)^2$-approximation to F_b.

After rank reduction, Theorem 1.5 of Musco et al. [14] uses $\widetilde{O}(r^{4-2\sqrt{2}}\,\mathrm{poly}(1/\varepsilon))$ queries. We ask for a design-only, preferably nonadaptive query set with linear rank exponent and the joint (n, r, ε) guarantee

$$F_b(\widehat{x})^{1/2} \le (1+\varepsilon) \min_{x \in \mathbb{R}^d} F_b(x)^{1/2}. \tag{2.2}$$

The first result supplies such a nonadaptive set; its explicit logarithms concern label complexity and do not imply a running-time bound.

Theorem 2.1 (Nonadaptive upper bound)**.** *Let $A \in \mathbb{R}^{n \times d}$, let $r = \mathrm{rank}(A)$, and let $0 < \varepsilon < 1/2$. There is a randomized algorithm such that the following holds for every fixed $b \in \mathbb{R}^n$. Its queried coordinates are fixed before any label is read and depend only on A, ε, and private randomness. It queries at most*

$$\min\left\{ n, Cr\left(\log^6(2n) + \frac{\log^4(2n)}{\varepsilon^2} \right) \right\} \tag{2.3}$$

distinct coordinates of b and, with probability at least 0.99 over its private randomness, returns $\widehat{x}$ satisfying

$$F_b(\widehat{x})^{1/2} \le (1+\varepsilon) \min_{x \in \mathbb{R}^d} F_b(x)^{1/2}.$$

In the small-error range below, the upper bound is optimal up to logarithmic factors throughout the (n, d, r) parameter range, even if adaptivity is allowed.

Theorem 2.2 (Matching adaptive lower bound)**.** *There is an absolute constant $c > 0$ such that the following holds. For all integers n, d, r with $1 \le r \le \min\{n, d\}$ and every $0 < \varepsilon \le 1/800$, there is a fixed matrix $A \in \mathbb{R}^{n \times d}$ of rank exactly r such that every adaptive randomized algorithm that, for every fixed $b \in \mathbb{R}^n$, returns $\widehat{x}$ satisfying $F_b(\widehat{x})^{1/2} \le (1+\varepsilon) \min_x F_b(x)^{1/2}$ with probability at least 0.99 obeys*

$$\sup_{b \in \mathbb{R}^n} \mathbb{E}_{\mathcal{R}} Q(A, b) \ge c \min\left\{ n, \frac{r}{\varepsilon^2} \right\}. \tag{2.4}$$

Here Q may be a stopping-time query count and $\mathcal{R}$ is the algorithm's private seed. Specifically, Q counts distinct queried coordinates. Repeated requests are permitted but can be answered from the algorithm's cache and are not charged twice; set $Q = \infty$ on a run that never returns an output.

Combining the two bounds gives the minimax statement in the common accuracy range. The response remains worst case, while the expectation below is over the algorithm's private randomness.

Corollary 2.3 (Minimax label complexity)**.** *Fix $0 < \varepsilon \le 1/800$ and $1 \le r \le \min\{n, d\}$. Let $\mathfrak{A}$ range over randomized algorithms that, for every fixed pair (A, b), achieve the $(1+\varepsilon)$-approximation to $F_b^{1/2}$ in (2.2) with probability at least 0.99, and define*

$$\mathcal{Q}_{n,d,r}(\varepsilon) := \inf_{\mathfrak{A}} \sup_{\substack{A \in \mathbb{R}^{n\times d}:\, \operatorname{rank}(A)=r \\ b \in \mathbb{R}^n}} \mathbb{E}_{\mathcal{R}} Q_{\mathfrak{A}}(A, b), \tag{2.5}$$

where the expectation is over all algorithmic randomness. For absolute constants $c, C > 0$,

$$\begin{aligned} c \min\left\{n, \frac{r}{\varepsilon^2}\right\} &\le \mathcal{Q}_{n,d,r}(\varepsilon) \\ &\le \min\left\{n, Cr\left(\log^6(2n) + \frac{\log^4(2n)}{\varepsilon^2}\right)\right\}. \end{aligned} \tag{2.6}$$

The upper algorithm is nonadaptive, whereas the lower bound holds against adaptive algorithms. Thus the minimax label complexity is $\widetilde{\Theta}(\min\{n, r/\varepsilon^2\})$.

Proof. Apply Theorem 2.2; for the reverse inequality use Theorem 2.1 and its full-query branch, then take the indicated infimum and supremum. □

For a width parameter $s > 0$, we use the fixed-tail-slope convention

$$H_s(t) := sH(t/s) = \begin{cases} t^2/(2s), & |t| \le s, \\ |t| - s/2, & |t| > s. \end{cases} \tag{2.7}$$

If $F_{s,A,b}(x) := \sum_i H_s(a_i^\mathsf{T} x - b_i)$, then

$$F_{s,A,b}(x) = sF_{1,A/s,b/s}(x).$$

Thus applying Theorem 2.1 to $(A/s, b/s)$, or multiplying both the rows and labels in the hard instance of Theorem 2.2 by s, proves the same approximation and query bounds for H_s.

Proof mechanism. One $O(r \log n)$-mass multiscale score vector controls unshifted Huber balls; after an independent pilot, the same scores preserve centered affine increments and localize the sampled minimizer. The lower bound uses noiseless recovery or biased-Bernoulli blocks with bounded information per queried label.

3 The nonadaptive upper bound

The upper bound has three components. We construct one design-only score vector for every unshifted radius, transfer it to centered affine increments, and combine that transfer with an independent pilot in a two-batch algorithm. The full proof is given here; the deterministic weight construction, endpoint algebra, and attainment details appear in Appendix A.

3.1 Unshifted geometry and label-oblivious scores

We begin with the metric-like square root $\psi_H(t) := \sqrt{H(t)}$. Its triangle and increment inequalities drive both the score construction and the later localization argument.

Lemma 3.1 (Huber triangle and increment inequalities)**.** *For every $s, t \in \mathbb{R}$,*

$$\psi_H(s+t) \le \psi_H(s) + \psi_H(t), \tag{3.1}$$

$$|\psi_H(s) - \psi_H(t)| \le \psi_H(s-t), \tag{3.2}$$

$$|H(s) - H(t)| \le \sqrt{H(s-t)}\big(\sqrt{H(s)} + \sqrt{H(t)}\big). \tag{3.3}$$

Consequently, for nonnegative weights w_i and vectors y, z of the same finite dimension,

$$\left(\sum_i w_i H(y_i + z_i)\right)^{1/2} \le \left(\sum_i w_i H(y_i)\right)^{1/2} + \left(\sum_i w_i H(z_i)\right)^{1/2}. \tag{3.4}$$

Proof. On $[0, \infty)$,

$$\psi_H(u) = \begin{cases} u/\sqrt{2}, & 0 \le u \le 1, \\ \sqrt{u - 1/2}, & u \ge 1 \end{cases}$$

is nonnegative, increasing, concave, and zero at the origin. Its increments are therefore decreasing, so $\psi_H(u+v) - \psi_H(u) \le \psi_H(v) - \psi_H(0)$ for $u, v \ge 0$. Together with $|s+t| \le |s| + |t|$, this proves (3.1). Apply that inequality to $s = t + (s-t)$, and then exchange s, t, to obtain (3.2). Factoring

$$|H(s) - H(t)| = |\psi_H(s) - \psi_H(t)| \, \big(\psi_H(s) + \psi_H(t)\big)$$

gives (3.3). Finally, apply (3.1) coordinatewise and use the weighted Euclidean triangle inequality to obtain (3.4). □

For the upper-bound lemmas, take a matrix $A \in \mathbb{R}^{N \times r}$ with full column rank and nonzero rows a_i^T. Define the unshifted objective and its sublevel sets by

$$\Phi(h) := \sum_{i=1}^N H(a_i^\mathsf{T} h), \qquad K_R := \{h \in \mathbb{R}^r : \Phi(h) \le R\}. \tag{3.5}$$

Because $c_A := \min_{\|h\|_2 = 1} \|Ah\|_1 > 0$ and $H(t) \ge |t| - 1/2$,

$$\Phi(h) \ge c_A \|h\|_2 - N/2. \tag{3.6}$$

Thus every K_R is compact, and the suprema of the continuous finite-sample processes below are finite and measurable.

For a metric space $(\mathcal{T}, d)$, let $\gamma_2(\mathcal{T}, d)$ denote Talagrand's chaining functional

$$\inf_{(\mathcal{A}_k)} \sup_{t \in \mathcal{T}} \sum_{k \ge 0} 2^{k/2} \operatorname{diam}_d(\mathcal{A}_k(t)), \tag{3.7}$$

where $(\mathcal{A}_k)_{k \ge 0}$ ranges over nested partitions of $\mathcal{T}$, $\mathcal{A}_k(t)$ is the cell containing t, $|\mathcal{A}_0| = 1$, and $|\mathcal{A}_k| \le 2^{2^k}$ [19].

We state explicitly the external covering fact used at the transition scales. For $w \in \mathbb{R}^N_{>0}$, write $W = \operatorname{diag}(w)$ and $v_i(w)^2 := a_i^\mathsf{T}(A^\mathsf{T} W A)^{-1} a_i$.

Lemma 3.2 (Imported multiscale covering primitive)**.** *Fix $\alpha \geq 1$, and let $\mathcal{I} \subset \mathbb{Z}$ be a finite contiguous interval with $|\mathcal{I}| \geq 4\log(2N)$. Suppose that $\{w^{(j)} : j \in \mathcal{I}\}$ is an α-approximate Huber weight scheme, meaning that for every i and $j \in \mathcal{I}$,*

$$\frac{2^j}{\alpha} \leq \frac{H(v_i(w^{(j)}))}{w_i^{(j)} v_i(w^{(j)})^2} \leq \alpha 2^j, \tag{3.8}$$

and $w_i^{(j+1)} \leq \alpha w_i^{(j)}$ whenever $j, j+1 \in \mathcal{I}$. Put

$$\lambda_i^{(j)} := w_i^{(j)} v_i(w^{(j)})^2, \qquad \tau_i^{\mathcal{I}} := \max_{j \in \mathcal{I}} \lambda_i^{(j)}, \qquad k_{\mathcal{I}} := \max \mathcal{I} + 1. \tag{3.9}$$

Define the associated deterministic metric by

$$d_{\tau^{\mathcal{I}}}(h,k) := \max_{i \leq N} \sqrt{\frac{H(a_i^{\mathsf{T}}(h-k))}{\tau_i^{\mathcal{I}}}}. \tag{3.10}$$

Then $\sum_i \lambda_i^{(j)} = r$, and for a constant C_α depending only on α,

$$\gamma_2(K_{2^{k_{\mathcal{I}}}}, d_{\tau^{\mathcal{I}}}) \leq C_\alpha 2^{k_{\mathcal{I}}/2} \log^{3/2}(2N). \tag{3.11}$$

There is a universal numerical α_0 for which such a scheme has a deterministic finite construction from $(A, \mathcal{I})$ in the exact-real model.

This is the specialization of the weight-scheme covering argument of Jambulapati et al. [12, Definition 1.8 and proof of Theorem 3.14] to $f_i = H$; the cited bibliography entry identifies the arXiv full version in which those items appear. The assumptions hold because H is even, $\psi_H = \sqrt{H}$ is even and one-auto-Lipschitz by (3.2), and ψ_H is 1-auto-Lipschitz and has lower-1/2, upper-1 scaling while H has lower-1, upper-2 scaling: for $\lambda \geq 1$,

$$\lambda H(t) \leq H(\lambda t) \leq \lambda^2 H(t). \tag{3.12}$$

The exact construction and its stopping criterion are recorded in Appendix A.1; no bit-complexity or running-time conclusion is used here.

The smallest and largest Huber radii reduce to ℓ_2^2 and ℓ_1 geometry. To avoid conflating Lewis scores with the auxiliary diagonal weights, for $p \in \{1, 2\}$ let $\ell^{(p)} \in \mathbb{R}^N_{>0}$ satisfy

$$\ell_i^{(p)} = \left(a_i^{\mathsf{T}} (A^{\mathsf{T}} L_p^{1-2/p} A)^{-1} a_i\right)^{p/2}, \qquad L_p := \operatorname{diag}(\ell^{(p)}). \tag{3.13}$$

Set $w_i^{(p)} := (\ell_i^{(p)})^{1-2/p}$ and $(v_i^{(p)})^2 := a_i^{\mathsf{T}} (A^{\mathsf{T}} \operatorname{diag}(w^{(p)}) A)^{-1} a_i$. Then the weighted leverage is

$$\lambda_i^{(p)} := w_i^{(p)} (v_i^{(p)})^2 = (v_i^{(p)})^p = \ell_i^{(p)}. \tag{3.14}$$

Thus $w^{(2)} \equiv 1$ and $\lambda^{(2)}$ is ordinary leverage, whereas $\lambda^{(1)}$ is the ℓ_1 Lewis score.

Lemma 3.3 (Endpoint coverings)**.** *For $p \in \{1, 2\}$, the scores in (3.14) have total mass r. With*

$$d_p(h,k) := \max_{i \leq N} \frac{\left|a_i^{\mathsf{T}}(h-k)\right|^{p/2}}{\sqrt{\lambda_i^{(p)}}}, \tag{3.15}$$

they satisfy, for every $S > 0$,

$$\gamma_2\big(\{h : \|Ah\|_p^p \leq S\}, d_p\big) \leq C\sqrt{S}\, \log^{3/2}(2N). \tag{3.16}$$

The proof verifies an exact scale scheme using $w_i^{(j)} = 2^{-2j/p} w_i^{(p)}$; see Appendix A.2. We now combine these endpoint scores with a finite transition-scale scheme.

Lemma 3.4 (All-radius label-oblivious Huber scores)**.** *There are scores $\tau_i = \tau_i(A) > 0$, computable from A alone, such that*

$$T := \sum_{i=1}^{N} \tau_i \le Cr \log(2N) \tag{3.17}$$

and, for

$$d_\tau(h,k) := \max_{i \le N} \sqrt{\frac{H(a_i^\mathsf{T}(h-k))}{\tau_i}}, \tag{3.18}$$

one has, simultaneously for every $R > 0$,

$$\gamma_2(K_R, d_\tau) \le C\sqrt{R}\, \log^{3/2}(2N). \tag{3.19}$$

Proof. Let

$$L := \lceil 4 \log(2N) \rceil, \qquad \mathcal{J} := [-2 - L,\, 2 + \lceil \log_2(2N) \rceil] \cap \mathbb{Z},$$

and construct the transition-scale schemes of Lemma 3.2 on $\mathcal{J}$, denoting their weighted leverages by $\lambda_{i,\mathrm{tr}}^{(j)}$. For $p = 1, 2$, let $c \ge 1$ be a fixed universal constant and let $\widetilde{\lambda}_i^{(p)}$ be a deterministic finite constant-factor approximation to the exact endpoint score in Lemma 3.3, with $\widetilde{\lambda}_i^{(p)} \in [\lambda_i^{(p)}/c, c\lambda_i^{(p)}]$, and set $\overline{\lambda}_i^{(p)} := c\widetilde{\lambda}_i^{(p)}$. Such approximations follow from the standard Lewis-weight construction [5]; for $p = 2$, ordinary leverage can instead be used exactly. Define

$$\tau_i^{\mathrm{tr}} := \max_{j \in \mathcal{J}} \lambda_{i,\mathrm{tr}}^{(j)}, \qquad \tau_i := \tau_i^{\mathrm{tr}} + \overline{\lambda}_i^{(1)} + \overline{\lambda}_i^{(2)}.$$

Each inflated endpoint vector dominates its exact counterpart and has mass at most $c^2 r$. Since every transition leverage vector has mass r,

$$\sum_i \tau_i \le |\mathcal{J}|\; r + 2c^2 r = O(r \log(2N)).$$

Suppose first that $1/2 \le R \le N/2$, let $k_R := \lceil \log_2 R \rceil$, and put $\mathcal{I}_R := \{k_R - L, \dots, k_R - 1\}$. Then $\mathcal{I}_R \subseteq \mathcal{J}$, $K_R \subseteq K_{2^{k_R}}$, and $d_\tau \le d_{\tau^{\mathcal{I}_R}}$. Therefore Lemma 3.2 and $2^{k_R} < 2R$ give

$$\gamma_2(K_R, d_\tau) \le \gamma_2(K_{2^{k_R}}, d_{\tau^{\mathcal{I}_R}}) \le C\sqrt{R} \log^{3/2}(2N).$$

If $R \le 1/2$, every coordinate on K_R lies in the quadratic region and

$$K_R = \{h : \|Ah\|_2^2 \le 2R\}, \qquad d_\tau \le 2^{-1/2} d_2.$$

If $R \ge N/2$, the bound $H(t) \ge |t| - 1/2$ gives

$$K_R \subseteq \{h : \|Ah\|_1 \le R + N/2\} \subseteq \{h : \|Ah\|_1 \le 2R\}, \qquad d_\tau \le d_1,$$

where the metric inequality uses $H(t) \le |t|$. Applying Lemma 3.3 in both cases completes the proof. □

Set $\pi_i^{\mathrm{main}} := \tau_i/T$, draw $\nu_1, \dots, \nu_m$ independently from π^{main}, and define

$$\widehat{\Phi}(h) := \frac{1}{m} \sum_{j=1}^{m} \frac{H(a_{\nu_j}^{\mathsf{T}} h)}{\pi_{\nu_j}^{\mathrm{main}}}. \tag{3.20}$$

Sampling from the all-radius scores yields the unshifted estimate that will control the random envelope in the affine transfer.

Lemma 3.5 (Unshifted expectation bound)**.** *There is an absolute constant C such that, whenever*

$$\alpha := C\sqrt{\frac{T \log^3(2N)}{m}} \le 1, \tag{3.21}$$

then, for every $R > 0$,

$$\mathbb{E} \sup_{h \in K_R} \left| \widehat{\Phi}(h) - \Phi(h) \right| \le 2\alpha R. \tag{3.22}$$

In particular,

$$\mathbb{E} \sup_{h \in K_R} \widehat{\Phi}(h) \le 3R. \tag{3.23}$$

Proof. Let the left side of (3.22) be X. Ghost-sample symmetrization gives

$$X \le 2\mathbb{E}_{\nu,\sigma} \sup_{h \in K_R} \left| \frac{1}{m} \sum_{j=1}^{m} \sigma_j \frac{H(a_{\nu_j}^{\mathsf{T}} h)}{\pi_{\nu_j}^{\mathrm{main}}} \right|,$$

where the σ_j's are independent Rademacher signs. Conditional on the indices, the increment metric is

$$d_\nu(h,k)^2 := \sum_{j=1}^{m} \frac{[H(a_{\nu_j}^{\mathsf{T}} h) - H(a_{\nu_j}^{\mathsf{T}} k)]^2}{m^2 (\pi_{\nu_j}^{\mathrm{main}})^2}.$$

Put $L_\nu := \sup_{u \in K_R} \widehat{\Phi}(u)$. By (3.3), $\pi_i^{\mathrm{main}} = \tau_i/T$, and (3.18),

$$d_\nu(h,k)^2 \le \frac{2T}{m} d_\tau(h,k)^2 \big(\widehat{\Phi}(h) + \widehat{\Phi}(k)\big) \le \frac{4TL_\nu}{m} d_\tau(h,k)^2.$$

The signed process is anchored at zero and has subgaussian increments in d_ν. The standard generic-chaining bound, followed by Lemma 3.4, therefore yields

$$\mathbb{E}_\sigma \sup_{h \in K_R} |Z_h| \le C\gamma_2(K_R, d_\nu) \le C\sqrt{\frac{T \log^3(2N)}{m}} \sqrt{RL_\nu},$$

where $Z_h = m^{-1} \sum_j \sigma_j H(a_{\nu_j}^{\mathsf{T}} h)/\pi_{\nu_j}^{\mathrm{main}}$. Taking expectation over the indices and using Jensen gives

$$X \le \alpha\sqrt{R}\, \mathbb{E}\sqrt{L_\nu} \le \alpha\sqrt{R}\sqrt{R + X},$$

because $L_\nu \le R + \sup_{K_R} \left|\widehat{\Phi} - \Phi\right|$. For $R > 0$, writing $y = X/R$ gives $y^2 \le \alpha^2(1 + y)$; its positive root is at most 2α when $\alpha \le 1$. The case $R = 0$ is immediate. Consequently $X \le 2\alpha R$ and $\mathbb{E}L_\nu \le R + X \le 3R$. □

3.2 Centered affine preservation

The preceding scores control an unshifted empirical process. We now show that the same distribution preserves centered increments around any fixed residual, which is precisely what the second batch will require.

Fix an arbitrary residual vector $z \in \mathbb{R}^N$, and define

$$F_z(h) := \sum_{i=1}^{N} H(z_i + a_i^\mathsf{T} h), \qquad R := F_z(0) = \sum_{i=1}^{N} H(z_i). \tag{3.24}$$

Using a fresh sample from π^{main}, define

$$\widehat{F}_z(h) := \frac{1}{m} \sum_{j=1}^{m} \frac{H(z_{\nu_j} + a_{\nu_j}^\mathsf{T} h)}{\pi^{\mathrm{main}}_{\nu_j}}. \tag{3.25}$$

Only sampled residual coordinates are needed. The same indices define $\widehat{F}_z$ and $\widehat{\Phi}$; this coupling is used below.

Lemma 3.6 (Centered affine transfer)**.** *Fix $B \geq 1$, $0 < \eta, \delta < 1/2$, and z with $R = F_z(0) > 0$. There is a constant $\overline{C}_B = O(B^2)$, depending only on B, such that if*

$$m \geq \overline{C}_B \frac{T \log^3(2N)}{\eta^2 \delta^2}, \tag{3.26}$$

then, with probability at least $1 - \delta$ over the fresh main sample,

$$\sup_{h \in K_{BR}} \left| [\widehat{F}_z(h) - \widehat{F}_z(0)] - [F_z(h) - F_z(0)] \right| \leq \eta R. \tag{3.27}$$

The distribution depends on A, but not on z.

Proof. Set $\varphi_i(h) := H(z_i + a_i^\mathsf{T} h) - H(z_i)$. Since $\varphi_i(0) = 0$, signed symmetrization gives

$$\begin{aligned} &\mathbb{E} \sup_{h \in K_{BR}} \left| \frac{1}{m} \sum_{j=1}^{m} \frac{\varphi_{\nu_j}(h)}{\pi^{\mathrm{main}}_{\nu_j}} - \sum_{i=1}^{N} \varphi_i(h) \right| \\ &\qquad \leq 2\mathbb{E}_{\nu,\sigma} \sup_{h \in K_{BR}} \left| \frac{1}{m} \sum_{j=1}^{m} \sigma_j \frac{\varphi_{\nu_j}(h)}{\pi^{\mathrm{main}}_{\nu_j}} \right|. \end{aligned}$$

Conditional on the sampled indices, let $d_{\nu,z}$ be the increment metric of this signed process and put $L_{\nu,z} := \sup_{u \in K_{BR}} \widehat{F}_z(u)$. Exactly as in the unshifted calculation,

$$d_{\nu,z}(h,k)^2 \leq \frac{2T}{m} d_\tau(h,k)^2 \big(\widehat{F}_z(h) + \widehat{F}_z(k)\big) \leq \frac{4TL_{\nu,z}}{m} d_\tau(h,k)^2.$$

The shared-sample weighted triangle inequality gives, pointwise,

$$\widehat{F}_z(u) \leq 2\widehat{F}_z(0) + 2\widehat{\Phi}(u).$$

After increasing the sample constant so that Lemma 3.5 applies on K_{BR}, unbiasedness and (3.23) yield

$$\mathbb{E} L_{\nu,z} \leq (2 + 6B)R.$$

Generic chaining and Lemma 3.4, followed by Jensen, now bound the expected centered error by

$$D_B R\sqrt{\frac{T\log^3(2N)}{m}}, \qquad D_B := C\sqrt{B(2+6B)} = O(B).$$

Choose $\overline{C}_B$ at least D_B^2 times a sufficiently large universal constant, also large enough for the unshifted condition $\alpha \le 1$. Under (3.26), the last display is at most $\delta\eta R$. Markov's inequality gives (3.27). □

Remark 3.7 (Why centering matters). The conclusion is additive, localized, and centered. It does not claim that $\widehat{F}_z$ multiplicatively preserves the full affine objective. An unseen residual may make the common baseline arbitrarily large, but it cannot destroy preservation of the increments needed for optimization.

3.3 Pilot, algorithm, and localization

We next obtain the center needed by the transfer lemma and then state the full nonadaptive procedure. The pilot adapts the constant-factor mechanism of Musco et al. [14, Lemma 5.1], instantiated with the all-radius A-only embedding above and with the finite-scale bookkeeping made explicit.

The finite-range pilot estimate must be extended without using the false implication $H(t) \le 1 \Rightarrow |t| \le 1$. The following Huber-specific lemma uses the sharp quadratic threshold $1/2$.

Lemma 3.8 (Huber all-scale extension)**.** *Let $N > 16$, $S_0 := 8N^3$, and $\zeta := 1/16$. Suppose*

$$G(h) := \sum_{i=1}^{N} \omega_i H(a_i^\mathsf{T} h), \qquad \omega_i \ge 0, \qquad \sum_i \omega_i \le 2N. \tag{3.28}$$

If $|G(h) - \Phi(h)| \le \zeta\Phi(h)$ whenever $\Phi(h) \in [1/2, S_0]$, then

$$|G(h) - \Phi(h)| \le 2\zeta\Phi(h) \qquad \text{for every } h. \tag{3.29}$$

Proof. If $\Phi(h) = 0$, then every $a_i^\mathsf{T} h = 0$, and hence $G(h) = 0$. Suppose $0 < \Phi(h) < 1/2$, and set $\beta := \sqrt{(1/2)/\Phi(h)} > 1$. Since $\beta^2 H(a_i^\mathsf{T} h) \le 1/2$ for every i, both $a_i^\mathsf{T} h$ and $\beta a_i^\mathsf{T} h$ lie in the quadratic region. Consequently

$$\Phi(\beta h) = \beta^2\Phi(h) = 1/2, \qquad G(\beta h) = \beta^2 G(h),$$

and the assumed relative error at βh scales back to h.

It remains to consider $\Phi(h) > S_0$. Put

$$c := \frac{S_0}{2\Phi(h)}, \qquad y := ch, \qquad \beta := c^{-1} > 1.$$

For every $t \in \mathbb{R}$ and $\beta \ge 1$, direct inspection of the two Huber pieces gives

$$0 \le H(\beta t) - \beta H(t) \le \frac{\beta - 1}{2}.$$

Consequently, using $\sum_i \omega_i \le 2N$ for the bound on G,

$$\beta\Phi(y) \le \Phi(h) \le \beta\Phi(y) + \frac{N}{2}(\beta - 1),$$
$$\beta G(y) \le G(h) \le \beta G(y) + N(\beta - 1).$$

It follows that

$$\frac{S_0 - N}{2} \le \Phi(y) \le \frac{S_0}{2},$$

so $\Phi(y) \in [1/2, S_0]$. Applying the assumed estimate at y, and comparing the two additive remainders, gives

$$|G(h) - \Phi(h)| \le \zeta\beta\Phi(y) + N(\beta - 1) \le \left(\zeta + \frac{2N}{S_0}\right)\Phi(h) \le 2\zeta\Phi(h),$$

because $S_0 = 8N^3$, $N > 16$, and $\zeta = 1/16$. □

Fix $\delta_0 := 10^{-3}$, $\rho := 1/8$, and $\zeta := 1/16$. Regularize the main distribution by

$$\pi_i^{\mathrm{pil}} := \frac{\pi_i^{\mathrm{main}}}{2} + \frac{1}{2N} = \frac{\tau_i + T/N}{2T}, \qquad T' := 2T. \tag{3.30}$$

Let

$$\mathcal{S} := \{2^j : -1 \le j \le \lceil \log_2(8N^3) \rceil\}, \qquad J := |\mathcal{S}|, \tag{3.31}$$

and choose a universal C_{pil} large enough for Lemma 3.5. Set

$$m_0 := \left\lceil C_{\mathrm{pil}} \frac{T' \log^3(2N) J^2}{\zeta^2 \delta_0^2} \right\rceil = O\big(r \log^6(2N)\big). \tag{3.32}$$

Lemma 3.9 (Label-oblivious constant-factor pilot)**.** *Fix $b \in \mathbb{R}^N$ and set $F_b(x) := \sum_{i=1}^N H(a_i^\mathsf{T} x - b_i)$. For independent $\mu_1, \dots, \mu_{m_0} \sim \pi^{\mathrm{pil}}$, define*

$$\widetilde{\Phi}_0(h) := \frac{1}{m_0} \sum_{j=1}^{m_0} \frac{H(a_{\mu_j}^\mathsf{T} h)}{\pi_{\mu_j}^{\mathrm{pil}}}, \tag{3.33}$$

$$\widetilde{F}_0(x) := \frac{1}{m_0} \sum_{j=1}^{m_0} \frac{H(a_{\mu_j}^\mathsf{T} x - b_{\mu_j})}{\pi_{\mu_j}^{\mathrm{pil}}}. \tag{3.34}$$

With probability at least $1 - \delta_0$,

$$(1 - \rho)\Phi(h) \le \widetilde{\Phi}_0(h) \le (1 + \rho)\Phi(h) \qquad \text{for every } h. \tag{3.35}$$

Let x_c be the unique minimum-Euclidean-norm point in $\arg\min \widetilde{F}_0$, selected using only the pilot indices and labels, and let $x_* \in \arg\min F_b$, $F_* := F_b(x_*)$. Except with total probability $2\delta_0$,

$$R := F_b(x_c) \le C_0 F_*, \qquad C_0 := \left(1 + \frac{2}{\sqrt{(1 - \rho)\delta_0}}\right)^2. \tag{3.36}$$

Proof. The probabilities in (3.30) correspond to the scores $\tau_i' := \tau_i + T/N$, which dominate τ_i and have total mass $T' = 2T$. Hence the all-radius metric only decreases, and Lemma 3.5 applies with T'. At every $S \in \mathcal{S}$, Markov's inequality applied to (3.22), with failure probability δ_0/J, gives

$$\sup_{h \in K_S} \left|\widetilde{\Phi}_0(h) - \Phi(h)\right| \le \frac{\zeta S}{2}.$$

The ceiling and constant in (3.32) make these bounds hold simultaneously with probability at least $1 - \delta_0$. If $\Phi(h) \in [1/2, 8N^3]$, choose the least dyadic $S \in \mathcal{S}$ with $S \ge \Phi(h)$. Since $S < 2\Phi(h)$, the relative error is at most ζ.

Let $M_i := \#\{j : \mu_j = i\}$ and $\widetilde{w}_i := M_i/(m_0\pi_i^{\mathrm{pil}})$. Then

$$\widetilde{\Phi}_0(h) = \sum_i \widetilde{w}_i H(a_i^{\mathsf{T}} h), \qquad \sum_i \widetilde{w}_i = \frac{1}{m_0}\sum_{j=1}^{m_0} \frac{1}{\pi_{\mu_j}^{\mathrm{pil}}} \le 2N.$$

The all-scale extension in Lemma 3.8 therefore upgrades the finite-range event to (3.35), with $2\zeta = \rho = 1/8$.

For any fixed optimizer x_*, the affine pilot is unbiased. Markov's inequality gives

$$\widetilde{F}_0(x_*) \le F_*/\delta_0$$

except with probability δ_0. On this event and the embedding event, sample optimality and (3.4) yield

$$\sqrt{(1-\rho)\Phi(x_c - x_*)} \le \sqrt{\widetilde{\Phi}_0(x_c - x_*)} \le \sqrt{\widetilde{F}_0(x_c)} + \sqrt{\widetilde{F}_0(x_*)} \le 2\sqrt{F_*/\delta_0}.$$

A second application of the full-objective triangle inequality gives

$$\sqrt{F_b(x_c)} \le \sqrt{F_*} + \sqrt{\Phi(x_c - x_*)} \le \left(1 + \frac{2}{\sqrt{(1-\rho)\delta_0}}\right)\sqrt{F_*},$$

which is (3.36). If $F_* = 0$, sample optimality makes both pilot residuals zero on every sampled row; the embedding then forces $\Phi(x_c - x_*) = 0$, the fact used in the zero-optimum branch of Theorem 2.1. □

We now instantiate the generic N-row lemmas for the original theorem input. Set $n_{\mathrm{orig}} := n$, $I_{\mathrm{nz}} := \{i \in [n_{\mathrm{orig}}] : a_i \neq 0\}$, and $n_{\mathrm{nz}} := |I_{\mathrm{nz}}|$. Enumerate $I_{\mathrm{nz}} = \{\iota(1) < \cdots < \iota(n_{\mathrm{nz}})\}$. Fix, as a deterministic function of A, a matrix $U \in \mathbb{R}^{d\times r}$ with orthonormal columns spanning $(\ker A)^{\perp}$, and put

$$\begin{aligned} \overline{A} &:= A_{I_{\mathrm{nz}},:}U, & \overline{b}_j &:= b_{\iota(j)}, \\ \overline{F}(y) &:= \sum_{j=1}^{n_{\mathrm{nz}}} H\big((\overline{A}y)_j - \overline{b}_j\big). \end{aligned} \tag{3.37}$$

When $n_{\mathrm{nz}} > 0$, $\overline{A}$ has full column rank and no zero row. We apply the preceding lemmas with $N = n_{\mathrm{nz}}$, $A = \overline{A}$, and $b = \overline{b}$; a sampled reduced index j means a query to the original coordinate $\iota(j)$. All random indices below are drawn before any such query is made.

Two-batch nonadaptive algorithm.

1. If $n_{\mathrm{nz}} = 0$, return zero without querying a label. If $n_{\mathrm{nz}} \le 16$, query all retained labels, minimize exactly in the reduced coordinates, and lift the result with U. Otherwise compute the scores and the main and pilot probabilities in Lemma 3.4 and Equation (3.30) for $\overline{A}$.

2. Set

$$\eta := \frac{2\varepsilon + \varepsilon^2}{2C_0}, \qquad \delta := 10^{-3}, \qquad \overline{C}_{16} := \overline{C}_B\big|_{B=16}, \qquad m := \left\lceil \overline{C}_{16}\frac{T\log^3(2n_{\mathrm{nz}})}{\eta^2\delta^2} \right\rceil.$$

   If $m_0 + m \ge n_{\mathrm{nz}}$, query all retained labels and return an exact lifted minimizer.

3. Otherwise, independently draw m_0 pilot indices from π^{pil} and m main indices from π^{main}. Query their union; duplicate occurrences reuse the cached label.

4. Let y_c be the unique minimum-Euclidean-norm point of $\arg\min \widetilde{F}_0$, using only $\overline{A}$, the pilot indices, and the pilot labels—not the main indices or labels. Form $z = \overline{A}y_c - \overline{b}$ on the queried main coordinates.

5. Let $\widehat{h}$ be the unique minimum-Euclidean-norm point of $\arg\min \widehat{F}_z$. Return $\widehat{x} := U(y_c + \widehat{h}) \in \mathbb{R}^d$.

Each finite sampled objective attains its minimum on the quotient by the kernel of its sampled rows. Its argmin is a nonempty closed convex set, so it has a unique minimum-norm point. The coercivity and selection details appear in Appendix A.3.

Proof of Theorem 2.1. The zero-row and full-query branches are exact. In the two-batch branch, retain the notation $b^{\mathrm{orig}} := b$, then write $A = \overline{A}$, $b = \overline{b}$, and $F_b(y) = \sum_{i=1}^{n_{\mathrm{nz}}} H(a_i^{\mathsf{T}} y - b_i)$ for the reduced variable objective. Fix $y_* \in \arg\min F_b$ and set $F_* := F_b(y_*)$. On the pilot event (3.36), define

$$z := Ay_c - b, \qquad R := F_z(0) = F_b(y_c), \qquad h_* := y_* - y_c. \tag{3.38}$$

First suppose $F_* > 0$. Then $R \geq F_* > 0$, and

$$\sqrt{\Phi(h_*)} \leq \sqrt{F_*} + \sqrt{R} \leq 2\sqrt{R}, \qquad h_* \in K_{4R}. \tag{3.39}$$

Let $\mathcal{F}_{\mathrm{pil}}$ be generated by the pilot indices and their labels. By the pilot-only selection rule, y_c, z, R are $\mathcal{F}_{\mathrm{pil}}$-measurable, while the main sample remains independent and i.i.d. from π^{main} conditionally on $\mathcal{F}_{\mathrm{pil}}$. Thus, for almost every pilot realization with $R > 0$, Lemma 3.6 applies with $B = 16$; integrating its conditional failure probability gives an unconditional failure probability at most δ. Neither the scores nor the sample size depends on R.

For any h with $\Phi(h) = 16R$, the reverse Huber triangle inequality gives

$$\sqrt{F_z(h)} \geq \sqrt{\Phi(h)} - \sqrt{F_z(0)} = 3\sqrt{R}, \qquad F_z(h) - F_z(0) \geq 8R. \tag{3.40}$$

On the event (3.27), the sampled centered objective is strictly positive on this boundary because

$$\widehat{F}_z(h) - \widehat{F}_z(0) \geq (8 - \eta)R > 0. \tag{3.41}$$

If a global minimizer of $\widehat{F}_z$ lay outside K_{16R}, the segment from zero to that minimizer would meet the boundary. Convexity would make the sampled objective there no larger than at zero, contradicting (3.41). Hence

$$\widehat{h} \in K_{16R}. \tag{3.42}$$

Both h_* and $\widehat{h}$ are in the preserved set. Sample optimality and two applications of (3.27) yield

$$F_z(\widehat{h}) - F_z(h_*) \leq 2\eta R \leq 2\eta C_0 F_* = (2\varepsilon + \varepsilon^2)F_*. \tag{3.43}$$

Since $F_z(h_*) = F_*$ and $\widehat{x} = U(y_c + \widehat{h})$,

$$F_b(y_c + \widehat{h}) = F_z(\widehat{h}) \leq (1 + \varepsilon)^2 F_*. \tag{3.44}$$

Taking square roots proves the approximation.

If $F_* = 0$, then $b = Ay_*$. The pilot optimum y_* has sampled cost zero, so every pilot minimizer has sampled cost zero. On (3.35), the pilot-distance argument in the proof of Lemma 3.9 forces $\Phi(y_c - y_*) = 0$, hence $R = 0$. In the second stage, $h = 0$ has sampled cost zero and is the unique minimum-norm point of the argmin, so $\widehat{h} = 0$. The returned prediction is exact, and Lemma 3.6 is not invoked at radius zero.

The pilot embedding, pilot Markov bound, and centered-transfer event fail with total probability at most $3 \cdot 10^{-3} < 0.01$. Also, Equations (3.17) and (3.32) give, for a universal C,

$$m_0 + m \le M(n_{\mathrm{nz}}), \qquad M(t) := Cr\left(\log^6(2t) + \varepsilon^{-2}\log^4(2t)\right). \tag{3.45}$$

The full-query branch and monotonicity of M therefore give

$$Q \le \min\{n_{\mathrm{nz}}, M(n_{\mathrm{nz}})\} \le \min\{n_{\mathrm{orig}}, M(n_{\mathrm{orig}})\}. \tag{3.46}$$

Both batches were selected before reading labels.

Finally let $C_{\mathrm{zr}} := \sum_{i \notin I_{\mathrm{nz}}} H(-b_i^{\mathrm{orig}})$. The original objective at a lifted point is $F^{\mathrm{orig}}_{b^{\mathrm{orig}}}(Uy) = C_{\mathrm{zr}} + F_b(y)$, and its optimum is $C_{\mathrm{zr}} + F_*$. Hence

$$F^{\mathrm{orig}}_{b^{\mathrm{orig}}}(\widehat{x}) \le C_{\mathrm{zr}} + (1+\varepsilon)^2 F_* \le (1+\varepsilon)^2(C_{\mathrm{zr}} + F_*).$$

Thus zero rows preserve the guarantee and (3.37) provides the promised lift. If $r = 0$, every row is zero and no label is needed to choose an optimizer. □

4 Matching lower bound against adaptive queries

This section proves Theorem 2.2, including algorithms with private randomness and stopping-time query counts. We first isolate the scalar Huber gap, then combine binary rate–distortion with a finite adaptive-transcript bound in two block-length regimes.

4.1 Scalar Huber gap

Lemma 4.1 (Exact scalar majority gap)**.** *Let $N \ge 1$ and $0 < \gamma \le 1/2$, and define*

$$F_\gamma(t) := N(1/2-\gamma)H(4t) + N(1/2+\gamma)H(4(t-1)). \tag{4.1}$$

Put

$$q_\gamma := \frac{1/2-\gamma}{1/2+\gamma}, \qquad t^*_\gamma := 1 - \frac{q_\gamma}{4}. \tag{4.2}$$

Then t^*_γ is a global minimizer, and every $t \le 1/2$ satisfies

$$F_\gamma(t) - F_\gamma(t^*_\gamma) \ge 2\gamma N \frac{1+4\gamma}{1+2\gamma} \ge 2\gamma N. \tag{4.3}$$

For a negative empirical majority, the symmetric statement holds with $t \mapsto 1-t$.

Proof. At t^*_γ, the zero-label residual is $4t^*_\gamma = 4 - q_\gamma > 1$, while the one-label residual is $4(t^*_\gamma - 1) = -q_\gamma \in [-1, 0]$. Hence

$$F'_\gamma(t^*_\gamma) = 4N\left[(1/2-\gamma) - (1/2+\gamma)q_\gamma\right] = 0,$$

and convexity makes t^*_γ globally optimal. Also $F'_\gamma(1/2) = -8\gamma N < 0$, so the minimum on the wrong half-line $(-\infty, 1/2]$ is attained at $1/2$. Direct substitution gives

$$\begin{aligned} F_\gamma(1/2) &= 3N/2, \\ F_\gamma(t^*_\gamma) &= N\left[\frac{7}{2}(1/2-\gamma) - \frac{(1/2-\gamma)^2}{2(1/2+\gamma)}\right], \\ F_\gamma(1/2) - F_\gamma(t^*_\gamma) &= 2\gamma N \frac{1+4\gamma}{1+2\gamma}. \end{aligned}$$

This proves (4.3); reflection about $1/2$ gives the negative-majority case. □

4.2 Information lemmas for adaptive queries

For discrete random variables, H, I, and D_{KL} denote Shannon entropy, mutual information, and KL divergence in nats. Define

$$h_2^{\mathrm{nat}}(u) := -u\log u - (1-u)\log(1-u), \quad 0\log 0 := 0, \qquad d_H(s,\widehat{s}) := \sum_{j=1}^{r} \mathbf{1}\{s_j \neq \widehat{s}_j\}. \tag{4.4}$$

Lemma 4.2 (Binary rate–distortion)**.** *Let S be uniform on $\{-1,+1\}^r$, and let $\widehat{S}$ be any $\{-1,+1\}^r$-valued estimate jointly distributed with S. If $\mathbb{E}d_H(S,\widehat{S}) < 0.04r$, then*

$$\mathrm{I}(S;\widehat{S}) \geq c_{\mathrm{RD}} r, \qquad c_{\mathrm{RD}} := \log 2 - h_2^{\mathrm{nat}}(0.04) > 0.525. \tag{4.5}$$

Proof. For $e_j := \mathbb{P}[\widehat{S}_j \neq S_j]$, the entropy chain rule and the binary Fano bound give

$$\mathsf{H}(S \mid \widehat{S}) \leq \sum_{j=1}^{r} h_2^{\mathrm{nat}}(e_j) \leq r h_2^{\mathrm{nat}}\left(r^{-1} \sum_j e_j \right).$$

Since $\mathsf{H}(S) = r\log 2$, the claim follows by concavity of h_2^{nat} [6]. □

The next lemma accounts for the seed, adaptive index choices, caching, and stopping without an infinite transcript. Once repetitions are served from the cache, at most n fresh coordinates can be queried.

Lemma 4.3 (Finite adaptive-transcript bound)**.** *Let S be any random variable, let $B = (B_1,\dots,B_n)$ be a random response vector, and let $\mathcal{R}$ be a private seed independent of (S,B). Suppose an adaptive algorithm returns almost surely after Q fresh coordinate queries to B. For $t = 1,\dots,n$, let $Z_t = \dagger$ if $Q < t$, and otherwise let $Z_t = (J_t, B_{J_t})$ record the t-th fresh index and response. Put $\mathcal{F}_{t-1} := \sigma(\mathcal{R}, Z_1,\dots,Z_{t-1})$. The active indicator and J_t on the active event are $\mathcal{F}_{t-1}$-measurable. For an active history h, let I_h denote mutual information under the regular conditional law given h. If $I_h(S; B_{J_t(h)}) \leq \kappa$ for almost every active history, then*

$$\mathrm{I}(S;\mathcal{R},Z_{1:n}) \leq \kappa\,\mathbb{E}Q. \tag{4.6}$$

Proof. The seed is independent of S, and a stopped slot is deterministic. For each slot, the pointwise hypothesis and measurability of its active indicator and index give

$$\mathrm{I}(S; Z_t \mid \mathcal{F}_{t-1}) \leq \kappa \mathbb{P}(Q \geq t).$$

The chain rule therefore gives

$$\mathrm{I}(S;\mathcal{R},Z_{1:n}) = \sum_{t=1}^{n} \mathrm{I}(S; Z_t \mid \mathcal{F}_{t-1}) \leq \kappa \sum_{t=1}^{n} \mathbb{P}(Q \geq t) = \kappa\,\mathbb{E}Q.$$

□

4.3 Proof of the lower bound

Proof of Theorem 2.2. We prove a distributional lower bound and then average to a fixed response. Fix n, d, r, ε, put $N := \lfloor n/r \rfloor$, and use only the first r columns of the design. If the expected query count under either hard distribution below is infinite, the conclusion is immediate; otherwise the algorithm returns almost surely and Lemma 4.3 applies. For either construction, decode an output by setting $\widehat{S}_j = +1$ when $\widehat{x}_j > 1/2$ and $\widehat{S}_j = -1$ otherwise. The decoder is measurable with respect to $(\mathcal{R}, Z_{1:n})$.

Short blocks: $N < 160$. Let the first r rows of A be the standard basis rows $e_1^\mathsf{T}, \dots, e_r^\mathsf{T}$, and set the remaining rows to zero. Draw S uniformly from $\{-1,+1\}^r$, set $b_j = S_j$ for $j \le r$, and set the padded responses to zero. The optimum is zero, so on the success event a multiplicative approximation must recover every S_j. The binary decoder therefore satisfies $\mathbb{E}d_H(S,\widehat{S}) \le 0.01r$. Each fresh informative response carries at most $\log 2$ nats, while padded rows are deterministic. Data processing and Lemmas 4.2 and 4.3 give

$$\mathbb{E}Q \ge (c_{\mathrm{RD}}/\log 2)r.$$

Because $n < 160r$ and $r/\varepsilon^2 \ge n$, this is $\Omega(\min\{n, r/\varepsilon^2\})$.

Long blocks: $N \ge 160$. Set

$$\beta := \max\left\{200\varepsilon, \sqrt{\frac{10}{N}}\right\}. \tag{4.7}$$

Then $N\beta^2 \ge 10$ and $\beta \le 1/4$. Conditional on independent uniform signs $S_1, \dots, S_r$, draw independent bits

$$Y_{jk} \sim \mathrm{Ber}(1/2 + \beta S_j), \qquad j \in [r],\ k \in [N],$$

and give row (j,k) the design vector $4e_j^\mathsf{T}$ and response $4Y_{jk}$. Pad the fewer than r remaining rows with zeros. A block is good when $S_j(\overline{Y}_j - 1/2) \ge \beta/2$, where $\overline{Y}_j := N^{-1}\sum_{k=1}^N Y_{jk}$; Hoeffding's inequality gives $\mathbb{P}(\text{block } j \text{ is not good} \mid S_j) \le e^{-5}$.

On a good block, Lemma 4.1 charges at least βN excess loss for a wrong sign. The vector whose first r coordinates equal $1/2$ has loss $3Nr/2$. For the realized response put $F_* := \min_x F_b(x)$; on the success event

$$F_b(\widehat{x}) - F_* \le (2\varepsilon + \varepsilon^2)F_* \le \frac{9}{2}\varepsilon N r.$$

Since $\beta \ge 200\varepsilon$, at most $9r/400$ good blocks are decoded incorrectly. Charging all nongood blocks and all blocks on the failure event yields

$$\mathbb{E}d_H(S,\widehat{S}) \le (e^{-5} + 9/400 + 0.01)r < 0.04r. \tag{4.8}$$

It remains to bound the information in one adaptive response. Let $P_s := \mathrm{Ber}(1/2+\beta s)$. For $\mathcal{F}_{t-1}$-almost every active history h, a padded-row response is deterministic and carries zero information. Otherwise the fresh index is the fixed pair $J_t(h) = (j,k)$. Under the corresponding regular conditional law, freshness gives the Markov relation $S_{-j} \to S_j \to Y_{jk}$, and the observed response $4Y_{jk}$ is a bijective function of Y_{jk}. Put $\theta_s(h) := \mathbb{P}(S_j = s \mid h)$. Convexity of KL in its second argument gives, under this conditional law,

$$\begin{aligned} I_h(S; Y_{jk}) &= I_h(S_j; Y_{jk}) \\ &\le \sum_{s,s' \in \{-1,+1\}} \theta_s(h)\theta_{s'}(h)\mathrm{D}_{\mathrm{KL}}(P_s \| P_{s'}) \le 16\beta^2, \end{aligned} \tag{4.9}$$

where $\mathrm{D}_{\mathrm{KL}}(P_+\|P_-) = 2\beta\log((1+2\beta)/(1-2\beta)) \le 16\beta^2$, and the reverse divergence is identical. Thus Lemma 4.3 applies with $\kappa = 16\beta^2$. Data processing, (4.8), and Lemma 4.2 imply

$$\mathbb{E}Q \ge \frac{c_{\mathrm{RD}}}{16\beta^2} r. \tag{4.10}$$

Finally, since $rN \ge n - r \ge 159n/160$,

$$\frac{r}{\beta^2} = \min\left\{\frac{r}{40000\varepsilon^2}, \frac{rN}{10}\right\} \ge c\min\left\{\frac{r}{\varepsilon^2}, n\right\}. \tag{4.11}$$

In either regime, the hard response distribution has finite support and the algorithm succeeds with probability at least 0.99 for every fixed response. Moreover, $\mathbb{E}_{S,Y,\mathcal{R}} Q = \mathbb{E}_b \mathbb{E}_{\mathcal{R}} Q(A, b)$, with Y absent in the short regime. Hence some fixed response in the support attains the distributional expected-query lower bound, proving Theorem 2.2. □

5 Related work and limitations

Active regression. For least squares, leverage sampling gives an $O(r \log r)$ constant-accuracy baseline and spectral sparsification reaches linear dependence on the effective dimension [4]. Lewis-weight sampling gives nearly tight guarantees for least absolute deviations [2, 15]. After column-space reduction, the nonadaptive Huber specialization of Musco et al. [14] uses $\widetilde{O}(r^{4-2\sqrt{2}} \operatorname{poly}(1/\varepsilon))$ labels; our gain is the linear rank exponent and the matching joint dependence on n, r, ε. Stratified, semi-supervised, online, and single-index regression use different models [3, 8, 10, 13, 16, 17]; statistical Huber regression instead studies population risk [9, 18].

Sparsification and the contribution boundary. Lewis weights and sensitivity sampling underlie subspace and objective summaries [5, 7]. Jambulapati et al. [11] developed common-basepoint signed chaining, and Jambulapati et al. [12] gave multiscale generalized-linear-model sparsifiers, including the unshifted Huber case. Full-data sparsification does not directly give active sampling because the affine augmentation contains hidden labels. The step specific to this paper is Lemma 3.6: after an independent pilot, design-only scores preserve the localized affine increments needed by the optimizer. We do not claim the pilot, localization, common-basepoint chaining, or unshifted sparsification separately. A fixed label-independent weighted proper-subset objective also cannot preserve every affine baseline, and optimizing either endpoint surrogate can be polynomially worse. The similarly titled work of Cavazza and Murino [1] concerns semi-supervised multi-view learning.

Proposition 5.1 (Simple surrogate minimizers can be polynomially bad)**.** *For all sufficiently large N, there are one-dimensional instances on which the exact least-squares minimizer has Huber-norm approximation ratio $\Theta(N^{1/4})$, and instances on which the exact ℓ_1-regression minimizer has ratio $\Theta(\sqrt{N})$.*

The explicit constructions are given in Appendix B.

Limitations. The result counts labels under exact real arithmetic and exact convex optimization; it gives neither a near-linear-time implementation, a strong affine coreset, nor optimal confidence dependence. Markov's inequality causes the δ^{-2} factor, while the scale union, chaining, and pilot cause the logarithms. Known Huber sensitivity examples have $\Omega(r \log\log(n/r))$ total sensitivity in some ranges [14], although this is not an active-query lower bound for those logarithms. Reducing the logarithmic and confidence factors, obtaining a fast implementation, and extending the centered transfer to other robust losses remain open.

References

[1] J. Cavazza and V. Murino. Active regression with adaptive Huber loss. *arXiv preprint*, 2016. arXiv:1606.01568.

[2] X. Chen and M. Dereziński. Query complexity of least absolute deviation regression via robust uniform convergence. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, volume 134 of *Proceedings of Machine Learning Research*, pages 1144–1179, 2021. arXiv:2102.02322.

[3] C. Chen, Y. Li, and Y. Sun. Online active regression. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 3320–3335, 2022. arXiv:2207.05945.

[4] X. Chen and E. Price. Active regression via linear-sample sparsification. In *Proceedings of the 32nd Conference on Learning Theory (COLT)*, volume 99 of *Proceedings of Machine Learning Research*, pages 663–695, 2019. arXiv:1711.10051.

[5] M. B. Cohen and R. Peng. ℓ_p row sampling by Lewis weights. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC)*, pages 183–192, 2015. doi:10.1145/2746539.2746567.

[6] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, second edition, 2006.

[7] D. Feldman and M. Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing (STOC)*, pages 569–578, 2011. doi:10.1145/1993636.1993712.

[8] A. Gajjar, W. M. Tai, X. Xu, C. Hegde, C. Musco, and Y. Li. Agnostic active learning of single index models with linear sample complexity. In *Proceedings of the 37th Conference on Learning Theory (COLT)*, volume 247 of *Proceedings of Machine Learning Research*, pages 1715–1754, 2024. arXiv:2405.09312.

[9] P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. doi:10.1214/aoms/1177703732.

[10] C. W. In, Y. Li, W. M. Tai, and X. Wu. Active regression for single-index models with unknown link functions. *arXiv preprint*, 2026. arXiv:2608.01287.

[11] A. Jambulapati, J. R. Lee, Y. P. Liu, and A. Sidford. Sparsifying sums of norms. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1953–1962, 2023. doi:10.1109/FOCS57990.2023.00119.

[12] A. Jambulapati, J. R. Lee, Y. P. Liu, and A. Sidford. Sparsifying generalized linear models. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC)*, pages 1665–1675, 2024. doi:10.1145/3618260.3649684. Full version: arXiv:2311.18145.

[13] Y. Li and W. M. Tai. Near-optimal active regression of single-index models. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025. arXiv:2502.18213.

[14] C. Musco, C. Musco, D. P. Woodruff, and T. Yasuda. Active linear regression for ℓ_p norms and beyond. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 744–753, 2022. doi:10.1109/FOCS54457.2022.00076. arXiv:2111.04888.

[15] A. Parulekar, A. Parulekar, and E. Price. ℓ_1 regression with Lewis weights subsampling. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2021)*, volume 207 of *Leibniz International Proceedings in Informatics*, article 49, 2021. doi:10.4230/LIPIcs.APPROX/RANDOM.2021.49. arXiv:2105.09433.

[16] N. Rajaraman, F. Devvrit, and P. Awasthi. Semi-supervised active linear regression. In *Advances in Neural Information Processing Systems 35*, pages 1294–1306, 2022. doi:10.52202/068431-0095.

[17] S. Sabato and R. Munos. Active regression by stratification. In *Advances in Neural Information Processing Systems 27*, pages 469–477, 2014. arXiv:1410.5920.

[18] Q. Sun, W.-X. Zhou, and J. Fan. Adaptive Huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020. doi:10.1080/01621459.2018.1543124.

[19] M. Talagrand. *Upper and Lower Bounds for Stochastic Processes*. Springer, 2014. doi:10.1007/978-3-642-54075-2.

A Technical details for the upper bound

This appendix records the construction, endpoint algebra, and attainment details used in Section 3. Throughout, $A \in \mathbb{R}^{N\times r}$ has full column rank and no zero row, and the notation is that of Section 3.1.

A.1 Deterministic weight schemes and endpoint covers

We spell out the exact-real construction asserted in Lemma 3.2. For positive vectors define

$$d_{\log}(u,w) := \max_{i\le N} |\log(u_i/w_i)|$$

and, at dyadic scale 2^j, define the exact leverage map

$$\mathcal{W}_j(w)_i := \frac{H(v_i(w))}{2^j v_i(w)^2}, \qquad v_i(w)^2 = a_i^{\mathsf{T}}(A^{\mathsf{T}}\operatorname{diag}(w)A)^{-1}a_i.$$

All coordinates are positive because A has full column rank and no zero row. The log-Lipschitzness of weighted leverage and the lower-1, upper-2 scaling of H imply

$$d_{\log}(\mathcal{W}_j(u),\mathcal{W}_j(w)) \le \tfrac{1}{2} d_{\log}(u,w), \qquad \mathcal{W}_{j+1}(w) = \tfrac{1}{2}\mathcal{W}_j(w).$$

Let $j_0 := \min \mathcal{I}$, set $u^{[0]} := \mathbf{1}$, and iterate $u^{[k+1]} := \mathcal{W}_{j_0}(u^{[k]})$. The contraction gives

$$d_{\log}(u^{[k+1]},u^{[k]}) \le 2^{-k} d_{\log}(u^{[1]},u^{[0]}),$$

and the right-hand side is finite. Stop at the first k for which it is at most $\log 2$, and set $w^{(j_0)} := u^{[k]}$. Recursively warm-start every later scale by the single update

$$w^{(j+1)} := \mathcal{W}_{j+1}(w^{(j)}), \qquad j = j_0,\dots,\max\mathcal{I}-1. \tag{A.1}$$

If $d_{\log}(\mathcal{W}_j(w^{(j)}),w^{(j)}) \le \log 4$, then

$$\frac{1}{8} \le \frac{w_i^{(j+1)}}{w_i^{(j)}} \le 2, \qquad d_{\log}\big(\mathcal{W}_{j+1}(w^{(j+1)}),w^{(j+1)}\big) \le \tfrac{1}{2}\log 8 < \log 4.$$

Induction therefore gives both inequalities in the weight-scheme definition with the universal choice $\alpha_0 = 4$. The construction makes finitely many exact-real updates and is deterministic in $(A,\mathcal{I})$. Moreover, for every scale,

$$\sum_i w_i^{(j)} v_i(w^{(j)})^2 = \operatorname{tr}\Big(A^{\mathsf{T}}\operatorname{diag}(w^{(j)})A(A^{\mathsf{T}}\operatorname{diag}(w^{(j)})A)^{-1}\Big) = r.$$

Finally, the residual bound is exactly

$$4^{-1} \le \frac{H(v_i(w^{(j)}))}{2^j w_i^{(j)} v_i(w^{(j)})^2} \le 4,$$

which is (3.8). This establishes the construction claim in Lemma 3.2. If a numerical implementation substitutes scores $\widetilde{\lambda}_i \in [\lambda_i/c, c\lambda_i]$ for a fixed $c \ge 1$, the safe sampling score is $c\widetilde{\lambda}_i$: it dominates λ_i coordinatewise and increases total mass by only a constant factor. No approximation is needed for the transition scheme; endpoint approximations are handled below.

We also use the following direct power-function specialization of the same source theorem. For $p \in \{1, 2\}$, let $\mathcal{I}$ be a finite contiguous integer interval of size at least $4\log(2N)$. If positive weights satisfy

$$\frac{\left|v_i(w^{(j)})\right|^p}{w_i^{(j)} v_i(w^{(j)})^2} = 2^j, \qquad w_i^{(j+1)} \le w_i^{(j)}, \tag{A.2}$$

set $\lambda_i^{(j)} := w_i^{(j)} v_i(w^{(j)})^2$, $\tau_i := \max_{j\in\mathcal{I}} \lambda_i^{(j)}$, and $k_{\mathcal{I}} := \max \mathcal{I} + 1$. Then

$$\gamma_2\left(\{h : \|Ah\|_p^p \le 2^{k_{\mathcal{I}}}\}, (h,k) \mapsto \max_i \frac{\left|a_i^{\mathsf{T}}(h-k)\right|^{p/2}}{\sqrt{\tau_i}}\right) \le C 2^{k_{\mathcal{I}}/2} \log^{3/2}(2N). \tag{A.3}$$

This is Jambulapati et al. [12, Definition 1.8 and the proof of Theorem 3.14] with $f_i(t) = |t|^p$; these functions meet the theorem's symmetry, scaling, and metric assumptions exactly.

A.2 Endpoint score bounds

Proof of Lemma 3.3. Fix $p \in \{1, 2\}$ and use the notation of (3.13)–(3.14). The trace identity gives

$$\sum_{i=1}^N \lambda_i^{(p)} = \operatorname{tr}\Big(A^{\mathsf{T}} \operatorname{diag}(w^{(p)}) A \big(A^{\mathsf{T}} \operatorname{diag}(w^{(p)}) A\big)^{-1}\Big) = r.$$

For an integer scale j, set $w_i^{(j)} := 2^{-2j/p} w_i^{(p)}$. If

$$(v_i^{(j)})^2 := a_i^{\mathsf{T}} (A^{\mathsf{T}} \operatorname{diag}(w^{(j)}) A)^{-1} a_i,$$

then $v_i^{(j)} = 2^{j/p} v_i^{(p)}$, and hence

$$\frac{(v_i^{(j)})^p}{w_i^{(j)} (v_i^{(j)})^2} = 2^j, \qquad w_i^{(j+1)} = 2^{-2/p} w_i^{(j)}, \qquad w_i^{(j)} (v_i^{(j)})^2 = \lambda_i^{(p)}.$$

Thus this is an exact weight scheme for the power function $f_p(t) = |t|^p$, with a scale-independent leverage vector. Let $L := \lceil 4\log(2N) \rceil$ and $\mathcal{I}_0 := \{-L, \dots, -1\}$, so $k_{\mathcal{I}_0} = 0$. Applying (A.3) to this block gives (3.16) for $S = 1$. The substitution $h = S^{1/p} u$ maps the unit ball onto $\{h : \|Ah\|_p^p \le S\}$ and multiplies d_p by $S^{1/2}$, which proves the stated bound for every $S > 0$. □

A.3 Attainment and deterministic selections

For completeness, consider a sampled objective

$$G(x) = \sum_{i \in S} \omega_i H(a_i^\mathsf{T} x - \beta_i), \qquad \omega_i > 0.$$

Represent a quotient class $[x]$ by its unique member in the orthogonal complement of the kernel of the sampled rows, and use that representative's Euclidean norm as $\|[x]\|_2$. Injectivity and compactness of the unit sphere give a constant $c_S > 0$. With $C_S := \sum_{i \in S} \omega_i (|\beta_i| + 1/2)$,

$$G(x) \geq \sum_{i \in S} \omega_i \left| a_i^\mathsf{T} x \right| - \sum_{i \in S} \omega_i (|\beta_i| + 1/2) \geq c_S \|[x]\|_2 - C_S.$$

Thus G is coercive on that quotient and attains its minimum. Its argmin in $\mathbb{R}^r$ is nonempty, closed, and convex, so the Euclidean projection of the origin onto it exists and is unique. This is the minimum-norm selection used for both batches. Equivalently, it is the limit, as $k \to \infty$, of the unique minimizers of $G(x) + k^{-1} \|x\|_2^2$. This characterization makes the selection a measurable deterministic function of the finite sample and its observed labels. In particular, the pilot selection uses only the pilot indices, pilot labels, and fixed design; the independent main sample cannot affect it.

B Surrogate counterexamples

Proof of Proposition 5.1. For least squares, take N rows $(a, b) = (1, 0)$ and one row $(a, b) = (\sqrt{N}, 4\sqrt{N})$. The squared-loss surrogate is $G_2(x) = Nx^2 + N(x-4)^2$, whose unique minimizer is $x = 2$. The Huber objective is $\Psi_2(x) = NH(x) + H(\sqrt{N}(x-4))$. At $x_* = N^{-1/2}$, both terms are differentiable and $\Psi_2'(x_*) = N/\sqrt{N} - \sqrt{N} = 0$; convexity makes x_* optimal. Moreover,

$$\Psi_2(2) = \frac{3}{2}N + 2\sqrt{N} - \frac{1}{2}, \qquad \Psi_2(x_*) = 4\sqrt{N} - 1,$$

so the ratio of square-root objectives is $\Theta(N^{1/4})$.

For ℓ_1, take N rows $(1, 0)$ and one row $(N/2, 1)$. The surrogate $G_1(x) = N|x| + |Nx/2 - 1|$ has slopes $-3N/2$, $N/2$, and $3N/2$ on its three linear intervals, so zero is its unique minimizer. The Huber objective $\Psi_1(x) = NH(x) + H(Nx/2 - 1)$ is minimized at $x_* = 2/(N+4)$, where both residuals are quadratic and the derivative is zero. Finally,

$$\Psi_1(0) = \frac{1}{2}, \qquad \Psi_1(x_*) = \frac{2}{N+4},$$

so the square-root-objective ratio is exactly $\sqrt{(N+4)/4} = \Theta(\sqrt{N})$. □