ROBUST COMPOSITE RECOVERY FOR TWO-COMPONENT MIXED LINEAR REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixed linear regression (MLR) models heterogeneous populations through multiple linear relationships with unobserved component memberships, with broad applications across market segmentation, trajectory clustering, and signal unmixing. A fundamental challenge in robust MLR is to reliably recover the underlying regression coefficients in the presence of both structured inlier noise and a small fraction of arbitrarily large response outliers, while providing rigorous finite-sample guarantees. We study the canonical two-component MLR setting with Gaussian covariates and component proportions bounded away from zero. We propose SARP-P (Segment-initialized Adaptive Robust Prox-linear estimation with Polishing), a complete estimator that unifies segment initialization, normalized composite iteration, and residual-calibrated response polishing in a single pipeline. With a sample size nearly linear in the covariate dimension, we establish a full finite-sample recovery guarantee: Provably correct initialization, a quadratic error recursion converging to an explicit noise floor, and valid recovery of the final polished estimate. Our bounds tolerate a bounded fraction of arbitrary response outliers independent of their amplitudes, and hold uniformly across all data-dependent adaptive weights and trimming steps. All stages of the estimator reuse the same observations, and all guarantees hold under a single high-probability event. A computable data-driven acceptance rule transfers recovery guarantees to the final polished estimate without requiring a known noise scale. Synthetic experiments demonstrate that response polishing improves estimation accuracy, with particularly strong gains under scale-mixture noise with response outliers.

1 INTRODUCTION

Mixed linear regression (MLR) is a foundational framework for modeling heterogeneous populations where each observation follows one of several linear relationships, with the component identity unobserved. Unlike a single regression that averages across subpopulations and masks group-specific behavior, MLR recovers distinct coefficient vectors, enabling applications ranging from market segmentation and trajectory clustering to signal unmixing and system identification (DeSarbo & Cron, 1988; Gaffney & Smyth, 1999). For practical deployment, reliable estimation must handle two ubiquitous sources of corruption simultaneously: Mild measurement noise affecting most observations, and occasional unreliable responses that can take arbitrarily large values. While existing literature on robust mixture estimators has developed heuristics and partial results (Neykov et al., 2007; Bai et al., 2012; Yao et al., 2014; Song et al., 2014), a central open theoretical question remains: How to achieve provably consistent recovery of the component regressions with finite-sample guarantees, under both inlier noise and adversarial response outliers.

The two-component MLR setting is already well studied, with a line of work on exact and noisy recovery (Yi et al., 2014; Chen et al., 2018; Kwon et al., 2019; Ghosh & Kannan, 2020). A major difficulty is that unobserved component labels tightly couple the quality of initialization with subsequent coefficient refinement. For uncontaminated data, spectral initialization combined with local iteration yields rigorous recovery guarantees (Yi et al., 2014; Zhong et al., 2016; Balakrishnan et al., 2017), and convex formulations achieve minimax-optimal recovery rates under additive noise (Chen et al., 2014; 2018). However, arbitrary response outliers can disrupt both initialization and local convergence. Among robust methods, iterative least trimmed squares (ILTS) provides recovery guarantees under response outliers, but its analysis excludes additive stochastic inlier noise (Shen &

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Sanghavi, 2019). Mix-IRLS, an iteratively reweighted least-squares approach, shows strong empirical performance but only admits a population-level analysis for sufficiently imbalanced mixtures (Zilber & Nadler, 2023). Notably, no existing method provides a complete finite-sample guarantee that spans initialization, local optimization, and final output, under simultaneous inlier noise and adversarial response outliers.

The quadratic formulation of two-component MLR (Chen et al., 2018) suggests a natural connection to composite optimization techniques developed for robust phase retrieval (Duchi & Ruan, 2019). However, this connection does not directly translate into recovery guarantees for MLR. A central structural difference is that in MLR, an observed response appears in both the residual and its derivative with respect to the coefficients, thus a single outlier can corrupt the entire local optimization landscape. While normalization can tame the derivative magnitude, adaptive normalization and sample trimming change the objective function at every iteration. A valid recovery analysis need to simultaneously control these data-dependent objectives and the initialization step computed from the same contaminated sample.

In this paper, we develop SARP-P, a complete robust estimator for two-component MLR with end-to-end finite-sample recovery guarantees. SARP-P operates in three stages, namely segment initialization to locate the line segment spanned by the two coefficient vectors and estimate both endpoints, normalized composite iteration for robust local refinement with provably quadratic convergence, and residual-calibrated response polishing to further improve accuracy. A uniform growth inequality ensures that the recovery guarantee remains valid despite data-dependent adaptive weights, sample trimming, and repeated use of the same observations across all stages. The final stage uses a computable data-driven acceptance rule that transfers the recovery bound to the polished estimate, with no requirement for a known noise scale. Our analysis handles independent inlier errors such as sub-exponential noises. Synthetic experiments demonstrate that response polishing delivers consistent accuracy gains, with particularly significant improvements under heavy-tailed and scale-mixture noise.

1.1 Related Work

Our work lies at the intersection of mixed linear regression (MLR) theory, robust statistical estimation, and composite optimization for inverse problems. We contextualize our contributions against the most relevant strands of prior research.

Statistical and computational theory of mixed regression. Mixed linear regression has a long history in statistics, with early work focusing on likelihood-based inference and algorithmic heuristics (Quandt & Ramsey, 1978; De Veaux, 1989; Viele & Tong, 2002; Chaganty & Liang, 2013; Anandkumar et al., 2014; Yi et al., 2014; 2016; Zhong et al., 2016; Sedghi et al., 2016). Further work studies the complexity of learning general mixtures (Li & Liang, 2018; Chen et al., 2020; Diakonikolas & Kane, 2020). For the two-component setting, convex estimation yields the quadratic representation and minimax noisy-recovery rates (Chen et al., 2014; 2018). Other convex guarantees concern well-separated data (Hand & Joshi, 2018), while approximate message passing admits asymptotic analysis (Tan & Venkataramanan, 2023). Analyses of expectation–maximization (EM) address local accuracy, global convergence, signal-to-noise dependence, and misspecification (Balakrishnan et al., 2017; Klusowski et al., 2019; Kwon et al., 2019; Kwon & Caramanis, 2020; Kwon et al., 2021; Ghosh & Mazumdar, 2024). For symmetric coefficients, Luo & Hashemi (2024) characterize EM trajectories, Luo & Hashemi (2025) analyze overspecification, and Luo & Hashemi (2026) establish an asymptotic score-matching connection. However, all of these results assume well-behaved additive noise and do not account for arbitrary response outliers, which are known to break both spectral initialization and least-squares-based refinement.

Robust mixed regression and robust estimation. A large literature has developed robust heuristics for mixture models to handle contaminated data. Early proposals include trimmed likelihood estimators, robust clusterwise regression, and robust EM variants (Neykov et al., 2007; García-Escudero et al., 2010; Bai et al., 2012), as well as mixture models based on heavy-tailed distributions such as Student's t and Laplace (Yao et al., 2014; Song et al., 2014). Among methods with theoretical guarantees, iterative least trimmed squares (ILTS) (Shen & Sanghavi, 2019) is the closest point of comparison. ILTS provides local recovery guarantees under response outliers, but its analysis

excludes additive stochastic inlier noise, and its error bound contains a contamination-dependent linear term. Mix-IRLS (Zilber & Nadler, 2023) demonstrates strong empirical performance with an outlier-handling option, but its theoretical analysis is limited to a population-level bounded-noise result for sufficiently imbalanced mixtures. Despite this progress, no existing robust MLR method provides a complete finite-sample guarantee that covers initialization, local optimization, and final output simultaneously under both inlier noise and adversarial response outliers.

Robust quadratic recovery and composite optimization. Our normalized composite iteration is inspired by work on robust composite optimization for phase retrieval. Duchi & Ruan (2019) established robust spectral initialization and quadratic convergence for composite phase retrieval, and follow-up work extended the framework to bilinear inverse problems and truncated-gradient methods (Charisopoulos et al., 2021; Huang et al., 2025). However, there is a critical structural gap between phase retrieval and MLR: In phase retrieval, a replaced intensity does not enter the residual derivative, whereas in MLR a response outlier contaminates both the residual and its gradient. Directly applying unnormalized composite methods to MLR therefore fails, as a single outlier can dominate the entire local landscape. Additionally, phase retrieval does not face the component-label ambiguity that requires segment initialization in MLR. This paper addresses both gaps via derivative normalization and segment-based initialization, adapting the composite optimization paradigm to the MLR setting.

Structured and alternative observation models. Extended settings such as sparse MLR, graph-structured mixtures, and query-based measurement have also been studied (Städler et al., 2010; Yi & Caramanis, 2015; Wang et al., 2015; Barik & Honorio, 2022; Fan et al., 2018; Arpino & Venkataramanan, 2023). We focus on the canonical two-component non-sparse setting with standard Gaussian covariates, which serves as the foundational benchmark for MLR theory. Our robustness techniques are potentially applicable to these structured extensions as well.

1.2 Main Contributions

Our main contributions are as follows.

- We design SARP-P as a complete estimator combining segment initialization, normalized composite iteration, and response polishing. Normalization prevents arbitrary response amplitudes from dominating the local derivatives. Residual calibration and a computable acceptance rule incorporate classical robust scores without requiring a known separation or noise scale.
- We establish finite-sample recovery for the complete SARP-P estimator in Theorems 4.1 and 4.2. With sample size nearly linear in the covariate dimension, a single high-probability event simultaneously guarantees successful initialization, a quadratic error recursion converging to an explicit noise floor, and valid recovery of the final polished estimate. All guarantees hold uniformly across data-dependent adaptive weights and trimmed sample sets, with all stages reusing the same observations. To our knowledge, this is the first end-to-end finite-sample recovery result for robust MLR under simultaneous inlier noise and adversarial response contamination.
- We evaluate SARP-P on synthetic data across multiple noise distributions and outlier mechanisms. Experiments show that response polishing consistently improves estimation accuracy over the composite-only baseline. Compared to relevant baseline methods, SARP-P achieves the best overall performance, with particularly pronounced gains under scale-mixture noise with response outliers. Ablation studies isolate the effects of composite iteration, response polishing, and residual calibration.

2 Problem Formulation

We specify two-component MLR with independent inlier noise and arbitrary response outliers. A common tail envelope includes bounded and unbounded errors, and a permutation-invariant loss measures coefficient recovery. Let n be the sample size, d the covariate dimension, and $[n] = \{1, \dots, n\}$. The two distinct regression coefficients are $\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^* \in \mathbb{R}^d$. Independently across $i \in [n]$, draw

$$\boldsymbol{x}_i \sim N(\mathbf{0}, \boldsymbol{I}_d), \qquad z_i \perp \boldsymbol{x}_i, \qquad \mathbb{P}(z_i = 1) = p, \quad \mathbb{P}(z_i = 2) = 1 - p, \quad p \in [p_0, 1 - p_0]. \quad (1)$$

Here $\boldsymbol{I}_d$ is the identity matrix and z_i is unobserved. Gaussian covariates are standard in MLR theory (Yi et al., 2014; Shen & Sanghavi, 2019; Zilber & Nadler, 2023) and provide the rotational identity for segment initialization. The supplied constant $p_0 \in (0, 1/2]$ lower-bounds both proportions, as required in related sample and initialization bounds (Yi et al., 2014, Theorem 4.4)(Shen & Sanghavi, 2019, Theorem 7). Fixing p_0 permits sample size nearly linear in d. Response outliers can conceal a sufficiently rare component, hence Theorem 4.1 additionally requires a small outlier fraction relative to p_0. The actual proportion p is unknown, and balance is allowed.

The errors $\xi_1, \ldots, \xi_n$ are mutually independent and independent of all covariates and labels. For every $i \in [n]$, the error distribution satisfies

$$\mathbb{P}(|\xi_i| > u) \le p_{\mathrm{tail}}(u) \qquad (i \in [n],\ u \ge 0), \tag{2}$$

where the deterministic envelope $p_{\mathrm{tail}} : [0, \infty) \to [0, 1]$ enters only the analysis. Neither identical error distributions nor symmetry is required.[1] The bounded-error envelope includes noiseless observations at $\sigma = 0$. Gaussian errors and finite centered Gaussian scale mixtures are sub-Gaussian. Laplace errors are sub-exponential. Independent error models are standard in noisy MLR and robust regression-mixture analysis (Chen et al., 2018; Song et al., 2014; Zilber & Nadler, 2023). The algorithm receives neither the envelope nor the noise scale.

We observe the covariates and contaminated responses

$$y_i = \begin{cases} \boldsymbol{x}_i^\top \boldsymbol{\beta}^*_{z_i} + \xi_i, & i \notin O, \\ v_i, & i \in O, \end{cases} \qquad |O| \le \lfloor \epsilon n \rfloor, \qquad 0 \le \epsilon \le \bar{\epsilon} < 1. \tag{3}$$

The response-outlier set $O \subseteq [n]$ and values $v_i \in \mathbb{R}$ may depend arbitrarily on $\{(\boldsymbol{x}_i, z_i, \xi_i)\}_{i=1}^n$. Only responses are modified, as in Shen & Sanghavi (2019). The supplied budget $\bar{\epsilon}$ determines the retained sample size. Error independence refers to the original sample, before the adversary selects O.

Write $y_i^0 = \boldsymbol{x}_i^\top \boldsymbol{\beta}^*_{z_i}$ for the noiseless response. Define the separation $\Delta = \|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\| > 0$, midpoint $\boldsymbol{m}^* = (\boldsymbol{\beta}_1^* + \boldsymbol{\beta}_2^*)/2$, direction $\boldsymbol{u}^* = (\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*)/\Delta$, and segment $S = [\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*]$. For a stacked pair $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top \in \mathbb{R}^{2d}$, define the permutation-invariant error by

$$\begin{aligned} \Theta^* &= \left\{ \left((\boldsymbol{\beta}^*_{\pi(1)})^\top, (\boldsymbol{\beta}^*_{\pi(2)})^\top \right)^\top : \pi \in \mathcal{S}_2 \right\}, \\ \mathcal{E}(\boldsymbol{\theta}) &= \mathrm{dist}(\boldsymbol{\theta}, \Theta^*) = \min_{\pi \in \mathcal{S}_2} \left(\sum_{j=1}^{2} \|\boldsymbol{\beta}_j - \boldsymbol{\beta}^*_{\pi(j)}\|^2 \right)^{1/2}. \end{aligned} \tag{4}$$

where $\mathcal{S}_2 = \{\pi_{\mathrm{id}}, \pi_{\mathrm{sw}}\}$ with $\pi_{\mathrm{id}}(j) = j$ and $\pi_{\mathrm{sw}}(j) = 3 - j$ for $j \in \{1, 2\}$. Vector norms are Euclidean, and matrix operator norms are denoted by $\|\cdot\|_{\mathrm{op}}$. No sparsity or bound on $\|\boldsymbol{m}^*\|$ is imposed. The recovery conditions compare a noise quantile with Δ, reflecting the difficulty of separating two regressions when noise is comparable to the difference between the component responses.

3 Adaptive Composite Recovery

We specify segment initialization, normalized composite iteration, and response polishing in the order used by SARP-P. Algorithm 1 combines the coefficient estimates, convex updates, and final acceptance test. Appendix B supplies subroutines and optimization certificates.

3.1 Segment initialization

Segment initialization estimates a point near S, determines the direction of S, and searches for both endpoints. A single absolute-loss fit need not identify either coefficient in a balanced mixture.

Let $\boldsymbol{b}$ minimize the least absolute deviations (LAD) loss and let τ be an upper residual quantile:

$$\boldsymbol{b} \in \arg\min_{\boldsymbol{v} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} |y_i - \boldsymbol{x}_i^\top \boldsymbol{v}|, \qquad \tau = Q_{1-p_0/2}\big(\{|y_i - \boldsymbol{x}_i^\top \boldsymbol{b}|\}_{i=1}^n\big). \tag{5}$$

[1] Examples include bounded errors with $p_{\mathrm{tail}}(u) = \mathbb{I}\{u < \sigma\}$ for $\sigma \ge 0$, sub-Gaussian errors $p_{\mathrm{tail}}(u) = \min\{1, 2e^{-u^2/(2v^2)}\}$ with $v > 0$, and sub-exponential errors $p_{\mathrm{tail}}(u) = \min\{1, 2e^{-u/v}\}$ with $v > 0$.

The empirical quantile Q_q has rank $1 + \lceil q(n-1) \rceil$, and observation ties use increasing index. With $c_{\mathrm{G}} = \sqrt{2/\pi}$, the noiseless population LAD loss equals $c_{\mathrm{G}}\{p\|\boldsymbol{v} - \boldsymbol{\beta}_1^*\| + (1-p)\|\boldsymbol{v} - \boldsymbol{\beta}_2^*\|\}$.[2] The upper residual quantile retains information from the smaller component when $\boldsymbol{b}$ approaches the majority coefficient. SARP-P declares initialization failure if $\tau = 0$.

With $g(s) = 1 - e^{-s^2/2}$, form the matrix

$$\widehat{\boldsymbol{T}} = \frac{1}{n}\sum_{i=1}^{n} g\left(\frac{y_i - \boldsymbol{x}_i^\top \boldsymbol{b}}{\tau}\right)(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \boldsymbol{I}_d). \tag{6}$$

Let $\widehat{\boldsymbol{u}}$ be a unit leading eigenvector of $\widehat{\boldsymbol{T}}$. Set $c_\tau = 1/32$ and $C_\tau = 144/\sqrt{p_0}$, and choose a dimensionless grid spacing $\delta_{\mathrm{g}} > 0$. For $N_{\mathrm{g}} = \lceil 4/(c_\tau \delta_{\mathrm{g}}) \rceil$, use

$$\mathcal{G} = \left\{ -\frac{2\tau}{c_\tau} + \frac{4k\tau}{c_\tau N_{\mathrm{g}}} : k = 0, \ldots, N_{\mathrm{g}} \right\}. \tag{7}$$

Adjacent grid points are separated by $4\tau/(c_\tau N_{\mathrm{g}}) \le \delta_{\mathrm{g}}\tau$, giving a maximum distance $\delta_{\mathrm{g}}\tau/2$ from any point of $[-2\tau/c_\tau, 2\tau/c_\tau]$ to $\mathcal{G}$. Theorem 4.1 states the sufficient spacing and proves $c_\tau \Delta \le \tau \le C_\tau \Delta$. Set

$$(\widehat{t}_1, \widehat{t}_2) \in \arg\min_{(t_1,t_2)\in\mathcal{G}^2} \frac{1}{n}\sum_{i=1}^{n} \min\{|y_i - \boldsymbol{x}_i^\top(\boldsymbol{b} + t_1\widehat{\boldsymbol{u}})|, |y_i - \boldsymbol{x}_i^\top(\boldsymbol{b} + t_2\widehat{\boldsymbol{u}})|, \tau\}, \tag{8}$$

$$\boldsymbol{\beta}_{j,0} = \boldsymbol{b} + \widehat{t}_j\widehat{\boldsymbol{u}}, \qquad j = 1, 2. \tag{9}$$

The vector $\boldsymbol{\theta}_0$ stacks $\boldsymbol{\beta}_{1,0}$ and $\boldsymbol{\beta}_{2,0}$. Appendix C.3, Proposition C.3, jointly analyzes the LAD estimate, bounded residual spectrum, and clipped endpoint search using the same observations.

3.2 NORMALIZED COMPOSITE ITERATION

Normalized composite iteration uses the two-component quadratic representation of Chen et al. (2018). The derivative normalization below bounds each weighted residual derivative independently of the response-outlier amplitude.

The label-elimination polynomial is $q_i(\boldsymbol{\theta}) = (y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}_1)(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}_2)$. For noiseless inliers, q_i vanishes at the true coefficient pair. At iterate $\boldsymbol{\theta}_t$, define $r_{ij,t} = y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}_{j,t}$. Let I_t index the $m = n - \lfloor \bar{\epsilon} n \rfloor$ smallest values of $\min_{j\in\{1,2\}} |r_{ij,t}|$. Fix $a \in (0, 1]$ and set

$$w_{i,t} = \frac{\mathbb{I}\{i \in I_t\}}{a\tau + |r_{i1,t}| + |r_{i2,t}|}, \qquad F_t(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} |w_{i,t} q_i(\boldsymbol{\theta})|. \tag{10}$$

The offset $a\tau$ prevents singular weights. For the row derivative $\boldsymbol{J}_{i,t} = -(r_{i2,t}\boldsymbol{x}_i^\top, r_{i1,t}\boldsymbol{x}_i^\top)$, normalization gives $\|w_{i,t}\boldsymbol{J}_{i,t}\| \le \|\boldsymbol{x}_i\|$, independently of y_i.[3] Each composite update holds $w_{i,t}$ fixed. The next composite update recomputes the weights at $\boldsymbol{\theta}_{t+1}$.

Let $\boldsymbol{X}$ have rows $\boldsymbol{x}_i^\top$, let $\boldsymbol{w}_t = (w_{i,t})_{i=1}^n$, and define $\Lambda = \|\boldsymbol{X}^\top\boldsymbol{X}/n\|_{\mathrm{op}}$, $\overline{L} = \Lambda/(a\tau)$, and $\ell_t = \|\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{w}_t)\boldsymbol{X}/n\|_{\mathrm{op}} \le \overline{L}$. Here diag forms a diagonal matrix. Set $\mu = p_0/\{2048(aC_\tau + 34)\}$ and use the fixed closed Euclidean ball $\Omega = B(\boldsymbol{\theta}_0, r)$ with $r = \mu/(16\overline{L})$. The radius r restricts composite updates and accepted response-polishing estimates to a common neighborhood of $\boldsymbol{\theta}_0$. Define

$$M_t(\boldsymbol{d}) = \frac{1}{n}\sum_{i=1}^{n} |w_{i,t}\{q_i(\boldsymbol{\theta}_t) + \boldsymbol{J}_{i,t}\boldsymbol{d}\}| + \frac{\ell_t}{2}\|\boldsymbol{d}\|^2, \qquad D_t = \Omega - \boldsymbol{\theta}_t. \tag{11}$$

Appendix B gives a conic formulation and a computable objective-gap certificate for M_t, proved in Appendix C, Proposition C.10. We use Clarabel (Goulart & Chen, 2026) to solve the convex model. The required objective gap $\eta_t \ge 0$ means $M_t(\boldsymbol{d}_t) \le \min_{\boldsymbol{d}\in D_t} M_t(\boldsymbol{d}) + \eta_t$.

[2] For $\boldsymbol{a} \in \mathbb{R}^d$ and $Z \sim N(0,1)$, $\boldsymbol{x}^\top\boldsymbol{a}$ has distribution $\|\boldsymbol{a}\|Z$. Thus $\mathbb{E}|\boldsymbol{x}^\top\boldsymbol{a}| = \|\boldsymbol{a}\|\sqrt{2/\pi}\int_0^\infty z e^{-z^2/2}\,\mathrm{d}z = c_{\mathrm{G}}\|\boldsymbol{a}\|$. Conditioning on z gives the displayed population loss. At $p = 1/2$, the triangle inequality gives minimum $c_{\mathrm{G}}\Delta/2$, attained exactly on S.

[3] The exact norm is $\mathbb{I}\{i \in I_t\}\|\boldsymbol{x}_i\|\sqrt{r_{i1,t}^2 + r_{i2,t}^2}/(a\tau + |r_{i1,t}| + |r_{i2,t}|)$. Apply $\sqrt{s^2+t^2} \le |s| + |t|$ and $a\tau > 0$.

3.3 RESPONSE POLISHING

Response polishing selects a robust loss from fitted residuals and applies weighted least squares. A final acceptance test preserves recovery without a supplied noise scale.

After $T \geq 1$ composite updates, set $\boldsymbol{c} = \boldsymbol{\theta}_T$. Write $a_i = \min_{j\in\{1,2\}} |y_i - \boldsymbol{x}_i^\top \boldsymbol{c}_j|$, where $\boldsymbol{c}_j$ is coefficient block j, and let J index the $m = n - \lfloor \bar{\epsilon} n \rfloor$ smallest a_i. Define

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-u^2/2}\, \mathrm{d}u, \qquad q = \Phi^{-1}(3/4), \qquad s_{\mathrm{raw}} = \frac{\mathrm{med}_{i\in J}\, a_i}{q}. \tag{12}$$

The median averages the middle two observations when m is even. If $s_{\mathrm{raw}} = 0$, return $\boldsymbol{c}$. Otherwise put $\widehat{s} = \max\{s_{\mathrm{raw}}, \upsilon\}$, where $\upsilon \geq 0$ is a numerical scale floor. The normal quantile q calibrates the residual median to Gaussian standard-deviation units. The algorithm also applies to non-Gaussian errors.

For $\lambda > 0$, use the Huber loss (Huber, 1992) and Tukey bisquare loss (Bai et al., 2012, Section 2.2), with their derivatives as score functions:

$$H_\lambda(u) = \begin{cases} u^2/2, & |u| \leq \lambda, \\ \lambda|u| - \lambda^2/2, & |u| > \lambda, \end{cases} \qquad \psi_{\mathrm{H},\lambda}(u) = \mathrm{clip}(u, [-\lambda, \lambda]), \tag{13}$$

$$T_\lambda(u) = \frac{\lambda^2}{6}\{1 - (1 - u^2/\lambda^2)_+^3\}, \qquad \psi_{\mathrm{T},\lambda}(u) = u(1 - u^2/\lambda^2)_+^2. \tag{14}$$

Here $\mathrm{clip}(u, [-\lambda, \lambda]) = \max(-\lambda, \min(\lambda, u))$, $v_+ = \max\{v, 0\}$, $\psi_{\mathrm{H},\lambda} = H'_\lambda$, and $\psi_{\mathrm{T},\lambda} = T'_\lambda$. For calibration use $\psi'_{\mathrm{H},\lambda}(u) = \mathbb{I}\{|u| \leq \lambda\}$ and $\psi'_{\mathrm{T},\lambda}(u) = (1 - u^2/\lambda^2)(1 - 5u^2/\lambda^2)\mathbb{I}\{|u| < \lambda\}$, including the negative values of the latter. The derivative assigned at the Huber breakpoints is a convention. Set

$$\mathcal{C}_{\mathrm{H}} = \{0.5, 0.75, 1, 1.345, 2, 3, 5, 10\}, \qquad \mathcal{C}_{\mathrm{T}} = \{2, 3, 4.685, 6, 10\}. \tag{15}$$

For $f \in \{\mathrm{H}, \mathrm{T}\}$ and $c \in \mathcal{C}_f$, let $A_{f,c} = m^{-1} \sum_{i\in J} \psi'_{f,c\widehat{s}}(a_i)$. On candidates with $A_{f,c} > 0$, define

$$V_{f,c} = \frac{m^{-1} \sum_{i\in J} \psi_{f,c\widehat{s}}(a_i)^2}{A_{f,c}^2}, \qquad (\widehat{f}, \widehat{c}) \in \arg\min_{(f,c):\, A_{f,c}>0} V_{f,c}. \tag{16}$$

Set $\widehat{\lambda} = \widehat{c}\widehat{s}$ and hold $\widehat{f}$, $\widehat{s}$, and $\widehat{\lambda}$ fixed throughout response polishing. Ties use Huber before Tukey, then increasing c. At least one Huber candidate has $A_{f,c} > 0$ whenever $s_{\mathrm{raw}} > 0$. Appendix E derives the variance ratio motivating (16) for known labels and independent errors.

At a response iterate $\boldsymbol{v} = (\boldsymbol{v}_1^\top, \boldsymbol{v}_2^\top)^\top \in \mathbb{R}^{2d}$, write $r_{ij}(\boldsymbol{v}) = y_i - \boldsymbol{x}_i^\top \boldsymbol{v}_j$ and $\lambda = \widehat{\lambda}$. Define

$$\ell_i^{\mathrm{H}}(\boldsymbol{v}) = -\widehat{s}^2 \log\left\{ \frac{1}{2} \sum_{j=1}^{2} e^{-H_\lambda(r_{ij}(\boldsymbol{v}))/\widehat{s}^2} \right\}, \qquad \omega_{ij}(\boldsymbol{v}) = \frac{e^{-H_\lambda(r_{ij}(\boldsymbol{v}))/\widehat{s}^2}}{\sum_{k=1}^{2} e^{-H_\lambda(r_{ik}(\boldsymbol{v}))/\widehat{s}^2}}. \tag{17}$$

Let $I^{\mathrm{H}}(\boldsymbol{v})$ index the m smallest $\ell_i^{\mathrm{H}}(\boldsymbol{v})$. For the Tukey loss, let $b_i(\boldsymbol{v}) = \arg\min_{j\in\{1,2\}} |r_{ij}(\boldsymbol{v})|$ and let $I^{\mathrm{T}}(\boldsymbol{v})$ index the m smallest $\min_{j\in\{1,2\}} |r_{ij}(\boldsymbol{v})|$. Observation and label ties use increasing index. The update weights are

$$w_{ij}^{\mathrm{H}}(\boldsymbol{v}) = \mathbb{I}\{i \in I^{\mathrm{H}}(\boldsymbol{v})\} \omega_{ij}(\boldsymbol{v}) \min\{1, \lambda/|r_{ij}(\boldsymbol{v})|\}, \tag{18}$$

$$w_{ij}^{\mathrm{T}}(\boldsymbol{v}) = \mathbb{I}\{i \in I^{\mathrm{T}}(\boldsymbol{v}),\ b_i(\boldsymbol{v}) = j\}(1 - r_{ij}(\boldsymbol{v})^2/\lambda^2)_+^2, \tag{19}$$

$$\boldsymbol{v}_j^{(k+1)} \in \arg\min_{\boldsymbol{b}\in\mathbb{R}^d} \sum_{i=1}^{n} w_{ij}^{\widehat{f}}(\boldsymbol{v}^{(k)})(y_i - \boldsymbol{x}_i^\top \boldsymbol{b})^2, \qquad j = 1, 2. \tag{20}$$

The Huber ratio at zero equals one. Weighted least squares uses the minimum-norm minimizer.[4] Each update minimizes a quadratic majorizer of a trimmed response loss. Appendix D, Proposition D.3, proves descent for both response losses. The equal working priors in the Huber update need not equal

[4]The implemented Tukey polishing returns $\boldsymbol{c}$ if either set $\{i : w_{ij}^{\mathrm{T}}(\boldsymbol{v}^{(k)}) > 0\}$ has fewer than d elements. Numerical solver failures also return $\boldsymbol{c}$. Theorem 4.2 covers both fallback returns.

the true mixture weights. Theorem 4.2 establishes recovery for either score family and any finite number of response-polishing updates.

Store $F = F_{T-1}$ from the last composite update. For the final response iterate $\boldsymbol{z}$, return

$$\boldsymbol{\theta}_{\mathrm{out}} = \begin{cases} \boldsymbol{z}, & \boldsymbol{z} \in \Omega \text{ and } F(\boldsymbol{z}) \le F(\boldsymbol{c}) + \widehat{s}, \\ \boldsymbol{c}, & \text{otherwise.} \end{cases} \tag{21}$$

The acceptance test uses $\Omega = B(\boldsymbol{\theta}_0, r)$ and weights computed at $\boldsymbol{\theta}_{T-1}$. Theorem 4.2 bounds $\widehat{s}$ and the resulting estimation error of $\boldsymbol{\theta}_{\mathrm{out}}$.

Algorithm 1 SARP-P: Segment initialization, normalized composite iteration, and response polishing

Require: Observations, p_0, $\bar{\epsilon}$, a, δ_{g}, integer $T \ge 1$, gaps $(\eta_t)_{t=0}^{T-1}$, response-polishing limits $L_{\mathrm{H}}, L_{\mathrm{T}} \in \{1, 2, \ldots\}$, $\epsilon_{\mathrm{r}} > 0$, and $\upsilon \ge 0$

$(\boldsymbol{b}, \tau) \leftarrow$ (5)
if $\tau = 0$ **then return** initialization failure
$\widehat{\boldsymbol{u}} \leftarrow \arg\max_{\|\boldsymbol{u}\|=1} \boldsymbol{u}^\top \widehat{\boldsymbol{T}} \boldsymbol{u}$ using (6)
$\boldsymbol{\theta}_0 \leftarrow$ (9), $r \leftarrow \mu a \tau / (16\Lambda)$, $\Omega \leftarrow B(\boldsymbol{\theta}_0, r)$
for $t = 0, \ldots, T-1$ **do**
    $I_t \in \arg\min_{I \subseteq [n],\, |I|=m} \sum_{i \in I} \min_{j=1,2} |y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_{j,t}|$
    $(w_{i,t})_{i=1}^n \leftarrow$ (10), $M_t, D_t \leftarrow$ (11)
    $\boldsymbol{d}_t \in D_t$ with $M_t(\boldsymbol{d}_t) - \min_{\boldsymbol{d} \in D_t} M_t(\boldsymbol{d}) \le \eta_t$
    $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \boldsymbol{d}_t$
end for
$\boldsymbol{c} \leftarrow \boldsymbol{\theta}_T$, $F \leftarrow F_{T-1}$, $s_{\mathrm{raw}} \leftarrow$ (12)
if $s_{\mathrm{raw}} = 0$ **then return** $\boldsymbol{c}$
$\widehat{s} \leftarrow \max\{s_{\mathrm{raw}}, \upsilon\}$, $(\widehat{f}, \widehat{c}) \leftarrow$ (16), $\widehat{\lambda} \leftarrow \widehat{c}\widehat{s}$, $\boldsymbol{v}^{(0)} \leftarrow \boldsymbol{c}$
for $k = 0, \ldots, L_{\widehat{f}} - 1$ **do**
    $w_{ij}^{\widehat{f}} \leftarrow$ (19) at $\boldsymbol{v}^{(k)}$
    $\boldsymbol{v}^{(k+1)} \leftarrow$ (20), $\boldsymbol{z} \leftarrow \boldsymbol{v}^{(k+1)}$
    if $\|\boldsymbol{v}^{(k+1)} - \boldsymbol{v}^{(k)}\| < \epsilon_{\mathrm{r}}$ **then break**
end for
return $\boldsymbol{\theta}_{\mathrm{out}}$ defined by the acceptance test (21)

4 FINITE-SAMPLE GUARANTEES

We first analyze segment initialization and the normalized composite iteration. We then bound the output of the complete SARP-P algorithm, including response polishing. In Section 4, $u \ge 0$ is a deterministic threshold used only in the analysis.

Theorem 4.1 (Segment initialization and normalized composite recovery)**.** *Fix $p_0 \in (0, 1/2]$ and $a \in (0, 1]$. Use c_τ, C_τ, μ from Section 3 and choose*

$$0 < \delta_{\mathrm{g}} \le \frac{p_0 \mu a c_\tau}{8192\sqrt{2} C_\tau}. \tag{22}$$

There exist constants $C, \epsilon_0, s_0 > 0$, depending only on p_0, a, for which the following holds. Let $\delta \in (0, 1/2)$ and suppose

$$n \ge C\{d + \log(1/\delta)\} \log(en/\delta), \qquad \bar{\epsilon} \le \epsilon_0, \qquad u \le s_0 \Delta. \tag{23}$$

Assume either of the two tail-budget conditions

$$\epsilon + p_{\mathrm{tail}}(u) + \sqrt{\frac{\log(2/\delta)}{2n}} \le \bar{\epsilon}, \qquad \textit{or} \qquad n p_{\mathrm{tail}}(u) \le \delta/2. \tag{24}$$

With probability at least $1 - \delta$ over the clean covariates, labels, and errors, the following statements hold simultaneously for every choice of response outliers allowed by (3)*.*

(i) The segment initializer satisfies

$$c_\tau \Delta \le \tau \le C_\tau \Delta, \qquad \mathcal{E}(\boldsymbol{\theta}_0) \le r/4. \tag{25}$$

(ii) Write $\mathcal{E}_t = \mathcal{E}(\boldsymbol{\theta}_t)$. Every feasible normalized composite update in Algorithm 1, with model objective gap at most η_t, satisfies

$$\mathcal{E}_{t+1} \le \frac{\ell_t}{\mu}\mathcal{E}_t^2 + \frac{6u+\eta_t}{\mu}, \qquad \ell_t \le \overline{L}. \tag{26}$$

For $\eta_t \le (\mu r/2)2^{-2^{t+1}}$, the iterates satisfy $\mathcal{E}_t \le r2^{-2^t} + 12u/\mu$ for $0 \le t \le T$.

Theorem 4.1 guarantees initialization and normalized composite iteration using the same observations. The recursion includes inlier noise without the contamination-dependent linear term in the ILTS guarantee (Shen & Sanghavi, 2019). Compared with robust phase retrieval (Duchi & Ruan, 2019), the proof establishes segment initialization and uniform local growth for response-dependent derivatives.

The alternatives in (24) count noise exceedances as additional outliers or bound every error by u. Bounded errors use $u = \sigma$. Appendix D.5 gives explicit Gaussian and Laplace thresholds with and without spare outlier budget. The resulting quantile-dependent floors do not establish a vanishing stochastic estimation rate. Appendix C proves Theorem 4.1 with explicit sufficient constants.

Theorem 4.2 (Recovery of the complete SARP-P estimator)**.** *Under the hypotheses of Theorem 4.1, let $T \ge 1$, $P_T = \ell_{T-1}\mathcal{E}_{T-1}^2 + 6u + \eta_{T-1}$, and $C_{\mathrm{g}} = 1 + 4/(q\mu)$, with q defined in* (12)*. On the probability event of Theorem 4.1, Algorithm 1 returns*

$$\mathcal{E}(\boldsymbol{\theta}_{\mathrm{out}}) \le \frac{C_{\mathrm{g}}}{\mu}P_T + \frac{u}{q\mu} + \frac{\upsilon}{\mu}. \tag{27}$$

The conclusion includes early returns and any measurable proposal satisfying (21)*. The score family, cutoff, scale, and proposal may use the observations from segment initialization and normalized composite iteration.*

Theorem 4.2 bounds the complete SARP-P output after selecting a score, scale, and cutoff from fitted residuals. Acceptance preserves recovery without implying improved error. Appendix D gives the complete proof, and Appendix E derives the calibration criterion under known labels.

Proof mechanism. Segment initialization uses a rank-one Gaussian residual identity and a clipped score separating the endpoints of S. Appendix C.3, Lemma C.5 and Proposition C.3, control anchor error, noise, and outliers uniformly over fitted anchors, scales, and coefficient pairs, proving initialization without sample splitting.

For normalized composite iteration, let $F_{\boldsymbol{\vartheta},I}$ denote (10) with weights frozen at $\boldsymbol{\vartheta} \in \Omega$ and retained set I. Write $F^{\star}_{\boldsymbol{\vartheta},I}$ for its common value on the truth orbit Θ^* in (4). The product residual is invariant to exchanging the two coefficients. Appendix C, Proposition C.8, proves simultaneously for every $\boldsymbol{\vartheta}, \boldsymbol{\theta} \in \Omega$ and $|I| \ge n - \lfloor \bar{\epsilon} n \rfloor$ that

$$F_{\boldsymbol{\vartheta},I}(\boldsymbol{\theta}) - F^{\star}_{\boldsymbol{\vartheta},I} \ge \mu\mathcal{E}(\boldsymbol{\theta}) - 6u. \tag{28}$$

Normalization bounds the outlier contribution independently of its amplitude, while uniformity permits adaptive weights and trimming. Put $F_t^{\star} = F^{\star}_{\boldsymbol{\theta}_t,I_t}$ and $f_t(\boldsymbol{d}) = M_t(\boldsymbol{d}) - \ell_t\|\boldsymbol{d}\|^2/2$. Quadratic expansion and the model objective-gap bound give

$$|F_t(\boldsymbol{\theta}_t + \boldsymbol{d}) - f_t(\boldsymbol{d})| \le \frac{1}{n}\sum_{i=1}^{n} w_{i,t}|\boldsymbol{x}_i^{\top}\boldsymbol{d}_1|\,|\boldsymbol{x}_i^{\top}\boldsymbol{d}_2| \le \frac{\ell_t}{2}\|\boldsymbol{d}\|^2, \tag{29}$$

$$F_t(\boldsymbol{\theta}_{t+1}) \le M_t(\boldsymbol{d}_t) \le F_t^{\star} + \ell_t\mathcal{E}_t^2 + \eta_t. \tag{30}$$

Here $\boldsymbol{d}_1, \boldsymbol{d}_2$ are the two blocks of $\boldsymbol{d}$. The final inequality in (29) uses $2|st| \le s^2 + t^2$. Initialization and Appendix C, Lemma C.11, ensure that a nearest true pair is feasible for each model comparison in (30). Combining (28) and (30) proves the recursion. For response polishing, Appendix D, Lemma D.4, gives $\widehat{s} \le (u + 4\mathcal{E}(\boldsymbol{c}))/q + \upsilon$. The acceptance test then yields $\mu\mathcal{E}(\boldsymbol{\theta}_{\mathrm{out}}) \le P_T + \widehat{s}$, proving Theorem 4.2.

5 EXPERIMENTS

We compare SARP-P with robust MLR algorithms at fixed noise variance and outlier fraction while varying the inlier distribution. An ablation study isolates response polishing.

Table 1: Coefficient recovery with simultaneous inlier noise and 5% response outliers. Entries are medians of the permutation-invariant error in (4) over thirty datasets, in units of 10^{-3}, with separation $\Delta = 1$ and inlier standard deviation $s = 0.03$. The column before polishing reports $\boldsymbol{\theta}_T$ from the same SARP-P run. Bold indicates the smallest median in each row.

Noise	Before polishing	SARP-P	Mix-IRLS	Mixregt	Laplace EM	ILTS
Laplace	7.71	**7.18**	8.81	7.31	7.89	8.85
Scale mixture	4.00	**3.23**	7.79	3.88	4.08	4.04
Gaussian	11.16	**9.34**	9.43	11.06	11.25	14.27
Uniform	15.94	9.33	**9.29**	16.49	15.85	23.47

Each dataset has $n = 1000$ independent $N(\mathbf{0}, \boldsymbol{I}_{20})$ covariates and balanced labels. For random orthonormal vectors $\boldsymbol{m}, \boldsymbol{u} \in \mathbb{R}^{20}$, set $\boldsymbol{\beta}_1^* = 2\boldsymbol{m} + \boldsymbol{u}/2$ and $\boldsymbol{\beta}_2^* = 2\boldsymbol{m} - \boldsymbol{u}/2$, giving $\Delta = 1$. Inlier errors have standard deviation $s = 0.03$ and one of four distributions: Laplace $(0, s/\sqrt{2})$, Gaussian $N(0, s^2)$, uniform on $[-\sqrt{3}s, \sqrt{3}s]$, or the scale mixture $0.9N(0, s^2/9) + 0.1N(0, 9s^2)$. Independently selected 5% of responses are replaced by $2\boldsymbol{x}_i^\top \boldsymbol{m} + 5Z_i$, with independent $Z_i \sim N(0, 1)$. Covariates are unchanged. Each comparison uses the same thirty independently generated datasets. All four noise distributions satisfy the noise condition in Section 2, and equal variances isolate distributional effects.

We compare with Mix-IRLS (Zilber & Nadler, 2023), ILTS (Shen & Sanghavi, 2019), Student t mixture regression (Mixregt) (Yao et al., 2014), and Laplace EM (Song et al., 2014). Mix-IRLS uses its published outlier-handling option and random initialization. ILTS, Mixregt, and Laplace EM use the same segment initialization as SARP-P. ILTS receives the true component proportion for its retained subset size. SARP-P receives $p_0 = 0.2$, $a = 0.1$, and $\bar{\epsilon} = 0.05$, but no true labels, coefficients, or noise scale. Mixregt and Laplace EM use their published likelihoods without added trimming. The practical grid spacing is 0.05τ on $[-3\tau, 3\tau]$, followed by scalar refinement, and the composite radius is 2τ. These practical parameters are not covered by the sufficient constants of Theorem 4.1. Appendix F.4 quantifies the gap. Appendix F specifies solver settings and reports additional noise levels, response outliers near one coefficient, and response outliers selected by covariate projections.

Table 1 isolates the gain from response polishing and shows the largest improvement over robust baselines under scale-mixture noise. The Laplace gain over Mixregt is smaller. Gaussian and uniform errors give accuracy close to Mix-IRLS. Every response proposal passes the acceptance test in the primary study. Appendix F.8 separates initialization, composite iteration, score selection, and acceptance, and Appendix F.9 varies the supplied proportion and contamination bounds. Appendix F, Figure 1 and Tables 2 and 4, report additional noise levels and outlier mechanisms with uncertainty summaries. Appendix F.3 specifies restart, tuning, and runtime conventions.

6 CONCLUSION

We developed SARP-P for two-component mixed linear regression with inlier noise and response outliers. A joint finite-sample analysis establishes segment initialization, a quadratic composite error recursion up to a noise floor, and recovery after response polishing, using the same observations.

We focus our analysis on the two-component setting with Gaussian covariates. Although this is the standard benchmark for MLR theory, a direction for future work is to extend the normalized composite recovery framework to more components and general covariate distributions.

AI USE STATEMENT

The authors prepared the draft of this work. Large language models were utilized for writing assistance and text polishing.

ETHICS STATEMENT

This work focuses on the theoretical analysis and synthetic numerical evaluation of robust mixed linear regression estimators. No human subjects, animal experiments, sensitive personal data, or potentially harmful experimental procedures are involved in this research. We are not aware of any ethical concerns associated with the content of this submission.

REPRODUCIBILITY STATEMENT

The supplementary material contains implementation details and complete numerical records. Appendix F states the data distributions and numerical conventions. The proofs include the common probability event, initialization analysis, uniform local geometry, and explicit sufficient constants.

REFERENCES

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15 (80):2773–2832, 2014.

Gabriel Arpino and Ramji Venkataramanan. Statistical-computational tradeoffs in mixed sparse linear regression. In *COLT*, 2023.

Xiuqin Bai, Weixin Yao, and John E. Boyer. Robust fitting of mixture regression models. *Computational Statistics & Data Analysis*, 56(7):2347–2359, 2012.

Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

Adarsh Barik and Jean Honorio. Sparse mixed linear regression with guarantees: Taming an intractable problem with invex relaxation. In *ICML*, 2022.

Arun Tejasvi Chaganty and Percy Liang. Spectral experts for estimating mixtures of linear regressions. In *ICML*, 2013.

Vasileios Charisopoulos, Damek Davis, Mateo Díaz, and Dmitriy Drusvyatskiy. Composite optimization for robust rank one bilinear sensing. *Information and Inference: A Journal of the IMA*, 10(2): 333–396, 2021.

Sitan Chen, Jerry Li, and Zhao Song. Learning mixtures of linear regressions in subexponential time via Fourier moments. In *STOC*, 2020.

Yudong Chen, Xinyang Yi, and Constantine Caramanis. A convex formulation for mixed regression with two components: Minimax optimal rates. In *COLT*, 2014.

Yudong Chen, Xinyang Yi, and Constantine Caramanis. Convex and nonconvex formulations for mixed regression with two components: Minimax optimal rates. *IEEE Transactions on Information Theory*, 64(3):1738–1766, 2018.

Richard D. De Veaux. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, 8 (3):227–245, 1989.

Wayne S. DeSarbo and William L. Cron. A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5(2):249–282, 1988.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*. springer New York, 1996.

Ilias Diakonikolas and Daniel M. Kane. Small covers for near-zero sets of polynomials and learning latent variable models. In *FOCS*, 2020.

John C. Duchi and Feng Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Information and Inference: A Journal of the IMA*, 8(3):471–529, 2019.

Jianqing Fan, Han Liu, Zhaoran Wang, and Zhuoran Yang. Curse of heterogeneity: Computational barriers in sparse mixture models and phase retrieval. *arXiv preprint arXiv:1808.06996*, 2018.

Gerald B Folland. *Real analysis: modern techniques and their applications*. John Wiley & Sons, 1999.

Scott Gaffney and Padhraic Smyth. Trajectory clustering with mixtures of regression models. In *KDD*, 1999.

L. A. García-Escudero, A. Gordaliza, A. Mayo-Iscar, and R. San Martín. Robust clusterwise linear regression through trimming. *Computational Statistics & Data Analysis*, 54(12):3057–3069, 2010.

Avishek Ghosh and Ramchandran Kannan. Alternating minimization converges super-linearly for mixed linear regression. In *AISTATS*, 2020.

Avishek Ghosh and Arya Mazumdar. Agnostic learning of mixed linear regressions with EM and AM algorithms. In *ICML*, 2024.

Paul J. Goulart and Yuwen Chen. Clarabel: An interior-point solver for conic programs with quadratic objectives. *Mathematical Programming Computation*, 2026.

Paul Hand and Babhru Joshi. A convex program for mixed linear regression with a recovery guarantee for well-separated data. *Information and Inference: A Journal of the IMA*, 7(3):563–579, 2018.

Gao Huang, Song Li, and Hang Xu. Robust outlier bound condition to phase retrieval with adversarial sparse outliers. *Applied and Computational Harmonic Analysis*, pp. 101819, 2025.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. 1992.

Jason M. Klusowski, Dana Yang, and W. D. Brinda. Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *IEEE Transactions on Information Theory*, 65(6): 3515–3524, 2019.

Jeongyeol Kwon and Constantine Caramanis. EM converges for a mixture of many linear regressions. In *AISTATS*, 2020.

Jeongyeol Kwon, Wei Qian, Constantine Caramanis, Yudong Chen, and Damek Davis. Global convergence of the EM algorithm for mixtures of two component linear regression. In *COLT*, 2019.

Jeongyeol Kwon, Nhat Ho, and Constantine Caramanis. On the minimax optimality of the EM algorithm for learning two-component mixed linear regression. In *AISTATS*, 2021.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

Yuanzhi Li and Yingyu Liang. Learning mixtures of linear regressions with nearly optimal complexity. In *COLT*, 2018.

Zhankun Luo and Abolfazl Hashemi. Unveiling the cycloid trajectory of EM iterations in mixed linear regression. In *ICML*, 2024.

Zhankun Luo and Abolfazl Hashemi. Characterizing evolution in expectation-maximization estimates for overspecified mixed linear regression. *arXiv preprint arXiv:2508.10154*, 2025.

Zhankun Luo and Abolfazl Hashemi. Connecting score matching, maximum likelihood, and expectation-maximization in mixed linear regression. *arXiv preprint arXiv:2609.05688*, 2026.

N. Neykov, Peter Filzmoser, R. Dimova, and P. Neytchev. Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics & Data Analysis*, 52(1):299–308, 2007.

Richard E. Quandt and James B. Ramsey. Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, 73(364):730–738, 1978.

Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1): 145–147, 1972.

Hanie Sedghi, Majid Janzamin, and Anima Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. In *AISTATS*, 2016.

Yanyao Shen and Sujay Sanghavi. Iterative least trimmed squares for mixed linear regression. In *NeurIPS*, 2019.

Weixing Song, Weixin Yao, and Yanru Xing. Robust mixture regression model fitting by Laplace distribution. *Computational Statistics & Data Analysis*, 71:128–137, 2014.

Nicolas Städler, Peter Bühlmann, and Sara van de Geer. ℓ_1-penalization for mixture regression models. *TEST*, 19(2):209–256, 2010.

Nelvin Tan and Ramji Venkataramanan. Mixed regression via approximate message passing. *Journal of Machine Learning Research*, 24(317):1–44, 2023.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity: festschrift for alexey chervonenkis*, pp. 11–30. 2015.

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

Kert Viele and Barbara Tong. Modeling with mixtures of linear regressions. *Statistics and Computing*, 12(4):315–330, 2002.

Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu. High dimensional EM algorithm: Statistical optimization and asymptotic normality. In *NeurIPS*, 2015.

Weixin Yao, Yan Wei, and Chun Yu. Robust mixture regression using the t-distribution. *Computational Statistics & Data Analysis*, 71:116–127, 2014.

Xinyang Yi and Constantine Caramanis. Regularized EM algorithms: A unified framework and statistical guarantees. In *NeurIPS*, 2015.

Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. In *ICML*, 2014.

Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Solving a mixture of many random linear equations by tensor decomposition and alternating minimization. *arXiv preprint arXiv:1608.05749*, 2016.

Kai Zhong, Prateek Jain, and Inderjit S. Dhillon. Mixed linear regression with multiple components. In *NeurIPS*, 2016.

Pini Zilber and Boaz Nadler. Imbalanced mixed linear regression. In *NeurIPS*, 2023.

A Auxiliary Probability and Geometric Results

The recovery proofs use scalar concentration, finite coverings, Gaussian identities, and deterministic projection bounds. We state the required versions and identify their sources before applying them to segment initialization and normalized composite iteration. The final subsection records the limit operations needed only for the oracle response-polishing calculation. Symbols $c, C > 0$ denote universal constants unless their dependence is stated.

A.1 Concentration and uniform empirical bounds

Fixed-parameter concentration is combined with finite coverings and uniform indicator bounds to control every fitted anchor, retained set, and iterate on one event. The bounds below concern the clean covariates and labels, before response outliers are introduced. The Orlicz norms are $\|U\|_{\psi_r} = \inf\{K > 0 : \mathbb{E}\exp(|U|^r/K^r) \le 2\}$ for $r \in \{1, 2\}$. The next statement collects the general Hoeffding and Bernstein inequalities in Vershynin (2018, Theorems 2.2.6, 2.6.2, and 2.8.1).

Lemma A.1 (Standard scalar concentration tools)**.** *Let $U_1, \dots, U_n$ be independent real random variables. If $\|U_i - \mathbb{E}U_i\|_{\psi_2} \le K$, then*

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n}(U_i - \mathbb{E}U_i)\right| > t\right) \le 2e^{-cnt^2/K^2}. \tag{31}$$

If $\|U_i - \mathbb{E}U_i\|_{\psi_1} \le K$, the bound is $2\exp\{-cn\min(t^2/K^2, t/K)\}$. If each variable belongs to an interval of length D, the Hoeffding bound is $2e^{-2nt^2/D^2}$, or e^{-2nt^2/D^2} for a specified one-sided deviation. Centering changes either Orlicz norm by at most a universal factor. Pointwise domination by $K|Z|$ or $K(Z^2+1)$, with $Z \sim N(0,1)$, implies respectively a sub-Gaussian or sub-exponential norm at most CK.

Lemma A.1 controls fixed parameters in the LAD, residual-spectrum, and endpoint-search processes. Domination does not require independence from Z, which is useful when a residual transform and a quadratic form share a covariate.

The domination assertions follow directly from $\mathbb{E}e^{Z^2/4} = \sqrt{2}$ and $e^{1/4}\sqrt{2} < 2$. Centering increases either norm by at most a universal factor. Pointwise domination does not require independence between a variable and the Gaussian envelope.

Lemma A.2 (Finite nets and matrix tails)**.** *A ball of radius R in $\mathbb{R}^d$ has an η_{net}-net of size at most $(1 + 2R/\eta_{\mathrm{net}})^d$, contained in the ball. The unit sphere has an η_{net}-net of size at most $(1 + 2/\eta_{\mathrm{net}})^d$. For a symmetric matrix $\boldsymbol{A}$ and such a sphere net $\mathcal{N}$, with $\eta_{\mathrm{net}} < 1/2$,*

$$\|\boldsymbol{A}\|_{\mathrm{op}} \le (1 - 2\eta_{\mathrm{net}})^{-1} \max_{\boldsymbol{w}\in\mathcal{N}} |\boldsymbol{w}^\top \boldsymbol{A}\boldsymbol{w}|. \tag{32}$$

For any rectangular matrix $\boldsymbol{D}$, the corresponding bound is $\|\boldsymbol{D}\|_{\mathrm{op}} \le (1 - \eta_{\mathrm{net}})^{-1} \max_{\mathcal{N}} \|\boldsymbol{D}\boldsymbol{w}\|$. For an $m \times d$ independent standard Gaussian matrix $\boldsymbol{G}$ and $t > 0$,

$$\mathbb{P}\{\|\boldsymbol{G}\|_{\mathrm{op}} > 2(\sqrt{m} + \sqrt{2(d\log 5 + t)})\} \le e^{-t}. \tag{33}$$

For an $n \times d$ such matrix $\boldsymbol{X}$, $n \ge C(d + \log(1/\delta))$ ensures

$$\mathbb{P}\{\|\boldsymbol{X}^\top\boldsymbol{X}/n - \boldsymbol{I}_d\|_{\mathrm{op}} > 1/2\} \le \delta, \quad \mathbb{P}\{\max_i \|\boldsymbol{x}_i\| > \sqrt{d} + \sqrt{2\log(n/\delta)}\} \le \delta. \tag{34}$$

Lemma A.2 extends scalar concentration to uniform parameter sets. The matrix tail also controls every possible replacement subset, which is necessary for an adversary that observes the sample.

The covering estimates and the rectangular extension inequality are Vershynin (2018, Corollary 4.2.13 and Lemma 4.4.1). The symmetric extension is also stated in Exercise 4.4.3(b) of the same reference. The symmetric extension follows from

$$|\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u} - \boldsymbol{w}^\top\boldsymbol{A}\boldsymbol{w}| \le 2\|\boldsymbol{u} - \boldsymbol{w}\|\,\|\boldsymbol{A}\|_{\mathrm{op}} \quad (\|\boldsymbol{u}\| = \|\boldsymbol{w}\| = 1). \tag{35}$$

For the precise constants in (33), Laurent & Massart (2000, Lemma 1, equations (4.1)–(4.3)) give

$$\mathbb{P}\{\|\boldsymbol{z}\|^2 > m + 2\sqrt{mt} + 2t\} \le e^{-t}, \qquad \boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{I}_m). \tag{36}$$

Thus $\mathbb{P}\{\|\boldsymbol{z}\| > \sqrt{m} + \sqrt{2t}\} \le e^{-t}$. For a sphere $1/2$-net $\mathcal{N}$ of size at most 5^d, put $s = t + d\log 5$. The rectangular extension and the union bound give

$$\mathbb{P}\{\|\boldsymbol{G}\|_{\mathrm{op}} > 2(\sqrt{m} + \sqrt{2s})\} \le \sum_{\boldsymbol{w}\in\mathcal{N}} \mathbb{P}\{\|\boldsymbol{G}\boldsymbol{w}\| > \sqrt{m} + \sqrt{2s}\} \le 5^d e^{-s} = e^{-t}, \tag{37}$$

which proves (33). The same norm bound and a union bound prove the maximum-row estimate. For the covariance estimate, Lemma A.1 at every point of a sphere $1/4$-net gives

$$\mathbb{P}\{\|\boldsymbol{X}^\top\boldsymbol{X}/n - \boldsymbol{I}_d\|_{\mathrm{op}} > 1/2\} \le 2\cdot 9^d e^{-cn} \le \delta \tag{38}$$

under the stated sample condition.

For a class $\mathcal{A}$ of sets, let $\Pi_{\mathcal{A}}(N)$ be the maximum number of membership patterns on N points. The quantity $\Pi_{\mathcal{A}}$ is the growth function of $\mathcal{A}$.

Lemma A.3 (Uniform indicator bound)**.** *A fixed Boolean combination of at most J affine halfspaces in $\mathbb{R}^d$ has growth function at most $(eN/(d+1))^{J(d+1)}$ when $N \ge d+1$. Intersecting with a fixed label indicator does not increase the pattern-count bound. For independent identically distributed observations W_i and $nb^2 \ge 2$,*

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}}\left|n^{-1}\sum_{i=1}^{n}\mathbb{I}\{W_i\in A\} - \mathbb{P}(W\in A)\right| > b\right\} \le 8\Pi_{\mathcal{A}}(2n)e^{-nb^2/32}. \tag{39}$$

The uniform indicator inequality (39) (Vapnik & Chervonenkis, 2015; Sauer, 1972) controls all slab intersections used in the proof. In particular, the slabs may be chosen after computing an iterate.

Affine halfspaces have Vapnik–Chervonenkis dimension $d+1$ by Devroye et al. (1996, Corollary 13.1). Their growth function is bounded by $\sum_{k=0}^{d+1}\binom{N}{k}$ by Devroye et al. (1996, Theorem 13.2). For $v = d+1 \le N$, the bound $(eN/v)^v$ follows from

$$(v/N)^v \sum_{k=0}^{v}\binom{N}{k} \le (1+v/N)^N \le e^v. \tag{40}$$

A Boolean formula with J inputs has no more membership patterns than the product of the J constituent pattern counts. Finally, Devroye et al. (1996, Theorem 12.5) give $8\Pi_{\mathcal{A}}(n)e^{-nb^2/32}$, which implies (39) by monotonicity of the growth function. The uniform-probability statements may be read in outer probability when necessary.

A.2 Gaussian identities used by the spectral initializer

The Gaussian calculations below identify the population residual spectrum and bound the probability of small projections. The spectral identity supports segment initialization, and the small-projection bounds support local growth.

Lemma A.4 (Gaussian integrals)**.** *For $Z\sim N(0,1)$ and $b\ge 0$,*

$$\mathbb{E}|Z| = c_{\mathrm{G}}, \quad \mathbb{E}Z^2 = 1, \quad \mathbb{E}Z^4 = 3, \quad \mathbb{P}(|Z|\le b) \le c_{\mathrm{G}}b, \quad \mathbb{P}(|Z|>b) \le 2e^{-b^2/2}, \tag{41}$$

$$\mathbb{E}e^{-b^2Z^2/2} = (1+b^2)^{-1/2}, \qquad \mathbb{E}Z^2e^{-b^2Z^2/2} = (1+b^2)^{-3/2}. \tag{42}$$

For $\boldsymbol{x}\sim N(\boldsymbol{0},\boldsymbol{I}_d)$, $\boldsymbol{v}\in\mathbb{R}^d$, and $\boldsymbol{w}\in\mathbb{S}^{d-1}$,

$$\mathbb{E}|\boldsymbol{x}^\top\boldsymbol{v}| = c_{\mathrm{G}}\|\boldsymbol{v}\|, \qquad \mathbb{E}\big[|\boldsymbol{x}^\top\boldsymbol{v}|\,|(\boldsymbol{x}^\top\boldsymbol{w})^2-1|\big] \le \sqrt{2}\|\boldsymbol{v}\|. \tag{43}$$

The function g in (50) *satisfies $0\le g\le 1$, $|g'|\le 1$, and $|sg'(s)|\le 2$.*

Lemma A.4 supplies the Gaussian identities for the rank-one residual spectrum and the simultaneous small-projection counts. In particular, no asymptotic spectral formula is needed.

The moments and tail bounds follow from the standard Gaussian density. For the two transform identities needed in the spectral calculation, the substitution $u = \sqrt{1+b^2}\,z$ gives directly

$$\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} z^{2k}e^{-(1+b^2)z^2/2}\,\mathrm{d}z = (1+b^2)^{-k-1/2}\mathbb{E}Z^{2k}, \qquad k\in\{0,1\}. \tag{44}$$

Gaussian rotational invariance and Cauchy–Schwarz inequality yield

$$\mathbb{E}\{|\boldsymbol{x}^\top\boldsymbol{v}|\,|(\boldsymbol{x}^\top\boldsymbol{w})^2-1|\} \le \sqrt{\mathbb{E}(\boldsymbol{x}^\top\boldsymbol{v})^2}\sqrt{\mathbb{E}((\boldsymbol{x}^\top\boldsymbol{w})^2-1)^2} = \sqrt{2}\|\boldsymbol{v}\|. \tag{45}$$

Finally, $g'(s) = se^{-s^2/2}$, with $\sup_s|g'(s)| = e^{-1/2}\le 1$ and $\sup_s|sg'(s)| = 2e^{-1}\le 2$.

A.3 Projection and rank-one perturbation

Projection onto the coefficient segment converts excess LAD loss into localization error. The rank-one perturbation estimate then converts a residual-matrix error into an error in the estimated segment direction.

Lemma A.5 (Deterministic geometric tools)**.** *For a closed segment S, let $\mathrm{Proj}_S(\boldsymbol{b}) = \arg\min_{\boldsymbol{z}\in S}\|\boldsymbol{b}-\boldsymbol{z}\|$ denote Euclidean projection. Put $\bar{\boldsymbol{b}} = \mathrm{Proj}_S(\boldsymbol{b})$. For every $\boldsymbol{v}\in S$,*

$$\langle \boldsymbol{b}-\bar{\boldsymbol{b}}, \boldsymbol{v}-\bar{\boldsymbol{b}}\rangle \le 0, \qquad \|\boldsymbol{b}-\boldsymbol{v}\|^2 \ge \|\bar{\boldsymbol{b}}-\boldsymbol{v}\|^2 + \|\boldsymbol{b}-\bar{\boldsymbol{b}}\|^2. \tag{46}$$

Let $\widehat{\boldsymbol{T}}$ be symmetric. If $\boldsymbol{T} = \lambda \boldsymbol{u}\boldsymbol{u}^\top$, $\|\boldsymbol{u}\| = 1$, $\lambda > 0$, and $\|\widehat{\boldsymbol{T}} - \boldsymbol{T}\|_{\mathrm{op}} \le E < \lambda/2$, then a top unit eigenvector $\widehat{\boldsymbol{u}}$ of $\widehat{\boldsymbol{T}}$ has a sign such that

$$\|\widehat{\boldsymbol{u}} - \boldsymbol{u}\| \le \frac{\sqrt{2}E}{\lambda - E}. \tag{47}$$

The first statement converts LAD loss growth into distance from the coefficient segment. The second transfers the residual matrix perturbation to its estimated direction.

Proof. Differentiate $\|\boldsymbol{b}-\bar{\boldsymbol{b}} - t(\boldsymbol{v}-\bar{\boldsymbol{b}})\|^2$ at its minimum over $t\in[0,1]$, attained at zero. Expansion then gives the second inequality. For the eigenvector assertion, the largest eigenvalue $\widehat{\lambda}$ is at least $\lambda - E$ by its Rayleigh quotient at $\boldsymbol{u}$. Put $\boldsymbol{P} = \boldsymbol{I}_d - \boldsymbol{u}\boldsymbol{u}^\top$. The eigenvector equation implies

$$\widehat{\lambda}\boldsymbol{P}\widehat{\boldsymbol{u}} = \boldsymbol{P}(\widehat{\boldsymbol{T}} - \boldsymbol{T})\widehat{\boldsymbol{u}}, \qquad \|\boldsymbol{P}\widehat{\boldsymbol{u}}\| \le E/(\lambda - E). \tag{48}$$

Choose the sign with $\langle \widehat{\boldsymbol{u}}, \boldsymbol{u}\rangle \ge 0$. Then $\|\widehat{\boldsymbol{u}} - \boldsymbol{u}\|^2 = 2(1 - \langle \widehat{\boldsymbol{u}}, \boldsymbol{u}\rangle) \le 2\|\boldsymbol{P}\widehat{\boldsymbol{u}}\|^2$, proving the claim. □

A.4 Limit operations in the oracle calculation

The oracle covariance proof uses convergence of empirical derivative matrices and Gaussian score sums. The following integration and limit statements specify the operations used in Appendix E. We use dominated convergence in the form of Folland (1999, Theorem 2.24): If measurable $f_N \to f$ almost everywhere on a fixed measure space with measure ν and $|f_N| \le g$ for an integrable g, then f is integrable and $\int f_N \,\mathrm{d}\nu \to \int f \,\mathrm{d}\nu$. Interchanging nonnegative integrals, or absolutely integrable product-space integrals, uses Folland (1999, Theorem 2.37). The empirical averages in Appendix E have finite variance, and their weak laws follow entrywise from the Chebyshev inequality. For matrix limit operations, Van der Vaart (2000, Lemma 2.8) implies the following: If $\boldsymbol{Z}_N \Rightarrow \boldsymbol{Z}$ and $\boldsymbol{A}_N \xrightarrow{\mathrm{p}} \boldsymbol{A}$ for deterministic invertible $\boldsymbol{A}$, then $\boldsymbol{A}_N^{-1}\boldsymbol{Z}_N \Rightarrow \boldsymbol{A}^{-1}\boldsymbol{Z}$. No independence between $\boldsymbol{A}_N$ and $\boldsymbol{Z}_N$ is required. The inverse can be assigned arbitrarily on the event of singularity, whose probability tends to zero.

B Detailed Algorithms and Optimization Certificates

The full SARP-P procedure has three stages: Segment initialization, normalized composite iteration, and response polishing with an acceptance test. Algorithm 1 in the main text uses the absolute loss, corresponding to $\kappa = 0$ below. We also specify an optional smoothed model for the general recovery argument. Appendix F records the practical grid, radius, stopping thresholds, and numerical safeguards used by the evaluated implementation.

B.1 Segment initialization on an affine line

Segment initialization first estimates an anchor near the coefficient segment, then estimates a direction and searches for both endpoints. The formulas below define the theoretical grid procedure. The implementation uses the same stages with the practical search interval and additional scalar search described in Appendix F. Compute a least absolute deviations (LAD) solution and a residual scale

$$\boldsymbol{b} \in \arg\min_{\boldsymbol{v}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} |y_i - \boldsymbol{x}_i^\top \boldsymbol{v}|, \qquad \tau = Q_{1-p_0/2}\big(\{|y_i - \boldsymbol{x}_i^\top \boldsymbol{b}|\}_{i=1}^n\big). \tag{49}$$

The noiseless population LAD loss is $c_{\mathrm{G}}\{p\|\boldsymbol{v}-\boldsymbol{\beta}_1^*\|+(1-p)\|\boldsymbol{v}-\boldsymbol{\beta}_2^*\|\}$. At balance, every point of S minimizes the population LAD loss. Thus the correct initialization target is the segment, not necessarily one coefficient. The upper residual quantile avoids a zero scale when LAD approaches a majority component.

On the recovery event $\tau > 0$. Off that event, a zero τ is declared an initialization failure before division. With $g(s) = 1 - e^{-s^2/2}$, form

$$\widehat{\boldsymbol{T}} = \frac{1}{n}\sum_{i=1}^{n} g\left(\frac{y_i - \boldsymbol{x}_i^\top \boldsymbol{b}}{\tau}\right)(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \boldsymbol{I}_d), \qquad \widehat{\boldsymbol{u}} \in \arg\max_{\|\boldsymbol{u}\|=1} \boldsymbol{u}^\top \widehat{\boldsymbol{T}}\boldsymbol{u}. \tag{50}$$

Choose a finite grid $\mathcal{G} \subset [-2\tau/c_\tau, 2\tau/c_\tau]$ with covering radius at most $\delta_{\mathrm{g}}\tau/2$. Here $c_\tau = 1/32$, and $\delta_{\mathrm{g}} > 0$ is a fixed grid spacing parameter. Select

$$(\widehat{t}_1, \widehat{t}_2) \in \arg\min_{(t_1,t_2)\in\mathcal{G}^2} V(\boldsymbol{b} + t_1\widehat{\boldsymbol{u}}, \boldsymbol{b} + t_2\widehat{\boldsymbol{u}}), \tag{51}$$

$$V(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{1}{n}\sum_{i=1}^{n} \min\{|y_i - \boldsymbol{x}_i^\top \boldsymbol{v}_1|, |y_i - \boldsymbol{x}_i^\top \boldsymbol{v}_2|, \tau\}, \qquad \boldsymbol{\beta}_{j,0} = \boldsymbol{b} + \widehat{t}_j\widehat{\boldsymbol{u}}. \tag{52}$$

The vector $\boldsymbol{\theta}_0$ stacks $\boldsymbol{\beta}_{1,0}$ and $\boldsymbol{\beta}_{2,0}$. Clipping both the spectrum and the search score bounds both perturbations independently of response-outlier amplitudes. The theoretical grid size may be large but depends only on fixed model constants, independently of d and $\|\boldsymbol{m}^*\|$.

B.2 Normalized composite iteration and its convex model

Each iteration freezes a retained set and response-dependent weights before minimizing a convex model. The normalization bounds the contribution of a response outlier to the model derivative. The fixed feasible ball allows the recovery proof to use one uniform local-growth inequality. Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ have rows $\boldsymbol{x}_i^\top$. Fix $a \in (0, 1]$ and the Huber smoothing level $\kappa \geq 0$. Define

$$\phi_\kappa(s) = \begin{cases} s^2/(2\kappa), & |s| \leq \kappa,\ \kappa > 0, \\ |s| - \kappa/2, & |s| > \kappa,\ \kappa > 0, \\ |s|, & \kappa = 0. \end{cases} \tag{53}$$

The function ϕ_κ is the Huber loss divided by the outer-branch slope (Huber, 1992). We retain $\kappa \geq 0$ in the composite recovery analysis to describe smoothing explicitly. The full SARP-P procedure uses $\kappa = 0$, thus the normalized composite iteration needs no noise scale. Response calibration follows the last normalized composite iteration.

At iteration t, let $r_{ij,t} = y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_{j,t}$. Retain the $n - \lfloor \bar{\epsilon} n \rfloor$ smallest values of $\min_{j\in\{1,2\}} |r_{ij,t}|$, and call the retained set I_t. Set

$$\begin{aligned} w_{i,t} &= \frac{\mathbb{I}\{i \in I_t\}}{a\tau + |r_{i1,t}| + |r_{i2,t}|}, \qquad q_i(\boldsymbol{\theta}) = (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_1)(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_2), \\ F_t(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} \phi_\kappa(w_{i,t} q_i(\boldsymbol{\theta})). \end{aligned} \tag{54}$$

The weights are frozen inside iteration t. We do not differentiate through the normalizer or the retained set. The row derivative of q_i at the current pair is

$$\boldsymbol{J}_{i,t} = (-r_{i2,t}\boldsymbol{x}_i^\top,\ -r_{i1,t}\boldsymbol{x}_i^\top), \quad c_{i,t} = w_{i,t} q_i(\boldsymbol{\theta}_t), \quad \boldsymbol{A}_{i,t} = w_{i,t}\boldsymbol{J}_{i,t}. \tag{55}$$

Stack $c_{i,t}$ into $\boldsymbol{c}_t \in \mathbb{R}^n$ and the rows $\boldsymbol{A}_{i,t}$ into $\boldsymbol{A}_t \in \mathbb{R}^{n\times 2d}$. With $\boldsymbol{w}_t = (w_{1,t}, \ldots, w_{n,t})^\top$, define

$$\ell_t = \left\| \frac{\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{w}_t)\boldsymbol{X}}{n} \right\|_{\mathrm{op}}, \qquad \Lambda = \|\boldsymbol{X}^\top\boldsymbol{X}/n\|_{\mathrm{op}}, \qquad \overline{L} = \frac{\Lambda}{a\tau}. \tag{56}$$

We use the fixed feasible ball $\Omega = B(\boldsymbol{\theta}_0, r)$ with $r = \mu/(16\overline{L})$, where the positive constant μ is specified below. The fixed constraint Ω places every feasible update in the domain of the uniform

Algorithm 2 Segment initialization and normalized composite iteration

Require: $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, p_0, $\bar{\epsilon}$, a, κ, δ_{g}, μ, $T \in \mathbb{N}$, gaps $(\eta_t)_{t=0}^{T-1}$
1: $\boldsymbol{b}, \tau \leftarrow$ solutions of (49)
2: **if** $\tau = 0$ **then**
3: **return** initialization failure
4: **end if**
5: $\widehat{\boldsymbol{T}}, \widehat{\boldsymbol{u}} \leftarrow$ (50), $\boldsymbol{\theta}_0 \leftarrow$ (52)
6: $\Lambda \leftarrow \|\boldsymbol{X}^\top \boldsymbol{X}/n\|_{\mathrm{op}}$, $\overline{L} \leftarrow \Lambda/(a\tau)$, $r \leftarrow \mu/(16\overline{L})$, $\Omega \leftarrow B(\boldsymbol{\theta}_0, r)$
7: **for** $t = 0, \ldots, T-1$ **do**
8: $r_{ij,t} \leftarrow y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_{j,t}$ for $i \in [n]$, $j \in \{1,2\}$
9: $I_t \leftarrow \arg\min_{I \subseteq [n],\ |I| = n - \lfloor \bar{\epsilon} n \rfloor} \sum_{i \in I} \min_{j \in \{1,2\}} |r_{ij,t}|$
10: $w_{i,t} \leftarrow \mathbb{I}\{i \in I_t\}/(a\tau + |r_{i1,t}| + |r_{i2,t}|)$
11: $\boldsymbol{c}_t \leftarrow (w_{i,t} r_{i1,t} r_{i2,t})_{i=1}^n$, $\boldsymbol{A}_{i,t} \leftarrow -w_{i,t}(r_{i2,t}\boldsymbol{x}_i^\top, r_{i1,t}\boldsymbol{x}_i^\top)$
12: $\ell_t \leftarrow \|\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{w}_t)\boldsymbol{X}/n\|_{\mathrm{op}}$, $D_t \leftarrow \Omega - \boldsymbol{\theta}_t$
13: $\boldsymbol{d}_t \in D_t$ such that $M_t(\boldsymbol{d}_t) - \min_{\boldsymbol{d} \in D_t} M_t(\boldsymbol{d}) \le \eta_t$
14: $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \boldsymbol{d}_t$
15: **end for**
16: **return** $\boldsymbol{\theta}_T$

growth bound. Both the center and radius are computable from observations and model constants. Define $D_t = \Omega - \boldsymbol{\theta}_t$ and

$$M_t(\boldsymbol{d}) = \frac{1}{n}\sum_{i=1}^n \phi_\kappa(c_{i,t} + \boldsymbol{A}_{i,t}\boldsymbol{d}) + \frac{\ell_t}{2}\|\boldsymbol{d}\|^2. \tag{57}$$

The update satisfies $\boldsymbol{d}_t \in D_t$ and $M_t(\boldsymbol{d}_t) - \min_{D_t} M_t \le \eta_t$, where $\eta_t \ge 0$ bounds the objective gap. Then $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \boldsymbol{d}_t$. The objectives F_t change with t, and the proof does not require $F_t(\boldsymbol{\theta}_t)$ to decrease.

B.3 An explicit subproblem solver and certificate

The absolute-loss iteration and the optional smoothed iteration require different numerical solvers but the same mathematical stopping condition. The conic formulation below implements the absolute loss used by SARP-P. The projected-gradient construction applies only to the optional choice $\kappa > 0$. The normalized composite iteration used by SARP-P has $\kappa = 0$. The convex model has the conic representation

$$\begin{aligned} \min_{\boldsymbol{d} \in \mathbb{R}^{2d},\, \boldsymbol{u} \in \mathbb{R}^n} \quad & \frac{1}{n}\mathbf{1}^\top \boldsymbol{u} + \frac{\ell_t}{2}\|\boldsymbol{d}\|^2 \\ \text{subject to} \quad & -\boldsymbol{u} \le \boldsymbol{c}_t + \boldsymbol{A}_t \boldsymbol{d} \le \boldsymbol{u}, \qquad \|\boldsymbol{d} - (\boldsymbol{\theta}_0 - \boldsymbol{\theta}_t)\| \le r. \end{aligned} \tag{58}$$

Here $\mathbf{1}$ is the all-ones vector, and vector inequalities are coordinatewise. The implementation solves (58) using Clarabel (Goulart & Chen, 2026). Proposition C.10 supplies a feasible-dual lower bound, including an explicit formula for its ball minimization. The mathematical update requires an objective gap, not a small coefficient change.

For the optional smoothed composite iteration with $\kappa > 0$, Algorithm 3 is a fully specified solver. The projection onto $D_t = B(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_t, r)$ is explicit:

$$\mathrm{Proj}_{B(\boldsymbol{v},r)}(\boldsymbol{u}) = \boldsymbol{v} + \min\{1, r/\|\boldsymbol{u} - \boldsymbol{v}\|\}(\boldsymbol{u} - \boldsymbol{v}), \tag{59}$$

with value $\boldsymbol{v}$ at $\boldsymbol{u} = \boldsymbol{v}$. Proposition C.9 proves the projected-gradient stopping certificate. Exact minimizers are also allowed in the analysis.

B.4 Calibrated response polishing and acceptance

The normalized product loss supplies local recovery geometry. At the true pair, the product for an inlier with label j equals $\xi_i\{\xi_i + \boldsymbol{x}_i^\top(\boldsymbol{\beta}_j^* - \boldsymbol{\beta}_{3-j}^*)\}$ and therefore contains a squared-error term.

Algorithm 3 Projected-gradient solver with an objective-gap certificate for (57), $\kappa > 0$

Require: $\boldsymbol{c}_t, \boldsymbol{A}_t, \ell_t > 0, D_t, \kappa > 0, \eta_t > 0$
$K_t \leftarrow \ell_t + \|\boldsymbol{A}_t\|_{\mathrm{op}}^2/(n\kappa)$, $\boldsymbol{d}^{(0)} \leftarrow \mathbf{0}$, $k \leftarrow 0$
loop
  $\boldsymbol{s}^{(k)} \leftarrow \mathrm{clip}((\boldsymbol{c}_t + \boldsymbol{A}_t\boldsymbol{d}^{(k)})/\kappa, [-1, 1])$
  $\boldsymbol{g}^{(k)} \leftarrow \boldsymbol{A}_t^\top \boldsymbol{s}^{(k)}/n$, $\boldsymbol{v}^{(k)} \leftarrow \mathrm{Proj}_{D_t}(-\boldsymbol{g}^{(k)}/\ell_t)$
  $G_k \leftarrow \frac{\ell_t}{2}(\|\boldsymbol{d}^{(k)}\|^2 - \|\boldsymbol{v}^{(k)}\|^2) + \langle \boldsymbol{g}^{(k)}, \boldsymbol{d}^{(k)} - \boldsymbol{v}^{(k)} \rangle$
  if $G_k \le \eta_t$ **then**
    return $\boldsymbol{d}^{(k)}$
  end if
  $\boldsymbol{d}^{(k+1)} \leftarrow \mathrm{Proj}_{D_t}(\boldsymbol{d}^{(k)} - (\ell_t\boldsymbol{d}^{(k)} + \boldsymbol{g}^{(k)})/K_t)$
  $k \leftarrow k + 1$
end loop

Response polishing instead fits scalar response residuals. The final acceptance test transfers the composite recovery bound to the polishing output.

Let $\boldsymbol{c} = \boldsymbol{\theta}_T$ for $T \ge 1$, let $m = n - \lfloor \bar{\epsilon} n \rfloor$, and retain the m smallest values $a_i = \min_{j \in \{1,2\}} |y_i - \boldsymbol{x}_i^\top \boldsymbol{c}_j|$ in a set J. Here $\boldsymbol{c}_j$ is block j of $\boldsymbol{c}$. Put $q = \Phi^{-1}(3/4)$, where Φ is the standard normal distribution function. Define

$$s_{\mathrm{raw}} = \mathrm{med}_{i \in J}(a_i)/q, \qquad \upsilon \ge 0. \tag{60}$$

The median averages the middle two values when m is even. If $s_{\mathrm{raw}} = 0$, return $\boldsymbol{c}$. Otherwise set $\widehat{s} = \max(s_{\mathrm{raw}}, \upsilon)$. The optional numerical floor υ prevents division by a tiny positive value. The recovery bound includes the floor υ explicitly. Under arbitrary bounded noise, $\widehat{s}$ is a scale proxy with a provable upper bound, rather than an asserted consistent standard-deviation estimate.

For a cutoff $\lambda > 0$, consider the losses

$$H_\lambda(u) = \begin{cases} u^2/2, & |u| \le \lambda, \\ \lambda|u| - \lambda^2/2, & |u| > \lambda, \end{cases} \qquad T_\lambda(u) = \frac{\lambda^2}{6}\left[1 - (1 - u^2/\lambda^2)_+^3\right], \tag{61}$$

where $v_+ = \max(v, 0)$. Write $\psi_{f,\lambda}$ for the derivative of family $f \in \{\mathrm{H}, \mathrm{T}\}$. Use multiplier grids $\mathcal{C}_{\mathrm{H}} = \{0.5, 0.75, 1, 1.345, 2, 3, 5, 10\}$ and $\mathcal{C}_{\mathrm{T}} = \{2, 3, 4.685, 6, 10\}$. With $\lambda = c\widehat{s}$, select the family and multiplier minimizing

$$\widehat{V}_{f,c} = \frac{m^{-1}\sum_{i \in J} \psi_{f,c\widehat{s}}(a_i)^2}{(m^{-1}\sum_{i \in J} \psi'_{f,c\widehat{s}}(a_i))^2}, \quad m^{-1}\sum_{i \in J} \psi'_{f,c\widehat{s}}(a_i) > 0. \tag{62}$$

If no score candidate has a positive empirical derivative mean, return $\boldsymbol{c}$. Ties use the displayed family and grid order. The derivative formulas, including the negative part of the Tukey derivative, appear in (218). Appendix E derives the variance criterion in (62) for independent correctly labeled data and gives a finite-grid concentration bound. The oracle concentration bound does not apply directly to fitted nearest-component residuals. Calibration is a specified data-driven choice whose output remains subject to the recovery guard.

For a proposal iterate $\boldsymbol{v}$, write $r_{ij} = y_i - \boldsymbol{x}_i^\top \boldsymbol{v}_j$. For the Huber loss, set

$$\ell_i = -\widehat{s}^2 \log\{\tfrac{1}{2}\textstyle\sum_{j=1}^2 e^{-H_\lambda(r_{ij})/\widehat{s}^2}\}, \qquad \omega_{ij} = \frac{e^{-H_\lambda(r_{ij})/\widehat{s}^2}}{\sum_{k=1}^2 e^{-H_\lambda(r_{ik})/\widehat{s}^2}}, \tag{63}$$

and let I retain the m smallest ℓ_i. For the Tukey loss, set $b_i = \arg\min_{j \in \{1,2\}} |r_{ij}|$, and let I retain the m smallest $\min_{j \in \{1,2\}} |r_{ij}|$. Use weights

$$w_{ij}^{\mathrm{H}} = \mathbb{I}\{i \in I\}\omega_{ij}\min(1, \lambda/|r_{ij}|), \qquad w_{ij}^{\mathrm{T}} = \mathbb{I}\{i \in I,\ b_i = j\}(1 - r_{ij}^2/\lambda^2)_+^2. \tag{64}$$

The Huber ratio at zero equals one. Huber polishing minimizes a trimmed soft-mixture objective with equal prior weights. Tukey polishing uses a trimmed hard-assignment risk. A bounded Tukey loss does not define a normalizable error density by exponentiation. Proposition D.3 proves descent for the trimmed soft Huber objective and the trimmed hard-assignment Tukey objective.

Algorithm 4 SARP-P: Noise-scale-free recovery and guarded response polishing

Require: Normalized composite inputs with $\kappa = 0$, $T \geq 1$, $L_{\mathrm{H}}, L_{\mathrm{T}} \geq 1$, $\epsilon_{\mathrm{r}} > 0$, $\upsilon \geq 0$

$\boldsymbol{c} \leftarrow \boldsymbol{\theta}_T$ from Algorithm 2, $F \leftarrow F_{T-1}$ with stored weights
$a_i \leftarrow \min_{j\in\{1,2\}} |y_i - \boldsymbol{x}_i^\top \boldsymbol{c}_j|$, $J \leftarrow \arg\min_{|I|=m} \sum_{i\in I} a_i$
$s_{\mathrm{raw}} \leftarrow \mathrm{med}_{i\in J}(a_i)/q$
if $s_{\mathrm{raw}} = 0$ **then**
    return $\boldsymbol{c}$
end if
$\widehat{s} \leftarrow \max(s_{\mathrm{raw}}, \upsilon)$, $\mathcal{A} \leftarrow \{(f, c) : c \in \mathcal{C}_f,\ \sum_{i\in J} \psi'_{f,c\widehat{s}}(a_i) > 0\}$
if $\mathcal{A} = \varnothing$ **then**
    return $\boldsymbol{c}$
end if
$(f, c) \leftarrow \arg\min_{(f,c)\in\mathcal{A}} \widehat{V}_{f,c}$, $\lambda \leftarrow c\widehat{s}$, $\boldsymbol{v}^{(0)} \leftarrow \boldsymbol{c}$, $\boldsymbol{z} \leftarrow \boldsymbol{c}$
for $k = 0, \ldots, L_f - 1$ **do**
    $w_{ij} \leftarrow w_{ij}^f(\boldsymbol{v}^{(k)})$ from (64)
    if $f = \mathrm{T}$ and $\min_{j=1,2} |\{i : w_{ij} > 0\}| < d$ **then**
        return $\boldsymbol{c}$
    end if
    $\boldsymbol{v}_j^{(k+1)} \in \arg\min_{\boldsymbol{b}\in\mathbb{R}^d} \sum_{i=1}^n w_{ij}(y_i - \boldsymbol{x}_i^\top \boldsymbol{b})^2$ for $j = 1, 2$
    $\boldsymbol{z} \leftarrow \boldsymbol{v}^{(k+1)}$
    if $\|\boldsymbol{v}^{(k+1)} - \boldsymbol{v}^{(k)}\| < \epsilon_{\mathrm{r}}$ **then**
        break
    end if
end for
if $\boldsymbol{z} \in \Omega$ and $F(\boldsymbol{z}) \leq F(\boldsymbol{c}) + \widehat{s}$ **then**
    return $\boldsymbol{z}$
else
    return $\boldsymbol{c}$
end if

Equal working prior weights in the Huber proposal are a design choice. Theorem 4.2 allows $p \in [p_0, 1 - p_0]$ without requiring a correctly specified proposal likelihood.

The guard uses the last reference $\boldsymbol{\theta}_{T-1}$, not recomputed weights at $\boldsymbol{c}$. The acceptance test controls coefficient error for an arbitrary proposal, including an unfavorable local solution of the response objective. A singular weighted least-squares problem admits a minimum-norm minimizer. Numerical failures return the composite estimate and are recorded.

C PROOF OF THEOREM 4.1

The proof first establishes a deterministic recovery bound on one high-probability event for the clean covariates and labels. Tail truncation then yields the independent-error theorem in the main text. The final subsection gives a separate impossibility result for the larger auxiliary class that permits design-dependent bounded perturbations.

C.1 PROOF OUTLINE AND AN EXTENSION WITH A SMOOTHED LOSS

The proof has four steps. A single event controls clean covariates and latent labels uniformly over parameter values and replacement subsets. On the same event, LAD localizes the coefficient segment, a bounded residual transform estimates the segment direction, and a clipped search identifies both endpoints. A separate deterministic argument establishes uniform growth of every admissible normalized objective and a quadratic bound on its affine-model error. Combining local growth, model accuracy, and the objective-gap bound gives the recovery recursion. Explicit simultaneous choices of constants close the initialization and local-radius conditions. The independent-noise statement follows by treating tail exceedances as additional response outliers.

Auxiliary bounded-perturbation model. The proof first establishes a deterministic perturbation statement on one event depending only on the clean covariates and labels. For the auxiliary statement, retain the observation equation (3) but allow an arbitrary vector of errors satisfying

$$|\xi_i| \le \sigma \qquad (i \in [n]), \qquad \sigma \ge 0, \tag{65}$$

without requiring independence from the covariates or labels. The larger perturbation class is used only for the intermediate theorem and the design-dependent lower bound. The main model remains the independent-error model in Section 2. The tail-exceedance argument below reduces the independent-error model to the auxiliary bounded-perturbation statement with $\sigma = u$.

For fixed p_0 and a, use the constants

$$c_\tau = \frac{1}{32}, \qquad C_\tau = \frac{144}{\sqrt{p_0}}, \qquad a_0 = \frac{1}{16}, \qquad \mu_0 = \frac{a_0^2 p_0}{4(aC_\tau + 34)}, \qquad \mu = \frac{\mu_0}{2}. \tag{66}$$

The grid spacing δ_{g}, noise threshold s_0, and contamination threshold ϵ_0 are constructed explicitly in Appendix C.6. Their conservatism is discussed there.

Theorem C.1 (Segment initialization and normalized composite recovery)**.** *Consider the auxiliary bounded-perturbation class in* (65)*. Fix $p_0 \in (0, 1/2]$ and $a \in (0, 1]$. There are $C, \epsilon_0, s_0, \delta_{\mathrm{g}} > 0$, depending only on p_0 and a, such that the following holds for $\delta \in (0, 1/2)$. Suppose*

$$n \ge C\{d + \log(1/\delta)\}\log(en/\delta), \qquad \bar{\epsilon} \le \epsilon_0, \qquad \sigma + \kappa \le s_0\Delta. \tag{67}$$

With probability at least $1 - \delta$ over (1)*, simultaneously for every error vector in* (65) *and every response replacement in* (3)*, the initializer satisfies*

$$c_\tau\Delta \le \tau \le C_\tau\Delta, \qquad \mathcal{E}(\boldsymbol{\theta}_0) \le r/4. \tag{68}$$

Let $\mathcal{E}_t = \mathcal{E}(\boldsymbol{\theta}_t)$. Lemma C.11 supplies one ordering $\boldsymbol{\theta}^*$ of the true coefficients for which $\mathcal{E}_t = \|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|$ throughout the feasible sequence. Every feasible update of Algorithm 2 obeys

$$\mathcal{E}_{t+1} \le \frac{\ell_t}{\mu}\mathcal{E}_t^2 + \frac{6\sigma + \kappa/2 + \eta_t}{\mu}, \qquad \ell_t \le \overline{L}. \tag{69}$$

If $\eta_t \le (\mu r/2)2^{-2^{t+1}}$, then, for $0 \le t \le T$,

$$\mathcal{E}_t \le r\,2^{-2^t} + \frac{2(6\sigma + \kappa/2)}{\mu}. \tag{70}$$

Theorem C.1 combines initialization from observations with a doubly exponential optimization envelope and an additive bounded-perturbation floor. The sample size is nearly linear in d for fixed p_0, a, with potentially large constants. Relative to ILTS (Shen & Sanghavi, 2019), the recursion includes bounded inlier perturbations and arbitrary response outliers. Relative to robust phase retrieval (Duchi & Ruan, 2019), the guarantee treats two unrestricted coefficients and changing normalized objectives. The error envelope concerns recovery of the true coefficients rather than convergence to a noisy empirical minimizer or a minimax stochastic rate.

C.2 A COMMON CLEAN-SAMPLE EVENT

The uniform event below depends only on the latent clean pairs $(\boldsymbol{x}_i, z_i)$. Set $B_0 = 16$, $B_1 = 17$, and $B_2 = B_0 + 2C_\tau/c_\tau$. Define the clean empirical LAD loss and its population counterpart by

$$L_n^0(\boldsymbol{b}) = \frac{1}{n}\sum_{i=1}^n |y_i^0 - \boldsymbol{x}_i^\top \boldsymbol{b}|, \qquad L^0(\boldsymbol{b}) = c_{\mathrm{G}}\sum_{j=1}^2 p_j\|\boldsymbol{\beta}_j^* - \boldsymbol{b}\|, \qquad p_1 = p,\ p_2 = 1 - p. \tag{71}$$

Define $\boldsymbol{T}_n^0(\boldsymbol{b}, \tau)$ by (50) with y_i^0 in place of y_i, and $\boldsymbol{T}^0(\boldsymbol{b}, \tau) = \mathbb{E}\boldsymbol{T}_n^0(\boldsymbol{b}, \tau)$. Define $V_n^0(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \tau)$ by the analogous substitution in (52), and let $V^0 = \mathbb{E}V_n^0$. The subset-covariance bound uses $\gamma = 100\{\bar{\epsilon}\log(e/\bar{\epsilon}) + (d + \log(n/\delta))/n\}$, with $0\log(e/0) = 0$.

Lemma C.2 (Uniform clean-sample event)**.** *Fix $0 < \alpha < 1$ and $\delta \in (0, 1/2)$. For $n \ge C(p_0, \alpha)(d + \log(1/\delta))\log(en/\delta)$ and $n \ge 64/p_0$, the following properties hold simultaneously with probability at least $1 - \delta$.*

(E1) $1/2 \le \Lambda \le 2$ *and* $n^{-1}\sum_{i=1}^n |\boldsymbol{x}_i^\top \boldsymbol{v}| \ge (c_{\mathrm{G}}/2)\|\boldsymbol{v}\|$ *for all* $\boldsymbol{v}$.

(E2) Every $A \subseteq [n]$ *with* $|A| \le \lfloor \bar{\epsilon} n \rfloor$ *satisfies*

$$\|\boldsymbol{X}_A^\top \boldsymbol{X}_A/n\|_{\mathrm{op}} \le \gamma, \qquad n^{-1}\sum_{i\in A} |\boldsymbol{x}_i^\top \boldsymbol{v}| \le \sqrt{\bar{\epsilon}\gamma}\|\boldsymbol{v}\|, \tag{72}$$

where $\boldsymbol{X}_A$ *is the row submatrix and* $0\log(e/0) = 0$.

(E3) Uniformly over $\boldsymbol{b} \in B(\boldsymbol{m}^*, B_0\Delta)$,

$$|L_n^0(\boldsymbol{b}) - L_n^0(\boldsymbol{m}^*) - L^0(\boldsymbol{b}) + L^0(\boldsymbol{m}^*)| \le \alpha\Delta. \tag{73}$$

(E4) For each $j \in \{1,2\}$ *and unit* $\boldsymbol{w}$, *at least* $5p_0n/8$ *indices satisfy* $z_i = j$ *and* $|\boldsymbol{x}_i^\top \boldsymbol{w}| \ge 1/8$. *For every such* $j, \boldsymbol{w}$ *and every* $\boldsymbol{v}_1, \boldsymbol{v}_2$ *of norm at most* 2, *at least* $p_0n/2$ *indices satisfy*

$$z_i = j, \quad |\boldsymbol{x}_i^\top \boldsymbol{u}^*| \ge a_0, \quad |\boldsymbol{x}_i^\top \boldsymbol{w}| \ge a_0, \quad |\boldsymbol{x}_i^\top \boldsymbol{v}_1| \le 16, \quad |\boldsymbol{x}_i^\top \boldsymbol{v}_2| \le 16. \tag{74}$$

(E5) Uniformly over $\boldsymbol{b} \in B(\boldsymbol{m}^*, B_0\Delta)$ *and* $\tau \in [c_\tau\Delta, C_\tau\Delta]$, $\|\boldsymbol{T}_n^0(\boldsymbol{b},\tau) - \boldsymbol{T}^0(\boldsymbol{b},\tau)\|_{\mathrm{op}} \le \alpha$.

(E6) Uniformly over $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in B(\boldsymbol{m}^*, B_2\Delta)$ *and the same scale interval,* $|V_n^0 - V^0| \le \alpha\Delta$.

Conditional on the event of Lemma C.2, the subsequent inequalities are deterministic and hold simultaneously for all admissible bounded perturbations, response-outlier sets, fitted anchors, and iterate sequences. The simultaneous quantifiers permit reuse of the observations.

Proof. Define eight bad events $\mathcal{B}_1, \dots, \mathcal{B}_8$ below, corresponding respectively to covariance control, the maximum row norm, the absolute-projection net, subset matrices, the LAD net, slab counts, the spectral net, and the clipped-score net. Each probability is bounded explicitly by $\delta/12$. Every net is deterministic and may depend on n, δ, p_0, α, and the fixed population parameters, but not on the observations. The constants $c_1, c_3, c_5, c_7 > 0$ below are universal concentration constants from Lemma A.1.

Design bounds. The sphere $1/4$-net argument in Lemma A.2, using Lemma A.1 at deviation $1/4$, gives

$$\begin{aligned} \mathcal{B}_1 &= \{\|\boldsymbol{X}^\top \boldsymbol{X}/n - \boldsymbol{I}_d\|_{\mathrm{op}} > 1/2\}, \\ \mathbb{P}(\mathcal{B}_1) &\le 2\cdot 9^d e^{-c_1 n} \le \delta/12. \end{aligned} \tag{75}$$

The covariance event bounds the eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}/n$ between $1/2$ and $3/2$, and hence $1/2 \le \Lambda \le 2$. For a fixed unit $\boldsymbol{w}$, $|\boldsymbol{x}_i^\top \boldsymbol{w}|$ has mean c_{G} and a universally bounded sub-Gaussian norm. Let $\mathcal{N}_1$ be a sphere net of radius $c_{\mathrm{G}}/(16\sqrt{2})$, with at most $N_1 = (1 + 32\sqrt{2}/c_{\mathrm{G}})^d$ points. Lemma A.1 and the union bound give

$$\begin{aligned} \mathcal{B}_3 &= \left\{ \max_{\boldsymbol{w}\in\mathcal{N}_1} \left| \frac{1}{n}\sum_{i=1}^n |\boldsymbol{x}_i^\top \boldsymbol{w}| - c_{\mathrm{G}} \right| > c_{\mathrm{G}}/4 \right\}, \\ \mathbb{P}(\mathcal{B}_3) &\le 2N_1 e^{-c_3 n} \le \delta/12. \end{aligned} \tag{76}$$

On the covariance event,

$$\frac{1}{n}\sum_{i=1}^n \big||\boldsymbol{x}_i^\top \boldsymbol{w}| - |\boldsymbol{x}_i^\top \boldsymbol{w}'|\big| \le \left(\frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i^\top(\boldsymbol{w} - \boldsymbol{w}'))^2\right)^{1/2} \le \sqrt{\Lambda}\|\boldsymbol{w} - \boldsymbol{w}'\|. \tag{77}$$

Extension from the net leaves at least $c_{\mathrm{G}} - c_{\mathrm{G}}/4 - c_{\mathrm{G}}/16 > c_{\mathrm{G}}/2$. Homogeneity proves (E1).

For (E2), set $k = \lfloor \bar{\epsilon} n \rfloor$. The case $k = 0$ has no nonempty subsets and is handled separately below. For $k \ge 1$, every smaller subset can be enlarged to size k, and adding rows increases its covariance in the positive semidefinite order. There are at most $\binom{n}{k} \le (en/k)^k$ size-k subsets, since

$$\binom{n}{k} \le \frac{n^k}{k!}, \qquad \log(k!) = \sum_{j=1}^k \log j \ge \int_1^k \log t \,\mathrm{d}t \ge k\log k - k. \tag{78}$$

For $k \geq 1$, put $t_4 = k\log(en/k) + \log(12/\delta)$ and apply (33) to each submatrix. Define

$$\mathcal{B}_4 = \bigcup_{|A|=k} \left\{ \|\boldsymbol{X}_A\|_{\mathrm{op}} > 2\big(\sqrt{k} + \sqrt{2(d\log 5 + t_4)}\big) \right\}, \qquad \mathbb{P}(\mathcal{B}_4) \leq \binom{n}{k} e^{-t_4} \leq \exp\{k\log(en/k) - t_4\} = \delta/12. \tag{79}$$

For $k = 0$, define $\mathcal{B}_4 = \varnothing$, hence (79) holds with zero failure probability. Outside $\mathcal{B}_4$, squaring the matrix threshold gives

$$\max_{|A|=k} \frac{\|\boldsymbol{X}_A\|_{\mathrm{op}}^2}{n} \leq \frac{8k + 16d\log 5 + 16k\log(en/k) + 16\log(12/\delta)}{n}. \tag{80}$$

Here $(u+v)^2 \leq 2u^2 + 2v^2$ was applied to the matrix-tail threshold. Since $\log(en/k) \geq 1$, the two k terms sum to at most $24k\log(en/k)$. The map $x \mapsto x\log(e/x)$ is increasing on $(0, 1]$, with derivative $\log(1/x) \geq 0$, hence

$$\frac{k}{n}\log(en/k) \leq \bar{\epsilon}\log(e/\bar{\epsilon}). \tag{81}$$

Also $16\log 5 < 100$ and $16\log(12/\delta) \leq 100\log(n/\delta)$ for $n \geq 128$. The covariance bound is therefore at most γ. For any allowed subset and vector,

$$\frac{1}{n}\sum_{i\in A} |\boldsymbol{x}_i^\top \boldsymbol{v}| \leq \sqrt{\frac{|A|}{n}}\sqrt{\boldsymbol{v}^\top \boldsymbol{X}_A^\top \boldsymbol{X}_A \boldsymbol{v}/n} \leq \sqrt{\bar{\epsilon}\gamma}\|\boldsymbol{v}\|. \tag{82}$$

LAD contrast. For fixed $\boldsymbol{b} \in B(\boldsymbol{m}^*, B_0\Delta)$, the single-observation contrast relative to $\boldsymbol{m}^*$ is bounded in absolute value by $|\boldsymbol{x}_i^\top(\boldsymbol{b} - \boldsymbol{m}^*)|$. Its sub-Gaussian norm is at most $CB_0\Delta$. Choose an anchor net of radius

$$h_b = \frac{\alpha\Delta}{8(\sqrt{2} + c_{\mathrm{G}})}, \qquad N_b \leq (1 + 2B_0\Delta/h_b)^d. \tag{83}$$

Write $\mathcal{N}_b$ for the anchor net and define $D_n(\boldsymbol{b}) = L_n^0(\boldsymbol{b}) - L_n^0(\boldsymbol{m}^*) - L^0(\boldsymbol{b}) + L^0(\boldsymbol{m}^*)$. Lemma A.1 at deviation $\alpha\Delta/2$ and a union bound give

$$\mathcal{B}_5 = \{\max_{\boldsymbol{b}\in\mathcal{N}_b} |D_n(\boldsymbol{b})| > \alpha\Delta/2\}, \qquad \mathbb{P}(\mathcal{B}_5) \leq 2N_b e^{-c_5 n\alpha^2/B_0^2} \leq \delta/12. \tag{84}$$

Changing an anchor by $\boldsymbol{v}$ changes the empirical contrast by at most $\sqrt{\Lambda}\|\boldsymbol{v}\|$ and its population expectation by at most $c_{\mathrm{G}}\|\boldsymbol{v}\|$. Extension from the net costs at most $(\sqrt{2} + c_{\mathrm{G}})h_b = \alpha\Delta/8$. The net extension proves (E3).

Uniform slab counts. For the first count at label j, the population probability is at least $p_j(1 - c_{\mathrm{G}}/8)$. A uniform empirical deviation of $p_0/4$ leaves at least

$$p_0(1 - c_{\mathrm{G}}/8) - p_0/4 = p_0(3/4 - c_{\mathrm{G}}/8) > 5p_0/8. \tag{85}$$

For the second count, conditional on label j, the union bound and the Markov inequality give

$$\mathbb{P}\{|\boldsymbol{x}^\top \boldsymbol{u}^*| \geq a_0,\ |\boldsymbol{x}^\top \boldsymbol{w}| \geq a_0,\ |\boldsymbol{x}^\top \boldsymbol{v}_1| \leq 16,\ |\boldsymbol{x}^\top \boldsymbol{v}_2| \leq 16\} \tag{86}$$

$$\geq 1 - 2c_{\mathrm{G}}a_0 - (\|\boldsymbol{v}_1\|^2 + \|\boldsymbol{v}_2\|^2)/16^2 \geq 1 - 2c_{\mathrm{G}}a_0 - 1/32 > 3/4. \tag{87}$$

Multiplication by p_j and subtraction of $p_0/4$ leave at least $p_0/2$. Let $\mathcal{A}_1, \ldots, \mathcal{A}_4$ be the four classes obtained by combining each count type with each label. Each class uses at most eight affine halfspaces and a fixed label indicator. Put $W_i = (\boldsymbol{x}_i, z_i)$ and $P_n(A) = n^{-1}\sum_{i=1}^n \mathbb{I}\{W_i \in A\}$. For $n(p_0/4)^2 \geq 2$, Lemma A.3, (39), and a union bound give

$$\mathcal{B}_6 = \bigcup_{h=1}^4 \left\{ \sup_{A\in\mathcal{A}_h} |P_n(A) - \mathbb{P}(W \in A)| > p_0/4 \right\}, \qquad \mathbb{P}(\mathcal{B}_6) \leq 32\left(\frac{2en}{d+1}\right)^{8(d+1)} e^{-np_0^2/512} \leq \delta/12. \tag{88}$$

The uniform deviation bound proves (E4).

Uniform residual spectrum. For fixed $\boldsymbol{b}, \tau, \boldsymbol{w}$ with $\|\boldsymbol{w}\| = 1$, the summand

$$g((y_i^0 - \boldsymbol{x}_i^\top \boldsymbol{b})/\tau)\{(\boldsymbol{x}_i^\top \boldsymbol{w})^2 - 1\} \tag{89}$$

is bounded in magnitude by $(\boldsymbol{x}_i^\top \boldsymbol{w})^2 + 1$. Its centered sub-exponential norm is universally bounded. Define

$$R_x = \sqrt{d} + \sqrt{2\log(12n/\delta)}. \tag{90}$$

The Gaussian norm tail in Lemma A.2 and a union bound give

$$\mathcal{B}_2 = \{\max_{i\in[n]} \|\boldsymbol{x}_i\| > R_x\}, \qquad \mathbb{P}(\mathcal{B}_2) \le n\exp\{-\log(12n/\delta)\} = \delta/12. \tag{91}$$

Take a sphere $1/4$-net and anchor and scale nets of radii

$$h_5 = \frac{\alpha c_\tau \Delta}{32(3R_x + 2)}, \qquad h_\tau = \frac{\alpha c_\tau \Delta}{320}. \tag{92}$$

Their joint cardinality is at most

$$N_5 = 9^d(1 + 2B_0\Delta/h_5)^d(2 + C_\tau\Delta/h_\tau). \tag{93}$$

Let $\mathcal{N}_5$ denote the product net of triples $(\boldsymbol{b}, \tau, \boldsymbol{w})$, and set $Q_n(\boldsymbol{b}, \tau, \boldsymbol{w}) = \boldsymbol{w}^\top\{\boldsymbol{T}_n^0(\boldsymbol{b}, \tau) - \boldsymbol{T}^0(\boldsymbol{b}, \tau)\}\boldsymbol{w}$. The sub-exponential Bernstein bound in Lemma A.1, with $0 < \alpha < 1$, yields

$$\begin{aligned} \mathcal{B}_7 &= \{\max_{(\boldsymbol{b},\tau,\boldsymbol{w})\in\mathcal{N}_5} |Q_n(\boldsymbol{b}, \tau, \boldsymbol{w})| > \alpha/4\}, \\ \mathbb{P}(\mathcal{B}_7) &\le 2N_5 e^{-c_7 n\alpha^2} \le \delta/12. \end{aligned} \tag{94}$$

We extend the spectral net bounds using the maximum-row and covariance events. The empirical anchor Lipschitz constant of each quadratic form is at most

$$\frac{R_x}{c_\tau\Delta}\frac{1}{n}\sum_{i=1}^n |(\boldsymbol{x}_i^\top \boldsymbol{w})^2 - 1| \le \frac{R_x(\Lambda + 1)}{c_\tau\Delta}. \tag{95}$$

Its population constant is at most $\sqrt{2}/(c_\tau\Delta)$ by Lemma A.4. The empirical-minus-population quadratic form therefore has anchor Lipschitz constant at most $(3R_x + \sqrt{2})/(c_\tau\Delta)$. For the scale, the derivative identity

$$\left|\frac{\mathrm{d}}{\mathrm{d}\tau} g(r/\tau)\right| = \frac{|(r/\tau)g'(r/\tau)|}{\tau} \le \frac{2}{c_\tau\Delta} \tag{96}$$

gives a centered scale constant at most

$$\frac{2(\Lambda + 1) + 2\mathbb{E}|Z^2 - 1|}{c_\tau\Delta} \le \frac{6 + 2\sqrt{2}}{c_\tau\Delta} \le \frac{10}{c_\tau\Delta}. \tag{97}$$

Anchor and scale extension therefore cost at most $\alpha/32$ each. The symmetric $1/4$-net extension costs a factor two, hence the operator error is at most $2(\alpha/4 + \alpha/32 + \alpha/32) < \alpha$. The net extension proves (E5) without conditioning independent summands on a sample-dependent row event.

Clipped pair score. Every summand lies in $[0, C_\tau\Delta]$. The inequality $|\min_{j\in\{1,2\}} a_j - \min_{j\in\{1,2\}} b_j| \le \max_j |a_j - b_j|$ shows that changing the parameters changes a summand by at most

$$|\boldsymbol{x}_i^\top(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1')| + |\boldsymbol{x}_i^\top(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_2')| + |\tau - \tau'|. \tag{98}$$

Use nets of radius $h_6 = \alpha\Delta/[16(\sqrt{2} + c_{\mathrm{G}} + 1)]$ for the two coefficient balls and the scale interval. Their joint size is at most

$$N_6 = (1 + 2B_2\Delta/h_6)^{2d}(2 + C_\tau\Delta/h_6). \tag{99}$$

Let $\mathcal{N}_6$ denote the product net of triples $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \tau)$. The Hoeffding bound in Lemma A.1, with interval length $C_\tau\Delta$ and deviation $\alpha\Delta/2$, gives

$$\begin{aligned} \mathcal{B}_8 &= \{\max_{(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\tau)\in\mathcal{N}_6} |V_n^0 - V^0| > \alpha\Delta/2\}, \\ \mathbb{P}(\mathcal{B}_8) &\le 2N_6\exp\{-n\alpha^2/(2C_\tau^2)\} \le \delta/12. \end{aligned} \tag{100}$$

The empirical and population extension costs together are at most

$$\{2(\sqrt{2}+c_{\mathrm{G}})+2\}h_6 = \alpha\Delta/8. \tag{101}$$

The clipped-score net bound proves (E6).

Finally choose the sample constant large enough that $n \geq d$ and all stated exponential failure bounds are at most $\delta/12$. The required dependence follows from the covering counts: The logarithms of N_1, N_b, N_6 are at most $C(p_0,\alpha)d$, and

$$\log N_5 \leq C(p_0,\alpha)d\log(en/\delta), \qquad 8(d+1)\log(2en/(d+1)) \leq Cd\log(en/\delta). \tag{102}$$

For the spectral term in (102), use $R_x \leq \sqrt{n}+\sqrt{2\log(12n/\delta)}$ when $d \leq n$ and substitute (92) into (93). The covering-count bounds imply that $n \geq C(p_0,\alpha)(d+\log(1/\delta))\log(en/\delta)$ satisfies every required exponential inequality after increasing its constant. The bounds (75), (91), (76), (79), (84), (88), (94), and (100) give

$$\mathbb{P}\left(\bigcup_{j=1}^{8}\mathcal{B}_j\right) \leq \sum_{j=1}^{8}\mathbb{P}(\mathcal{B}_j) \leq 8\frac{\delta}{12} = \frac{2\delta}{3} \leq \delta. \tag{103}$$

Outside $\bigcup_{j=1}^{8}\mathcal{B}_j$, all net extensions above apply simultaneously and establish (E1)–(E6). □

C.3 Localization of the segment and its endpoints

Segment initialization is analyzed in three steps. LAD localizes an anchor and estimates the separation scale, the bounded residual spectrum estimates the segment direction, and the clipped grid score identifies both endpoints. Proposition C.3 collects the resulting errors.

Proposition C.3 (Segment localization and endpoint initialization)**.** *On the probability event in Lemma C.2, put*

$$s = \sigma/\Delta, \qquad \gamma = 100\left\{\bar{\epsilon}\log(e/\bar{\epsilon}) + \frac{d+\log(n/\delta)}{n}\right\}, \tag{104}$$

$$\zeta = \frac{4B_1}{c_{\mathrm{G}}}\sqrt{\bar{\epsilon}\gamma} + \sqrt{\frac{4B_1}{c_{\mathrm{G}}}(\alpha+s)}, \qquad B_1 = 17, \tag{105}$$

$$E_* = \alpha+\gamma+\bar{\epsilon}+\frac{2\zeta+3s}{c_\tau}, \qquad H_*^{\mathrm{err}} = \zeta+\frac{4E_*}{\kappa_0}+\frac{\delta_{\mathrm{g}}C_\tau}{2}+2\alpha+2s+2\bar{\epsilon}C_\tau. \tag{106}$$

Here $\alpha > 0$ is the uniform deviation parameter, $\kappa_0 > 0$ is the spectral gap in (124)*, and* $\mathrm{dist}(\boldsymbol{b}, S) = \inf_{\boldsymbol{v}\in S}\|\boldsymbol{b}-\boldsymbol{v}\|$*. Under the numerical conditions verified in Appendix C.6,*

$$\|\boldsymbol{b}-\boldsymbol{m}^*\| \leq 16\Delta, \qquad \mathrm{dist}(\boldsymbol{b}, S) \leq \zeta\Delta, \tag{107}$$

$$\min_{\omega\in\{-1,1\}}\|\widehat{\boldsymbol{u}}-\omega\boldsymbol{u}^*\| \leq 4E_*/\kappa_0, \qquad \mathcal{E}(\boldsymbol{\theta}_0) \leq \frac{16\sqrt{2}}{p_0}H_*^{\mathrm{err}}\Delta. \tag{108}$$

Proposition C.3 quantifies the errors accumulated during the three steps of segment initialization. Segment localization avoids a bound involving $\|\boldsymbol{m}^*\|$. The rank-one residual identity and clipped endpoint comparison convert segment localization into estimates of both coefficients. The argument tolerates sample-dependent LAD estimates and replacement locations.

We work throughout on the event of Lemma C.2. Write $L_n(\boldsymbol{b}) = n^{-1}\sum_{i=1}^{n}|y_i - \boldsymbol{x}_i^\top\boldsymbol{b}|$. For every $\boldsymbol{b}, \boldsymbol{b}'$, the reverse triangle inequality gives

$$|\{L_n(\boldsymbol{b})-L_n(\boldsymbol{b}')\}-\{L_n^0(\boldsymbol{b})-L_n^0(\boldsymbol{b}')\}| \tag{109}$$

$$\leq \frac{2}{n}\sum_{i\in O}|\boldsymbol{x}_i^\top(\boldsymbol{b}-\boldsymbol{b}')| + 2\sigma \leq 2\sqrt{\bar{\epsilon}\gamma}\|\boldsymbol{b}-\boldsymbol{b}'\| + 2\sigma. \tag{110}$$

For replaced rows, both contrasts are Lipschitz with respect to the fitted response, regardless of v_i. For other rows, each of the two losses changes by at most σ. The two rowwise bounds prove (110) independently of the response-outlier amplitudes.

Lemma C.4 (LAD localization and a separation-scale quantile)**.** *If $s = \sigma/\Delta \le 1/32$, $\bar{\epsilon} \le p_0/16$, and $2\sqrt{\bar{\epsilon}\gamma} \le c_{\mathrm{G}}/4$, every LAD solution satisfies*

$$\|\boldsymbol{b} - \boldsymbol{m}^*\| \le 16\Delta, \qquad \mathrm{dist}(\boldsymbol{b}, S) \le \zeta\Delta, \qquad c_\tau\Delta \le \tau \le C_\tau\Delta, \tag{111}$$

with ζ defined in (106)*.*

LAD need not identify either endpoint. Lemma C.4 locates the LAD output near the coefficient segment and obtains a nondegenerate scale even when a majority component is nearly interpolated.

Proof. Put $H = \|\boldsymbol{b} - \boldsymbol{m}^*\|$. For every candidate $\boldsymbol{b}$,

$$L_n^0(\boldsymbol{b}) - L_n^0(\boldsymbol{m}^*) \ge n^{-1}\sum_{i=1}^n |\boldsymbol{x}_i^\top(\boldsymbol{b} - \boldsymbol{m}^*)| - 2n^{-1}\sum_{i=1}^n |y_i^0 - \boldsymbol{x}_i^\top\boldsymbol{m}^*| \tag{112}$$

$$\ge (c_{\mathrm{G}}/2)H - \sqrt{\Lambda}\Delta. \tag{113}$$

Here $|y_i^0 - \boldsymbol{x}_i^\top\boldsymbol{m}^*| = (\Delta/2)|\boldsymbol{x}_i^\top\boldsymbol{u}^*|$ exactly. Combining (113), (110), and LAD optimality gives

$$(c_{\mathrm{G}}/2 - 2\sqrt{\bar{\epsilon}\gamma})H \le \sqrt{2}\Delta + 2\sigma, \qquad H \le 4(\sqrt{2} + 1/16)\Delta/c_{\mathrm{G}} < 16\Delta. \tag{114}$$

The same coercive inequality guarantees existence of a minimizer. Let $\bar{\boldsymbol{b}} = \mathrm{Proj}_S(\boldsymbol{b})$ and $t = \|\boldsymbol{b} - \bar{\boldsymbol{b}}\|$. Lemma A.5 gives

$$\|\boldsymbol{b} - \boldsymbol{\beta}_j^*\| - \|\bar{\boldsymbol{b}} - \boldsymbol{\beta}_j^*\| = \frac{\|\boldsymbol{b} - \boldsymbol{\beta}_j^*\|^2 - \|\bar{\boldsymbol{b}} - \boldsymbol{\beta}_j^*\|^2}{\|\boldsymbol{b} - \boldsymbol{\beta}_j^*\| + \|\bar{\boldsymbol{b}} - \boldsymbol{\beta}_j^*\|} \ge \frac{t^2}{2B_1\Delta}. \tag{115}$$

If a denominator is zero, $t = 0$ and the claimed projection inequality is immediate. Since both anchors lie in the ball covered by (E3), optimality and (110) imply

$$\frac{c_{\mathrm{G}}t^2}{2B_1\Delta} \le L^0(\boldsymbol{b}) - L^0(\bar{\boldsymbol{b}}) \le 2\alpha\Delta + 2\sigma + 2\sqrt{\bar{\epsilon}\gamma}\,t. \tag{116}$$

For $x^2 \le bx + c$ with $x, b, c \ge 0$, $x \le b + \sqrt{c}$. Applying the scalar inequality to $x = t/\Delta$ proves $t \le \zeta\Delta$.

For the quantile bounds, let $k = 1 + \lceil (1 - p_0/2)(n-1) \rceil$ be the rank defining τ. The integer identity $\lceil N - x \rceil = N - \lfloor x \rfloor$ gives

$$n - k = \lfloor p_0(n-1)/2 \rfloor. \tag{117}$$

For the upper bound, set $C_p = 8/\sqrt{p_0}$ and $R = C_p(H + \Delta)$. The squared residual satisfies

$$|y_i^0 - \boldsymbol{x}_i^\top\boldsymbol{b}|^2 \le 2|\boldsymbol{x}_i^\top(\boldsymbol{b} - \boldsymbol{m}^*)|^2 + (\Delta^2/2)|\boldsymbol{x}_i^\top\boldsymbol{u}^*|^2. \tag{118}$$

Consequently

$$\frac{1}{n}\#\{i : |y_i^0 - \boldsymbol{x}_i^\top\boldsymbol{b}| > R\} \le \frac{4H^2 + \Delta^2}{C_p^2(H + \Delta)^2} \le \frac{p_0}{16}. \tag{119}$$

An observed residual can exceed $R + \sigma$ only on these rows or on O. There are at most $p_0 n/8$ such rows. Since $n \ge 64/p_0$,

$$n - k \ge \frac{p_0(n-1)}{2} - 1 \ge \frac{p_0 n}{8}, \tag{120}$$

where the second inequality follows from $3p_0 n/8 \ge p_0/2 + 1$. Thus at least k residuals are at most $R + \sigma$, and

$$\tau \le C_p(H + \Delta) + \sigma \le (136/\sqrt{p_0} + 1/32)\Delta \le C_\tau\Delta. \tag{121}$$

For the lower bound, at least one endpoint satisfies $\|\boldsymbol{\beta}_j^* - \boldsymbol{b}\| \ge \Delta/2$. Apply the first count in (E4) in its direction. At least $5p_0 n/8$ clean-design rows of label j have noiseless residual at least $\Delta/16$. Removing O leaves at least $9p_0 n/16$ rows with observed residual at least $\Delta/16 - \sigma \ge \Delta/32$. Moreover,

$$n - k + 1 \le \frac{p_0(n-1)}{2} + 1 \le \frac{p_0 n}{2} + 1 \le \frac{9p_0 n}{16}, \tag{122}$$

since $p_0 n/16 \ge 4$. Thus at least $n - k + 1$ residuals are at least $\Delta/32$, forcing the rank-k residual to be at least $c_\tau\Delta$. □

Lemma C.5 (Rank-one residual identity and perturbation)**.** *For any $\boldsymbol{b} \in \mathbb{R}^d$ and $\tau > 0$, writing $\boldsymbol{a}_j = \boldsymbol{\beta}_j^* - \boldsymbol{b}$,*

$$\boldsymbol{T}^0(\boldsymbol{b}, \tau) = \sum_{j=1}^{2} p_j \frac{\boldsymbol{a}_j \boldsymbol{a}_j^\top / \tau^2}{(1 + \|\boldsymbol{a}_j\|^2/\tau^2)^{3/2}}. \tag{123}$$

If $\bar{\boldsymbol{b}} \in S$ and $\tau \in [c_\tau \Delta, C_\tau \Delta]$, then $\boldsymbol{T}^0(\bar{\boldsymbol{b}}, \tau) = \lambda \boldsymbol{u}^* \boldsymbol{u}^{*\top}$ with

$$\lambda \geq \kappa_0 := p_0 \min_{v \in [1/(2C_\tau), 1/c_\tau]} \frac{v^2}{(1+v^2)^{3/2}} > 0. \tag{124}$$

For the fitted $\boldsymbol{b}$ and $\bar{\boldsymbol{b}} = \mathrm{Proj}_S(\boldsymbol{b})$,

$$\|\widehat{\boldsymbol{T}} - \boldsymbol{T}^0(\bar{\boldsymbol{b}}, \tau)\|_{\mathrm{op}} \leq E_*. \tag{125}$$

If $E_* \leq \kappa_0/4$, the direction error is at most $4E_*/\kappa_0$ after a sign choice.

Identity (123) produces a rank-one population matrix from any anchor on the coefficient segment. Bounded preprocessing controls the direction error under arbitrary response outliers and bounded inlier perturbations.

Proof. For $\boldsymbol{a} \neq \boldsymbol{0}$, put $\boldsymbol{u} = \boldsymbol{a}/\|\boldsymbol{a}\|$, $Z = \boldsymbol{u}^\top \boldsymbol{x}$, and $\boldsymbol{z}_\perp = (\boldsymbol{I}_d - \boldsymbol{u}\boldsymbol{u}^\top)\boldsymbol{x}$. In an orthogonal coordinate system with first axis $\boldsymbol{u}$, the Gaussian density factors into the product of its one-dimensional densities. Hence Z and $\boldsymbol{z}_\perp$ are independent, with

$$\mathbb{E}\boldsymbol{z}_\perp = \boldsymbol{0}, \qquad \mathbb{E}\boldsymbol{z}_\perp \boldsymbol{z}_\perp^\top = \boldsymbol{I}_d - \boldsymbol{u}\boldsymbol{u}^\top. \tag{126}$$

Write $v = \|\boldsymbol{a}\|/\tau$. Independence gives the two vanishing cross blocks

$$\mathbb{E}[g(vZ) Z \boldsymbol{z}_\perp^\top] = \mathbb{E}[g(vZ)Z]\mathbb{E}\boldsymbol{z}_\perp^\top = \boldsymbol{0}^\top, \qquad \mathbb{E}[g(vZ)\boldsymbol{z}_\perp Z] = \boldsymbol{0}. \tag{127}$$

Expanding $\boldsymbol{x} = Z\boldsymbol{u} + \boldsymbol{z}_\perp$ therefore yields

$$\begin{aligned} &\mathbb{E}[g(vZ)(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{I}_d)] && (128)\\ &\quad = \mathbb{E}[g(vZ)Z^2]\boldsymbol{u}\boldsymbol{u}^\top + \mathbb{E}g(vZ)(\boldsymbol{I}_d - \boldsymbol{u}\boldsymbol{u}^\top) - \mathbb{E}g(vZ)\boldsymbol{I}_d && (129)\\ &\quad = \mathbb{E}[g(vZ)(Z^2-1)]\boldsymbol{u}\boldsymbol{u}^\top. && (130)\end{aligned}$$

Lemma A.4 gives

$$\begin{aligned} \mathbb{E}[g(vZ)(Z^2-1)] &= \mathbb{E}(Z^2-1) - \mathbb{E}[e^{-v^2Z^2/2}(Z^2-1)] && (131)\\ &= (1+v^2)^{-1/2} - (1+v^2)^{-3/2} = \frac{v^2}{(1+v^2)^{3/2}}. && (132)\end{aligned}$$

Substituting (132) into (130) and using $v^2\boldsymbol{u}\boldsymbol{u}^\top = \boldsymbol{a}\boldsymbol{a}^\top/\tau^2$ gives the single-component matrix in (123). For $\boldsymbol{a} = \boldsymbol{0}$, $g(0) = 0$ and the same matrix formula is zero. The label probabilities are p_j and the label is independent of the Gaussian design. Summing the two conditional expectations proves (123).

For a segment anchor all $\boldsymbol{a}_j$ are parallel to $\boldsymbol{u}^*$. At least one has norm at least $\Delta/2$ and both have norm at most Δ. The component with the larger offset has weight at least p_0, giving (124).

For replacements, the difference of two g values belongs to $[-1, 1]$. Consequently (E2) gives

$$\left\| \frac{1}{n} \sum_{i \in O} (g_i - g_i^0)(\boldsymbol{x}_i \boldsymbol{x}_i^\top - \boldsymbol{I}_d) \right\|_{\mathrm{op}} \leq \gamma + \bar{\epsilon}. \tag{133}$$

On other rows, the Lipschitz property gives $|g_i - g_i^0| \leq \sigma/\tau$. Their contribution is at most $\sigma(\Lambda + 1)/\tau \leq 3s/c_\tau$. For the population anchor perturbation, take a unit $\boldsymbol{w}$ and apply Lemma A.4 to obtain

$$\begin{aligned} &|\boldsymbol{w}^\top \{\boldsymbol{T}^0(\boldsymbol{b}, \tau) - \boldsymbol{T}^0(\bar{\boldsymbol{b}}, \tau)\}\boldsymbol{w}| \\ &\quad \leq \frac{1}{\tau}\mathbb{E}\big[|\boldsymbol{x}^\top(\boldsymbol{b} - \bar{\boldsymbol{b}})||(\boldsymbol{x}^\top \boldsymbol{w})^2 - 1|\big] \\ &\quad \leq \frac{1}{\tau}\sqrt{\mathbb{E}|\boldsymbol{x}^\top(\boldsymbol{b} - \bar{\boldsymbol{b}})|^2}\sqrt{\mathbb{E}\{(\boldsymbol{x}^\top \boldsymbol{w})^2 - 1\}^2} \\ &\quad = \frac{\sqrt{2}\|\boldsymbol{b} - \bar{\boldsymbol{b}}\|}{\tau} \leq \frac{2\zeta}{c_\tau}. \end{aligned} \tag{134}$$

The replacement estimate, the good-row estimate, (134), and property (E5) of Lemma C.2 give

$$\begin{aligned}\|\widehat{\boldsymbol{T}} - \boldsymbol{T}^0(\bar{\boldsymbol{b}}, \tau)\|_{\mathrm{op}} &\le \|\widehat{\boldsymbol{T}} - \boldsymbol{T}_n^0(\boldsymbol{b}, \tau)\|_{\mathrm{op}} + \|\boldsymbol{T}_n^0(\boldsymbol{b}, \tau) - \boldsymbol{T}^0(\boldsymbol{b}, \tau)\|_{\mathrm{op}} \\ &\quad + \|\boldsymbol{T}^0(\boldsymbol{b}, \tau) - \boldsymbol{T}^0(\bar{\boldsymbol{b}}, \tau)\|_{\mathrm{op}} \\ &\le \gamma + \bar{\epsilon} + \frac{3s}{c_\tau} + \alpha + \frac{2\zeta}{c_\tau} = E_*.\end{aligned} \tag{135}$$

Lemma A.5 yields $\sqrt{2}E_*/(\kappa_0 - E_*) \le 4E_*/\kappa_0$. □

Lemma C.6 (Clipped score identifies both endpoints)**.** *For any candidate pair and any $\tau > 0$, put $D_j = \min_{k=1,2} \|\boldsymbol{\beta}_k - \boldsymbol{\beta}_j^*\|$. Then*

$$V^0(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \tau) \ge \frac{p_0}{16} \max_{j=1,2} \min(D_j, \tau). \tag{136}$$

If $\zeta \le 1$, $E_* \le \kappa_0/4$, and $16H_*^{\mathrm{err}}/p_0 < \min(c_\tau, 1/2)$, the grid output satisfies $\mathcal{E}(\boldsymbol{\theta}_0) \le 16\sqrt{2}H_*^{\mathrm{err}}\Delta/p_0$, as asserted in Proposition C.3.

The score cannot be small if a candidate pair misses one true component. Clipping preserves the endpoint-identification argument under unbounded response outliers.

Proof. Fix j with $D_j > 0$ and set $b = \min(D_j, \tau)$. The two vectors $\boldsymbol{\beta}_j^* - \boldsymbol{\beta}_k$ have norm at least D_j. Gaussian anti-concentration gives

$$\mathbb{P}\{\min_k |\boldsymbol{x}^\top(\boldsymbol{\beta}_j^* - \boldsymbol{\beta}_k)| \ge b/8\} \ge 1 - 2c_{\mathrm{G}}/8 > 1/2. \tag{137}$$

The clipped residual is at least $b/8$ on this event, hence its conditional expectation is at least $b/16$. Multiplication by $p_j \ge p_0$ proves (136). If $D_j = 0$ the bound is trivial.

Choose $\omega \in \{-1, 1\}$ such that $\|\widehat{\boldsymbol{u}} - \omega\boldsymbol{u}^*\| \le 4E_*/\kappa_0$. Write $\boldsymbol{\beta}_j^* = \bar{\boldsymbol{b}} + t_j^*\omega\boldsymbol{u}^*$ with $|t_j^*| \le \Delta$. Since $\tau \ge c_\tau\Delta$, the offsets t_j^* belong to the grid interval. There exist grid offsets t_j' with $|t_j' - t_j^*| \le \delta_{\mathrm{g}}\tau/2$. The corresponding pair $\boldsymbol{\beta}_j' = \boldsymbol{b} + t_j'\widehat{\boldsymbol{u}}$ satisfies

$$\|\boldsymbol{\beta}_j' - \boldsymbol{\beta}_j^*\| \le \left(\zeta + \frac{4E_*}{\kappa_0} + \frac{\delta_{\mathrm{g}}C_\tau}{2}\right)\Delta =: H_g\Delta. \tag{138}$$

Both the constructed grid pair and the selected pair lie in the (E6) ball by the definition of B_2. The observed scores compare the population scores of the constructed and selected pairs. For every fixed pair,

$$|V - V_n^0| \le \sigma + \bar{\epsilon}\tau, \qquad V^0(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \tau) \le \sum_{j=1}^{2} p_j \mathbb{E}|\boldsymbol{x}^\top(\boldsymbol{\beta}_j' - \boldsymbol{\beta}_j^*)| \le H_g\Delta. \tag{139}$$

Using (E6) twice and the minimization in (52) gives

$$\begin{aligned}V^0(\boldsymbol{\beta}_{1,0}, \boldsymbol{\beta}_{2,0}, \tau) &\le V(\boldsymbol{\beta}_{1,0}, \boldsymbol{\beta}_{2,0}) + \alpha\Delta + \sigma + \bar{\epsilon}\tau \\ &\le V(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2') + \alpha\Delta + \sigma + \bar{\epsilon}\tau \\ &\le V^0(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \tau) + 2\alpha\Delta + 2\sigma + 2\bar{\epsilon}\tau \\ &\le (H_g + 2\alpha + 2s + 2\bar{\epsilon}C_\tau)\Delta = H_*^{\mathrm{err}}\Delta.\end{aligned} \tag{140}$$

By (136), each $\min(D_j, \tau) \le 16H_*^{\mathrm{err}}\Delta/p_0 < c_\tau\Delta \le \tau$. Thus $D_j \le 16H_*^{\mathrm{err}}\Delta/p_0 < \Delta/2$. The nearest candidate to each true coefficient must be different, since otherwise the triangle inequality would give $\Delta \le D_1 + D_2 < \Delta$. The distinct nearest candidates define a permutation. Hence $\mathcal{E}(\boldsymbol{\theta}_0) \le \sqrt{2}\max_j D_j \le 16\sqrt{2}H_*^{\mathrm{err}}\Delta/p_0$, proving the initialization error bound in Proposition C.3. □

Proof of Proposition C.3. Apply Lemma C.4, followed by Lemmas C.5 and C.6. Appendix C.6 verifies the numerical conditions of all three lemmas simultaneously. Each result holds on the same event and requires no conditioning on the fitted anchor. □

C.4 UNIFORM GROWTH OF NORMALIZED OBJECTIVES

An unnormalized product objective can fail after one response outlier. We first demonstrate that failure, then prove that normalized objectives retain local growth uniformly over all admissible references and retained sets.

Proposition C.7 (An unnormalized product can lose local minimality)**.** *Fix a finite noiseless sample with at least one nonzero covariate. There exists a replacement of a single response for which the true pair is not a local minimizer of* $n^{-1}\sum_{i=1}^{n}|q_i(\boldsymbol{\theta})|$.

One arbitrary response can overwhelm the local directional derivative of the raw product objective. The denominator in (54) removes the derivative dependence on response-outlier amplitude.

Proof. Replace row i_0 by $y_{i_0} = R > 0$ and take $\boldsymbol{d}_1 = \boldsymbol{d}_2 = \boldsymbol{x}_{i_0}/\|\boldsymbol{x}_{i_0}\|$. For sufficiently large R, both replaced residuals at the truth are positive. The directional derivative of the absolute product on this row is

$$-\{2R - \boldsymbol{x}_{i_0}^\top(\boldsymbol{\beta}_1^* + \boldsymbol{\beta}_2^*)\}\|\boldsymbol{x}_{i_0}\|. \tag{141}$$

For every remaining row the right directional derivative at the truth is finite and bounded by $\Delta|\boldsymbol{x}_i^\top\boldsymbol{u}^*|\max_j|\boldsymbol{x}_i^\top\boldsymbol{d}_j|$. Their sum is independent of R. Thus the total directional derivative is negative for sufficiently large R, excluding local minimality. □

For the local expansions, define the two possible true stacks and choose the ordering nearest the initializer:

$$\boldsymbol{\theta}^{*,\pi} = \big((\boldsymbol{\beta}^*_{\pi(1)})^\top, (\boldsymbol{\beta}^*_{\pi(2)})^\top\big)^\top, \qquad \pi_0 \in \arg\min_{\pi\in\mathcal{S}_2}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^{*,\pi}\|. \tag{142}$$

Lemma C.11 proves that, on the recovery event,

$$\mathcal{E}(\boldsymbol{\theta}) = \|\boldsymbol{\theta} - \boldsymbol{\theta}^{*,\pi_0}\| \qquad (\boldsymbol{\theta}\in\Omega). \tag{143}$$

Within the appendix local expansions, $\boldsymbol{\theta}^*$ abbreviates the explicitly ordered vector $\boldsymbol{\theta}^{*,\pi_0}$ in (142). The main recovery bounds remain expressed through $\mathcal{E}$.

For the local analysis, choose $\pi_0 \in \arg\min_{\pi\in\mathcal{S}_2}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^{*,\pi}\|$ and put

$$\widetilde{\boldsymbol{\beta}}^*_j = \boldsymbol{\beta}^*_{\pi_0(j)}, \qquad \widetilde{p}_j = p_{\pi_0(j)}, \qquad \widetilde{z}_i = \pi_0^{-1}(z_i), \qquad \boldsymbol{x}_i^\top\widetilde{\boldsymbol{\beta}}^*_{z_i} = \boldsymbol{x}_i^\top\boldsymbol{\beta}^*_{z_i} = y_i^0. \tag{144}$$

Use the relabeled coefficients, probabilities, and labels in all subsequent local expansions, with tildes suppressed. The event in Lemma C.2 is unchanged since its label-specific properties hold for both labels. Lemma C.11 proves that the resulting stack $\boldsymbol{\theta}^*$ is the uniquely minimizing ordering throughout Ω.

Proposition C.8 (Uniform local growth for changing objectives)**.** *On the event and under the constants of Theorem C.1, every reference pair* $\boldsymbol{\vartheta}\in\Omega$ *and every retained set* I *with* $|I| \ge n - \lfloor\bar{\epsilon}n\rfloor$ *define weights as in* (54)*. Let* $F_{\boldsymbol{\vartheta},I}$ *be the resulting frozen objective. Then*

$$F_{\boldsymbol{\vartheta},I}(\boldsymbol{\theta}) - F_{\boldsymbol{\vartheta},I}(\boldsymbol{\theta}^*) \ge \mu\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| - (6\sigma + \kappa/2) \qquad (\boldsymbol{\theta}\in\Omega). \tag{145}$$

Proposition C.8 supplies the uniform growth bound. The retained set and normalizer may depend on the same observations used in the objective. No independence between iterations is needed. The additive loss term in (145) becomes the noise floor in (69), while the amplitude of a response replacement never appears.

Proof of Proposition C.8. Fix an arbitrary reference pair $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1^\top, \boldsymbol{\vartheta}_2^\top)^\top \in \Omega$ and any allowed retained set I. Write

$$d_i = a\tau + |y_i - \boldsymbol{x}_i^\top\boldsymbol{\vartheta}_1| + |y_i - \boldsymbol{x}_i^\top\boldsymbol{\vartheta}_2|, \quad w_i = \mathbb{I}\{i\in I\}/d_i, \quad \ell = \|\boldsymbol{X}^\top\operatorname{diag}(\boldsymbol{w})\boldsymbol{X}/n\|_{\mathrm{op}} \le \overline{L}. \tag{146}$$

The constants ensure $\mathcal{E}_0 \le r/4$ and $5r/4 \le \Delta$. Therefore every point of Ω has aligned error at most $5r/4$, including $\boldsymbol{\vartheta}$. For good rows use $\widetilde{y}_i = y_i^0$, and for replaced rows use $\widetilde{y}_i = y_i$. Define the auxiliary frozen loss

$$F^0(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} w_i|(\widetilde{y}_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}_1)(\widetilde{y}_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}_2)|. \tag{147}$$

The weights still use the observed responses. Let $\boldsymbol{d}_j = \boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*$, stack them as $\boldsymbol{d}$, and set $\mathcal{E} = \|\boldsymbol{d}\|$.

Clean first-order growth. For label j, the clean product vanishes at the truth. Its first-order term has absolute value $\Delta|\boldsymbol{x}_i^\top \boldsymbol{u}^*|\,|\boldsymbol{x}_i^\top \boldsymbol{d}_j|$. If $\boldsymbol{d}_j \neq \mathbf{0}$, apply (E4) with $\boldsymbol{w} = \boldsymbol{d}_j/\|\boldsymbol{d}_j\|$ and

$$\boldsymbol{v}_k = (\boldsymbol{\beta}_j^* - \boldsymbol{\vartheta}_k)/\Delta \quad (k = 1, 2). \tag{148}$$

Both $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ have norm at most 2. On the resulting $p_0 n/2$ rows,

$$d_i \le aC_\tau\Delta + \Delta(|\boldsymbol{x}_i^\top \boldsymbol{v}_1| + |\boldsymbol{x}_i^\top \boldsymbol{v}_2|) + 2\sigma \le (aC_\tau + 34)\Delta. \tag{149}$$

Removing the replaced and discarded rows deletes at most $2\bar{\epsilon}n \le p_0 n/8$ indices. At least $p_0 n/4$ remain. Their total first-order contribution, divided by n, is at least

$$\frac{p_0 a_0^2}{4(aC_\tau + 34)}\|\boldsymbol{d}_j\| = \mu_0\|\boldsymbol{d}_j\|. \tag{150}$$

For $\boldsymbol{d}_j = \mathbf{0}$, the bound in (150) is zero. Summing (150) over the two disjoint label groups gives at least $\mu_0(\|\boldsymbol{d}_1\| + \|\boldsymbol{d}_2\|) \ge \mu_0\mathcal{E}$.

Replaced first-order terms. For a replaced row and $k \in \{1, 2\}$,

$$w_i|y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_k^*| \le 1 + \frac{|\boldsymbol{x}_i^\top(\boldsymbol{\vartheta}_k - \boldsymbol{\beta}_k^*)|}{a\tau}. \tag{151}$$

The absolute first-order sum from replaced rows is thus at most

$$\frac{1}{n}\sum_{i\in O}\sum_{j=1}^{2}|\boldsymbol{x}_i^\top \boldsymbol{d}_j| + \frac{1}{na\tau}\sum_{i\in O}\sum_{j=1}^{2}|\boldsymbol{x}_i^\top(\boldsymbol{\vartheta}_{3-j} - \boldsymbol{\beta}_{3-j}^*)|\,|\boldsymbol{x}_i^\top \boldsymbol{d}_j| \tag{152}$$

$$\le \sqrt{2}\sqrt{\bar{\epsilon}\gamma}\,\mathcal{E} + \frac{\gamma}{a\tau}\|\boldsymbol{\vartheta} - \boldsymbol{\theta}^*\|\,\mathcal{E} \le D\mathcal{E}, \qquad D := \sqrt{2}\left(\sqrt{\bar{\epsilon}\gamma} + \frac{\gamma}{ac_\tau}\right). \tag{153}$$

The first inequality in (153) uses (E2) and the Cauchy–Schwarz inequality across rows and then across coefficient blocks. The final inequality in (153) uses $\|\boldsymbol{\vartheta} - \boldsymbol{\theta}^*\| \le \Delta$ and $\tau \ge c_\tau\Delta$.

Second-order remainder. The exact quadratic expansion has remainder $(\boldsymbol{x}_i^\top \boldsymbol{d}_1)(\boldsymbol{x}_i^\top \boldsymbol{d}_2)$. Hence

$$\frac{1}{n}\sum_{i=1}^{n} w_i|(\boldsymbol{x}_i^\top \boldsymbol{d}_1)(\boldsymbol{x}_i^\top \boldsymbol{d}_2)| \le \frac{1}{2n}\sum_{i=1}^{n} w_i\{(\boldsymbol{x}_i^\top \boldsymbol{d}_1)^2 + (\boldsymbol{x}_i^\top \boldsymbol{d}_2)^2\} \le \frac{\ell}{2}\mathcal{E}^2. \tag{154}$$

The triangle inequality on good rows and the reverse triangle inequality on replaced rows now give

$$F^0(\boldsymbol{\theta}) - F^0(\boldsymbol{\theta}^*) \ge (\mu_0 - D)\mathcal{E} - \frac{\ell}{2}\mathcal{E}^2 \ge \mu\mathcal{E}. \tag{155}$$

For the last step, Appendix C.6 ensures $D \le \mu_0/4$, while $\mathcal{E} \le 5r/4 \le \mu/\bar{L}$ and $\mu = \mu_0/2$.

Bounded-noise perturbation with observed weights. Let F^{abs} be the observed frozen loss with ϕ_0. On a good row put $r_{ik}^0(\boldsymbol{\theta}) = y_i^0 - \boldsymbol{x}_i^\top \boldsymbol{\beta}_k$. Expanding the noisy product gives

$$q_i(\boldsymbol{\theta}) - r_{i1}^0(\boldsymbol{\theta})r_{i2}^0(\boldsymbol{\theta}) = \xi_i\{r_{i1}^0(\boldsymbol{\theta}) + r_{i2}^0(\boldsymbol{\theta})\} + \xi_i^2. \tag{156}$$

Also

$$|r_{i1}^0(\boldsymbol{\theta})| + |r_{i2}^0(\boldsymbol{\theta})| \le d_i + 2\sigma + \sum_{k=1}^{2}|\boldsymbol{x}_i^\top(\boldsymbol{\vartheta}_k - \boldsymbol{\beta}_k)|. \tag{157}$$

Since $d_i \ge a\tau$ and $n^{-1}\sum_{i=1}^n\sum_{k=1}^2|\boldsymbol{x}_i^\top(\boldsymbol{\vartheta}_k - \boldsymbol{\beta}_k)| \le \sqrt{2\Lambda}\|\boldsymbol{\vartheta} - \boldsymbol{\theta}\|$, it follows that

$$|F^{\mathrm{abs}}(\boldsymbol{\theta}) - F^0(\boldsymbol{\theta})| \le \sigma\left(1 + \frac{3\sigma}{a\tau} + \frac{\sqrt{2\Lambda}\|\boldsymbol{\vartheta} - \boldsymbol{\theta}\|}{a\tau}\right) \le 3\sigma. \tag{158}$$

Indeed, $\sigma \le a\tau/3$, $\|\boldsymbol{\vartheta} - \boldsymbol{\theta}\| \le 2r$, and

$$\frac{2r\sqrt{2\Lambda}}{a\tau} = \frac{\mu}{4\sqrt{2\Lambda}} \le \frac{\mu}{4} < 1. \tag{159}$$

Bound (158) also holds at $\boldsymbol{\theta}^* \in \Omega$. Finally, $|s| - \kappa/2 \le \phi_\kappa(s) \le |s|$. Combining the two scalar-loss inequalities at both arguments, (155), and (158) yields

$$\begin{aligned} F_{\boldsymbol{\vartheta},I}(\boldsymbol{\theta}) - F_{\boldsymbol{\vartheta},I}(\boldsymbol{\theta}^*) &\ge F^{\mathrm{abs}}(\boldsymbol{\theta}) - F^{\mathrm{abs}}(\boldsymbol{\theta}^*) - \kappa/2 \\ &\ge F^0(\boldsymbol{\theta}) - F^0(\boldsymbol{\theta}^*) - 6\sigma - \kappa/2 \\ &\ge \mu\mathcal{E} - 6\sigma - \kappa/2. \end{aligned} \tag{160}$$

The first-order bound (153), auxiliary growth bound (155), and perturbation bound (158) hold uniformly over $\boldsymbol{\vartheta}, I, \boldsymbol{\theta}$. Their combination proves Proposition C.8. □

C.5 Convex model accuracy and optimization certificates

The recovery recursion needs a quadratic bound on the model error and an explicit bound on the model optimization error. Proposition C.9 supplies both for the smoothed formulation. Proposition C.10 supplies the corresponding objective-gap certificate for the absolute-loss formulation used by SARP-P.

Proposition C.9 (Model accuracy and observable solver gap)**.** *For fixed nonnegative weights in* (54)*, the unregularized affine model $f_t(\boldsymbol{d}) = n^{-1}\sum_{i=1}^n \phi_\kappa(c_{i,t} + \boldsymbol{A}_{i,t}\boldsymbol{d})$ satisfies*

$$|F_t(\boldsymbol{\theta}_t + \boldsymbol{d}) - f_t(\boldsymbol{d})| \le \frac{\ell_t}{2}\|\boldsymbol{d}\|^2. \tag{161}$$

Assume $\ell_t > 0$. For $\kappa > 0$ and any $\boldsymbol{d} \in D_t$, define $\boldsymbol{s} = \mathrm{clip}((\boldsymbol{c}_t + \boldsymbol{A}_t\boldsymbol{d})/\kappa, [-1,1])$, $\boldsymbol{g} = \boldsymbol{A}_t^\top \boldsymbol{s}/n$, and $\boldsymbol{v} = \mathrm{Proj}_{D_t}(-\boldsymbol{g}/\ell_t)$. Then

$$0 \le M_t(\boldsymbol{d}) - \min_{D_t} M_t \le G(\boldsymbol{d}) := \frac{\ell_t}{2}(\|\boldsymbol{d}\|^2 - \|\boldsymbol{v}\|^2) + \langle \boldsymbol{g}, \boldsymbol{d} - \boldsymbol{v}\rangle. \tag{162}$$

Algorithm 3 terminates for every $\eta_t > 0$.

Equation (161) bounds the quadratic model error used in composite optimization (Duchi & Ruan, 2019). Equation (162) turns the abstract inexactness condition into a computable certificate. A small change in consecutive coefficients is not used as a mathematical substitute for the objective-gap certificate.

Proposition C.10 (A certificate for the absolute-composite conic model)**.** *Fix a model with $\kappa = 0$, $\ell_t > 0$, and $D_t \ne \varnothing$. For any $\boldsymbol{s} \in [-1,1]^n$, put*

$$\boldsymbol{g} = \boldsymbol{A}_t^\top \boldsymbol{s}/n, \qquad \boldsymbol{v} = \mathrm{Proj}_{D_t}(-\boldsymbol{g}/\ell_t), \qquad D(\boldsymbol{s}) = \boldsymbol{s}^\top \boldsymbol{c}_t/n + \ell_t\|\boldsymbol{v}\|^2/2 + \boldsymbol{g}^\top \boldsymbol{v}. \tag{163}$$

For every feasible $\boldsymbol{d} \in D_t$,

$$0 \le M_t(\boldsymbol{d}) - \min_{D_t} M_t \le M_t(\boldsymbol{d}) - D(\boldsymbol{s}). \tag{164}$$

If $s_i \in \partial|c_{i,t} + \boldsymbol{A}_{i,t}\boldsymbol{d}|$, the right side equals $G(\boldsymbol{d})$ in (162)*, for the specified subgradient vector $\boldsymbol{s}$.*

Proposition C.10 gives a directly checkable objective-gap upper bound for the nonsmooth normalized composite iteration used by SARP-P. Any approximate dual vector can be clipped to the box before evaluating $D(\boldsymbol{s})$. Numerical feasibility and arithmetic error should still be reported separately.

Proof. For $|s_i| \le 1$ and every real u_i, $s_i u_i \le |u_i|$. Consequently, for all $\boldsymbol{u} \in D_t$,

$$M_t(\boldsymbol{u}) \ge \frac{1}{n}\boldsymbol{s}^\top(\boldsymbol{c}_t + \boldsymbol{A}_t\boldsymbol{u}) + \frac{\ell_t}{2}\|\boldsymbol{u}\|^2 \tag{165}$$

$$= \frac{1}{n}\boldsymbol{s}^\top \boldsymbol{c}_t - \frac{\|\boldsymbol{g}\|^2}{2\ell_t} + \frac{\ell_t}{2}\|\boldsymbol{u} + \boldsymbol{g}/\ell_t\|^2 \tag{166}$$

$$\ge \frac{1}{n}\boldsymbol{s}^\top \boldsymbol{c}_t - \frac{\|\boldsymbol{g}\|^2}{2\ell_t} + \frac{\ell_t}{2}\|\boldsymbol{v} + \boldsymbol{g}/\ell_t\|^2 = D(\boldsymbol{s}). \tag{167}$$

Minimize the left side over D_t and subtract from $M_t(\boldsymbol{d})$. When $s_i \in \partial|c_{i,t} + \boldsymbol{A}_{i,t}\boldsymbol{d}|$, the scalar identity $|u_i| = s_i u_i$ gives

$$M_t(\boldsymbol{d}) - D(\boldsymbol{s}) = \frac{\ell_t}{2}(\|\boldsymbol{d}\|^2 - \|\boldsymbol{v}\|^2) + \langle \boldsymbol{g}, \boldsymbol{d} - \boldsymbol{v}\rangle. \tag{168}$$

The absolute-value subdifferential is $\{1\}$ for a positive argument, $\{-1\}$ for a negative argument, and $[-1, 1]$ at zero, as follows directly from its two affine branches. Finally, the adaptive model retains at least one nonzero Gaussian row and assigns it a positive weight. Thus its positive semidefinite weighted covariance is nonzero and $\ell_t > 0$ almost surely. □

Proof of Proposition C.9. The scalar function ϕ_κ is 1-Lipschitz. The exact expansion

$$q_i(\boldsymbol{\theta}_t + \boldsymbol{d}) = q_i(\boldsymbol{\theta}_t) + \boldsymbol{J}_{i,t}\boldsymbol{d} + (\boldsymbol{x}_i^\top \boldsymbol{d}_1)(\boldsymbol{x}_i^\top \boldsymbol{d}_2) \tag{169}$$

and (154), with $(\boldsymbol{w}, \ell, \boldsymbol{d}) = (\boldsymbol{w}_t, \ell_t, \boldsymbol{d})$, prove (161).

For the gap, completing the square in a scalar maximization gives the exact identity

$$\phi_\kappa(z) = \max_{|s|\le 1}\{sz - \kappa s^2/2\}, \qquad s = \mathrm{clip}(z/\kappa, [-1, 1]) \text{ is a maximizer.} \tag{170}$$

Fix the vector $\boldsymbol{s}$ computed at the feasible $\boldsymbol{d}$. For every $\boldsymbol{u} \in D_t$,

$$M_t(\boldsymbol{u}) \ge \frac{\boldsymbol{s}^\top \boldsymbol{c}_t}{n} - \frac{\kappa\|\boldsymbol{s}\|^2}{2n} + \boldsymbol{g}^\top \boldsymbol{u} + \frac{\ell_t}{2}\|\boldsymbol{u}\|^2. \tag{171}$$

Equality holds in the scalar-loss part at $\boldsymbol{u} = \boldsymbol{d}$. The quadratic right side is minimized over D_t by $\boldsymbol{v} = \mathrm{Proj}_{D_t}(-\boldsymbol{g}/\ell_t)$. Subtracting its value at $\boldsymbol{v}$ from $M_t(\boldsymbol{d})$ gives exactly (162). The square-completion argument gives a dual lower bound without assuming numerical stationarity.

For termination, M_t has gradient $\ell_t\boldsymbol{d} + \boldsymbol{A}_t^\top \mathrm{clip}((\boldsymbol{c}_t + \boldsymbol{A}_t\boldsymbol{d})/\kappa, [-1, 1])/n$, which is K_t-Lipschitz. Let $\boldsymbol{d}^+$ be its projected step and $\boldsymbol{z} \in D_t$. The projection optimality inequality is

$$\langle \nabla M_t(\boldsymbol{d}) + K_t(\boldsymbol{d}^+ - \boldsymbol{d}), \boldsymbol{z} - \boldsymbol{d}^+\rangle \ge 0. \tag{172}$$

Smoothness at $\boldsymbol{d}$ and convexity at $\boldsymbol{z}$ imply

$$M_t(\boldsymbol{d}^+) - M_t(\boldsymbol{z}) \le \langle \nabla M_t(\boldsymbol{d}), \boldsymbol{d}^+ - \boldsymbol{z}\rangle + \tfrac{K_t}{2}\|\boldsymbol{d}^+ - \boldsymbol{d}\|^2 \tag{173}$$

$$\le \tfrac{K_t}{2}\{\|\boldsymbol{d} - \boldsymbol{z}\|^2 - \|\boldsymbol{d}^+ - \boldsymbol{z}\|^2\}. \tag{174}$$

Taking $\boldsymbol{z} = \boldsymbol{d}$ shows monotonicity. Taking $\boldsymbol{z}$ to be the unique minimizer $\boldsymbol{d}^*$ and summing for k steps gives

$$M_t(\boldsymbol{d}^{(k)}) - M_t(\boldsymbol{d}^*) \le \frac{K_t\|\boldsymbol{d}^{(0)} - \boldsymbol{d}^*\|^2}{2k}. \tag{175}$$

Strong convexity gives $\|\boldsymbol{d}^{(k)} - \boldsymbol{d}^*\|^2 \le 2\{M_t(\boldsymbol{d}^{(k)}) - M_t(\boldsymbol{d}^*)\}/\ell_t \to 0$. At the minimizer the optimality condition gives $\boldsymbol{d}^* = \mathrm{Proj}_{D_t}(-\boldsymbol{g}(\boldsymbol{d}^*)/\ell_t)$, hence $G(\boldsymbol{d}^*) = 0$. Continuity of the projection and gradient implies $G(\boldsymbol{d}^{(k)}) \to 0$, proving finite termination for every positive requested gap. □

C.6 Simultaneous choices of constants

The following constants satisfy the initialization, local-growth, and recursion conditions simultaneously. All quantities depend only on the supplied p_0 and the chosen a. Use (66), (124), and $B_1 = 17$, and set

$$\rho = \frac{\mu a c_\tau}{128}, \qquad H_{\max} = \min\left\{\frac{p_0\rho}{32\sqrt{2}}, \frac{p_0 c_\tau}{32}, \frac{p_0}{64}\right\}, \qquad \delta_{\mathrm{g}} = \frac{H_{\max}}{2C_\tau}, \tag{176}$$

$$K_E = 3 + 5/c_\tau, \qquad K_H = 5 + 4K_E/\kappa_0 + 2C_\tau, \tag{177}$$

$$\lambda_* = \frac{1}{2}\min\left\{1, \frac{p_0}{16}, \frac{c_{\mathrm{G}}}{8}, \frac{\kappa_0}{4K_E}, \frac{\mu_0}{4\sqrt{2}(1 + 1/(ac_\tau))}, \frac{H_{\max}}{2K_H}\right\}, \tag{178}$$

$$\alpha = \min\{\lambda_*, c_{\mathrm{G}}\lambda_*^2/(32B_1)\}, \tag{179}$$

$$\epsilon_{\mathrm{aux}} = \min\{\lambda_*, c_{\mathrm{G}}^2\lambda_*/(64B_1^2)\}, \qquad \epsilon_0 = \min\{\epsilon_{\mathrm{aux}}, \lambda_*^2/(16 \cdot 100^2)\}, \tag{180}$$

$$s_0 = \min\left\{\frac{1}{32}, \lambda_*, \frac{c_{\mathrm{G}}\lambda_*^2}{32B_1}, \frac{ac_\tau}{3}, \frac{\mu^2 a c_\tau}{3072}\right\}. \tag{181}$$

Every minimum is strictly positive. Increase the sample constant such that Lemma C.2 holds and

$$100\{d + \log(n/\delta)\}/n \le \lambda_*/2. \tag{182}$$

The sample condition in (67) permits the required bounds. For $0 < x \le 1$, $x \log(e/x) \le 2\sqrt{x}$. Thus

$$\gamma \le 200\sqrt{\epsilon_0} + \lambda_*/2 \le \lambda_*. \tag{183}$$

In particular, $\bar{\epsilon} \le p_0/16$, $2\sqrt{\bar{\epsilon}\gamma} \le 2\lambda_* \le c_{\mathrm{G}}/4$, and $s \le 1/32$. The anchor and scale lemma applies. Moreover,

$$\frac{4B_1}{c_{\mathrm{G}}}\sqrt{\bar{\epsilon}\gamma} \le \frac{4B_1}{c_{\mathrm{G}}}\sqrt{\frac{c_{\mathrm{G}}^2\lambda_*^2}{64B_1^2}} = \lambda_*/2, \tag{184}$$

$$\sqrt{\frac{4B_1}{c_{\mathrm{G}}}(\alpha + s)} \le \sqrt{\frac{4B_1}{c_{\mathrm{G}}}\frac{c_{\mathrm{G}}\lambda_*^2}{16B_1}} = \lambda_*/2. \tag{185}$$

Hence $\zeta \le \lambda_* \le 1$. Next,

$$E_* \le K_E\lambda_* \le \kappa_0/4, \qquad D \le \sqrt{2}(1 + 1/(ac_\tau))\lambda_* \le \mu_0/4. \tag{186}$$

The initialization error bound satisfies

$$H_*^{\mathrm{err}} \le K_H\lambda_* + \delta_{\mathrm{g}}C_\tau/2 \le H_{\max}/2 + H_{\max}/4 = 3H_{\max}/4. \tag{187}$$

It follows that $16H_*^{\mathrm{err}}/p_0 < \min(c_\tau, 1/2)$ and

$$\mathcal{E}_0 \le \frac{16\sqrt{2}}{p_0}H_{\max}\Delta \le \rho\Delta/2, \qquad r = \frac{\mu a\tau}{16\Lambda} \ge \frac{\mu a c_\tau \Delta}{32} = 4\rho\Delta. \tag{188}$$

In particular $\mathcal{E}_0 \le r/4$. Conversely, $\Lambda \ge 1/2$ and $\tau \le C_\tau\Delta$ imply

$$\frac{5r}{4} \le \frac{5\mu a C_\tau}{32}\Delta < \frac{5\Delta}{32} < \frac{\Delta}{\sqrt{2}}, \qquad \mu a C_\tau \le \frac{a_0^2 p_0}{8} < 1. \tag{189}$$

The bounds on $\mathcal{E}_0$ and r verify the geometric conditions in the local argument. The condition $\sigma \le a\tau/3$ follows from $s_0 \le ac_\tau/3$. Finally,

$$\frac{6\sigma + \kappa/2}{\mu} \le \frac{6s_0\Delta}{\mu} \le \frac{\mu a c_\tau \Delta}{512} \le \frac{r}{16}. \tag{190}$$

All initialization, local geometry, and recursion conditions have now been checked.

Scope of the sufficient constants. The simultaneous constant choices give positive thresholds and a sample size nearly linear in dimension. The sufficient thresholds are extremely conservative. The grid and trust radius obtained from them are not practical parameter recommendations. In particular, the theorem does not certify the 5% contamination level or the practical radius in our experiments. Improving the constants and providing a computable basin diagnostic would materially strengthen the result.

Lemma C.11 (One alignment throughout the feasible ball)**.** *Suppose $\mathcal{E}(\boldsymbol{\theta}_0) \le r/4$ and $5r/4 < \Delta/\sqrt{2}$. Let $\boldsymbol{\theta}^*$ be a stacked ordering of the true pair that attains $\mathcal{E}(\boldsymbol{\theta}_0)$. Then*

$$\mathcal{E}(\boldsymbol{\theta}) = \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \qquad (\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, r)), \tag{191}$$

and the ordering $\boldsymbol{\theta}^*$ uniquely minimizes the error throughout the ball.

The lemma permits one fixed stack of the true coefficients in the local expansions while keeping every stated recovery error invariant to the component labels.

Proof. Let $\boldsymbol{\theta}^{\mathrm{sw}}$ be the other ordering of the true pair. Their distance is

$$\|\boldsymbol{\theta}^{\mathrm{sw}} - \boldsymbol{\theta}^*\|^2 = \|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\|^2 + \|\boldsymbol{\beta}_2^* - \boldsymbol{\beta}_1^*\|^2 = 2\Delta^2. \tag{192}$$

For every $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, r)$,

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \le \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\| \le 5r/4 < \Delta/\sqrt{2}. \tag{193}$$

The reverse triangle inequality gives

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{sw}}\| \ge \sqrt{2}\Delta - \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| > \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|. \tag{194}$$

There are exactly two label permutations. Inequality (194) proves (191) and uniqueness. □

C.7 Completion of the deterministic recovery argument

The initialization, local-growth, and model-accuracy results now apply on the same event. Comparing each convex model with the feasible true pair gives the error recursion. A separate induction establishes the stated doubly exponential optimization bound.

Proof of Theorem C.1. The constants in Appendix C.6 imply the hypotheses of Proposition C.3 and $\mathcal{E}_0 \le r/4$. Choose $\boldsymbol{\theta}^*$ as in Lemma C.11. Then $\boldsymbol{\theta}^* \in \Omega$ and $\mathcal{E}_t = \|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\| \le 5r/4$ for every feasible iterate. Proposition C.8 is therefore applicable at every iteration, regardless of previous updates. Write $B = 6\sigma + \kappa/2$. Model accuracy and the subproblem objective-gap bound give

$$F_t(\boldsymbol{\theta}_{t+1}) \le M_t(\boldsymbol{d}_t) \le M_t(\boldsymbol{\theta}^* - \boldsymbol{\theta}_t) + \eta_t \tag{195}$$

$$\le F_t(\boldsymbol{\theta}^*) + \ell_t \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_t\|^2 + \eta_t. \tag{196}$$

Combining (196) with (145) at $\boldsymbol{\theta} = \boldsymbol{\theta}_{t+1}$ yields $\mu\mathcal{E}_{t+1} \le \ell_t \mathcal{E}_t^2 + B + \eta_t$, which proves (69).

For the explicit rate, put $b = B/\mu$ and $b_t^{\mathrm{opt}} = r2^{-2^t}$. The constants give $b \le r/16$, and $\overline{L}/\mu = 1/(16r)$. Assume inductively $\mathcal{E}_t \le b_t^{\mathrm{opt}} + 2b$. Since $b_t^{\mathrm{opt}} \le r/2$ and $b_{t+1}^{\mathrm{opt}} = (b_t^{\mathrm{opt}})^2/r$,

$$\mathcal{E}_{t+1} \le \frac{(b_t^{\mathrm{opt}} + 2b)^2}{16r} + b + \frac{b_{t+1}^{\mathrm{opt}}}{2} \tag{197}$$

$$\le \frac{9b_{t+1}^{\mathrm{opt}}}{16} + \frac{b}{8} + \frac{b}{64} + b \le b_{t+1}^{\mathrm{opt}} + 2b. \tag{198}$$

The base case holds since $\mathcal{E}_0 \le r/4 \le b_0^{\mathrm{opt}} + 2b$. The induction proves (70). □

C.8 Independent noise and completion of the main theorem

A tail threshold converts independent errors into bounded perturbations plus additional response outliers. A maximum-noise event gives an alternative reduction without increasing the contamination budget. Applying the auxiliary bounded-perturbation theorem on either event proves Theorem 4.1.

Corollary C.12 (Independent random inlier noise)**.** *Suppose the ξ_i are independent and satisfy $\mathbb{P}(|\xi_i| > u) \le p_{\mathrm{tail}}(u)$. They are independent of the clean design and labels. Fix $u \ge 0$ and let*

$$b_n(u) = p_{\mathrm{tail}}(u) + \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{199}$$

If $\epsilon + b_n(u) \le \bar{\epsilon} \le \epsilon_0$ and $u + \kappa \le s_0\Delta$, then Theorem C.1 holds with $\sigma = u$ and total probability at least $1 - \delta$, using failure probability $\delta/2$ in its sample condition. Sub-Gaussian and sub-exponential tails can respectively be represented by

$$p_{\mathrm{tail}}(u) = \min\{1, 2e^{-u^2/(2v^2)}\} \quad \textit{or} \quad p_{\mathrm{tail}}(u) = \min\{1, 2e^{-u/v}\}, \tag{200}$$

where $v > 0$ is a scalar tail scale. Alternatively, if $np_{\mathrm{tail}}(u) \le \delta/2$, the conclusion holds with $\bar{\epsilon} \ge \epsilon$ and no extra trimming budget.

The exceedance-count reduction treats large errors as additional response outliers and increases the contamination budget. The maximum-error reduction leaves the budget unchanged. Both reductions yield recovery under independent errors, without the sharp stochastic rates established in the different model of Chen et al. (2018).

Proof of Corollary C.12. Let $A_u = \{i : |\xi_i| > u\}$. Its independent indicator variables have expectations at most $p_{\mathrm{tail}}(u)$. The one-sided Hoeffding bound gives

$$\mathbb{P}\{|A_u|/n > p_{\mathrm{tail}}(u) + \sqrt{\log(2/\delta)/(2n)}\} \le \delta/2. \tag{201}$$

On the complementary event, $|O \cup A_u| \le n\{\epsilon + b_n(u)\} \le n\bar{\epsilon}$. Since cardinality is an integer, it is at most $\lfloor n\bar{\epsilon} \rfloor$. Outside $O \cup A_u$ the noise is bounded by u. Responses in $O \cup A_u$ are treated as arbitrary response outliers. The event of Theorem C.1, invoked with failure probability $\delta/2$, is uniform over all such sets and values. Intersecting the clean-sample and exceedance-count events proves the first claim with total failure at most δ. For the alternative claim, the union bound gives $\mathbb{P}(A_u \ne \varnothing) \le np_{\mathrm{tail}}(u) \le \delta/2$. No additional replacements are then needed. Substitution of either tail form in (200) proves the stated examples. □

Proof of Theorem 4.1. Apply the auxiliary Theorem C.1 with $\kappa = 0$ and clean-sample failure probability $\delta/2$. The auxiliary event depends only on the covariates and labels. Choose the constant in (23) to satisfy the auxiliary sample-size requirement with failure probability $\delta/2$.

First verify the grid spacing. The constants in (181) give

$$\mu a \le \mu a C_\tau \le a_0^2 p_0/8 < 1, \qquad \rho = \mu a c_\tau/128 < c_\tau/128. \tag{202}$$

The first entry of the minimum defining $H_{\max}$ is therefore the smallest. Hence

$$\frac{H_{\max}}{2C_\tau} = \frac{p_0 \mu a c_\tau}{8192\sqrt{2}C_\tau}. \tag{203}$$

A smaller grid spacing only decreases the nonnegative discretization-error bounds in Proposition C.3. Thus (22) is sufficient.

Let $A_u = \{i \in [n] : |\xi_i| > u\}$. Under the first alternative of (24), independence and the one-sided Hoeffding inequality yield

$$\mathbb{P}\left\{\frac{|A_u|}{n} > p_{\mathrm{tail}}(u) + \sqrt{\frac{\log(2/\delta)}{2n}}\right\} \le \delta/2. \tag{204}$$

On the complementary event, every allowed original replacement set O satisfies

$$|O \cup A_u| \le n\left\{\epsilon + p_{\mathrm{tail}}(u) + \sqrt{\frac{\log(2/\delta)}{2n}}\right\} \le n\bar{\epsilon}. \tag{205}$$

The cardinality $|O \cup A_u|$ is at most $\lfloor n\bar{\epsilon}\rfloor$. Outside $O \cup A_u$ the errors are bounded by u. Every response in $O \cup A_u$ can be treated as a response outlier. The auxiliary theorem applies with $\sigma = u$. Under the second alternative, the union bound gives

$$\mathbb{P}(A_u \neq \varnothing) \le n p_{\mathrm{tail}}(u) \le \delta/2. \tag{206}$$

The auxiliary theorem then applies with the original replacement set and $\sigma = u$. Both arguments also hold for $u = 0$.

Intersect the clean-sample event with the selected noise event. The combined failure probability is at most δ. Theorem C.1 gives initialization, the objective-gap recursion, and the error envelope on the intersection. Uniformity over bounded perturbations and replacement sets makes the conclusion simultaneous over every allowed adversary. Lemma C.11, with the radius inequality from Appendix C.6, identifies the local Euclidean errors with $\mathcal{E}_t = \mathcal{E}(\boldsymbol{\theta}_t)$. The initialization, recursion, and permutation-invariant error envelope therefore prove Theorem 4.1. □

C.9 A LIMITATION OF THE AUXILIARY BOUNDED-PERTURBATION CLASS

The auxiliary recovery result permits bounded errors to depend on the covariates. The following construction shows why a nonvanishing error floor can be unavoidable in that larger class. The construction does not satisfy the independence assumption of the main stochastic model.

Proposition C.13 (A noise floor for the auxiliary perturbation class)**.** *Let $\sigma > 0$, $0 < \epsilon < 1/2$, $n \ge 4/\epsilon$, and $A_\epsilon = \sqrt{2\log(4/\epsilon)}$. For every estimator there are two balanced experiments in the auxiliary bounded-perturbation class* (65)*, with the same separation and noise bound σ, such that on at least one experiment*

$$\mathbb{P}\left\{\mathcal{E}(\widehat{\boldsymbol{\theta}}) \ge \frac{\sigma}{\sqrt{2}A_\epsilon}\right\} \ge \frac{1 - e^{-n\epsilon^2/8}}{2}. \tag{207}$$

Proposition C.13 limits uniform guarantees over errors that may depend on the covariates and explains the nonvanishing floor in the auxiliary perturbation theorem. The construction uses covariate-dependent errors, and the proposition is not a lower bound for the independent-error model of Section 2. The proposition therefore does not establish optimality of the stochastic bounds in Theorems 4.1 and 4.2.

Proof of Proposition C.13. Choose any unit vector $\boldsymbol{v}$ and any distinct pair $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$. Put $b = 2\sigma/A_\epsilon$. World 0 has coefficients $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$. World 1 has $\boldsymbol{\beta}_1 + b\boldsymbol{v}, \boldsymbol{\beta}_2 + b\boldsymbol{v}$. Use the same Gaussian design and independent balanced labels in both worlds. The separation is unchanged. Define the tail set

$$A = \{i : |\boldsymbol{x}_i^\top \boldsymbol{v}| > A_\epsilon\}. \tag{208}$$

For $i \notin A$, choose noise $b\boldsymbol{x}_i^\top \boldsymbol{v}/2$ in world 0 and its negative in world 1. For $i \in A$, set the pre-replacement noise to zero in both worlds. Both error assignments have magnitude at most σ and depend on the covariates. If $|A| \le \lfloor \epsilon n \rfloor$, replace all responses in A by zero in both worlds. Otherwise replace no responses. The response-replacement rule satisfies the contamination budget on every sample realization.

On the event $E = \{|A| \le \lfloor \epsilon n \rfloor\}$ the entire observed sample is identical in the two worlds. Outside A its common response is $\boldsymbol{x}_i^\top \boldsymbol{\beta}_{z_i} + b\boldsymbol{x}_i^\top \boldsymbol{v}/2$. Inside A it is zero. By Lemma A.4, $\mathbb{P}(i \in A) \le 2e^{-A_\epsilon^2/2} = \epsilon/2$. Since $n \ge 4/\epsilon$, $\lfloor \epsilon n \rfloor / n \ge 3\epsilon/4$. The Hoeffding inequality therefore gives

$$\mathbb{P}(E^c) \le \exp\{-2n(\epsilon/4)^2\} = e^{-n\epsilon^2/8}. \tag{209}$$

The permutation-invariant distance between the two parameter pairs is $\sqrt{2}b$. Indeed, the aligned squared distance is $2b^2$, while the swapped squared distance is

$$\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 - b\boldsymbol{v}\|^2 + \|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1 - b\boldsymbol{v}\|^2 = 2\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|^2 + 2b^2. \tag{210}$$

The permutation-invariant distance obeys the triangle inequality by composing the two minimizing permutations and applying the Euclidean triangle inequality. Consequently a single output cannot have error less than $b/(2\sqrt{2})$ in both worlds. Couple any internal randomness of the estimator as well. On E its outputs coincide, hence the sum of the two error-event probabilities is at least $\mathbb{P}(E)$. One probability is at least $\mathbb{P}(E)/2$. Substitution of $b/(2\sqrt{2}) = \sigma/(\sqrt{2}A_\epsilon)$ proves the proposition. □

D Proof of Theorem 4.2 and Response Polishing

The response-polishing objective need not have the local geometry proved for the normalized product loss. The final acceptance test transfers the composite recovery bound to the full SARP-P output. We establish the transfer under auxiliary bounded perturbations and then apply the independent-error reduction used for Theorem 4.1.

D.1 Proof outline

The residual median bounds the observable acceptance slack by the bounded-perturbation level and the preceding coefficient error. The final normalized composite iteration bounds $F_{T-1}(\boldsymbol{c}) - F_{T-1}(\boldsymbol{\theta}^*)$. The acceptance rule transfers the objective bound to the returned pair, and uniform local growth converts the bound into coefficient error. The majorization results justify the response-polishing updates. The proposal enters the recovery proof only through the final acceptance test.

Run Algorithm 2 with $\kappa = 0$, the stated segment initialization, the feasible ball Ω, and the required objective-gap bounds. Residual calibration and response polishing produce one proposal. The deterministic acceptance test supplies recovery for any such data-dependent proposal.

Corollary D.1 (Unknown-scale guarded recovery)**.** *Assume the hypotheses of Theorem C.1 for the normalized composite iteration with $\kappa = 0$. Write*

$$P_T = \ell_{T-1}\mathcal{E}_{T-1}^2 + 6\sigma + \eta_{T-1}, \qquad C_{\mathrm{g}} = 1 + \frac{4}{q\mu}. \tag{211}$$

On the uniform clean-sample event of Theorem C.1, the output of Algorithm 1 obeys

$$\boxed{\|\boldsymbol{\theta}_{\mathrm{out}} - \boldsymbol{\theta}^*\| \le \frac{C_{\mathrm{g}}}{\mu} P_T + \frac{\sigma}{q\mu} + \frac{\upsilon}{\mu}.} \tag{212}$$

The acceptance rule gives the bound for the specified response-polishing proposal and for any other measurable proposal. In particular, the cutoff and family selection can use the same sample as the composite estimate.

Corollary D.1 preserves quadratic dependence on the preceding coefficient error and an explicit bounded-perturbation floor, with larger constants than the normalized composite bound. The guard protects data-dependent response polishing without requiring an optimal family or cutoff.

D.2 Residual scale and a finite calibration family

The retained residual median sets the response scale without a supplied noise level. A finite family of Huber and Tukey scores then determines the proposal objective. The calibrated scale, score family, and cutoff are fixed during response polishing.

Run normalized composite iteration for $T \geq 1$ iterations and denote its output by $\boldsymbol{c} = \boldsymbol{\theta}_T$. Keep the last frozen objective F_{T-1}, including the weights computed at $\boldsymbol{\theta}_{T-1}$. Let

$$m = n - \lfloor \bar{\epsilon} n \rfloor, \qquad a_i = \min_{j=1,2} |y_i - \boldsymbol{x}_i^\top \boldsymbol{c}_j|, \qquad J = \arg \min_{I \subseteq [n],\ |I|=m} \sum_{i \in I} a_i. \tag{213}$$

The two blocks of $\boldsymbol{c}$ are $\boldsymbol{c}_1, \boldsymbol{c}_2 \in \mathbb{R}^d$. Ties are resolved by index. For ordered real values $b_{(1)} \leq \cdots \leq b_{(m)}$, define their median as $b_{((m+1)/2)}$ when m is odd and $(b_{(m/2)} + b_{(m/2+1)})/2$ when m is even. The averaged-middle convention matches the implementation and never exceeds the upper median.

Let Φ be the standard normal distribution function, put $q = \Phi^{-1}(3/4) > 0$, and define

$$s_{\mathrm{raw}} = \frac{\mathrm{med}(\{a_i : i \in J\})}{q}, \qquad \widehat{s} = \begin{cases} 0, & s_{\mathrm{raw}} = 0, \\ \max\{s_{\mathrm{raw}}, \upsilon\}, & s_{\mathrm{raw}} > 0, \end{cases} \qquad \upsilon \geq 0. \tag{214}$$

The scalar υ is an optional numerical floor in response units. The recovery bound retains the term contributed by υ. The value $\widehat{s}$ is a residual scale proxy. Under auxiliary bounded perturbations, $\widehat{s}$ is not asserted to estimate a standard deviation. If $s_{\mathrm{raw}} = 0$, return $\boldsymbol{c}$ and skip the proposal before applying any positive numerical floor.

For a cutoff $\lambda > 0$, define the Huber and Tukey losses by

$$H_\lambda(u) = \begin{cases} u^2/2, & |u| \leq \lambda, \\ \lambda|u| - \lambda^2/2, & |u| > \lambda, \end{cases} \tag{215}$$

$$T_\lambda(u) = \frac{\lambda^2}{6} \left[1 - \left(1 - \frac{u^2}{\lambda^2} \right)_+^3 \right], \qquad v_+ = \max(v, 0). \tag{216}$$

The Huber and Tukey score functions and derivatives are

$$\psi_{\mathrm{H},\lambda}(u) = \mathrm{clip}(u, [-\lambda, \lambda]), \qquad \psi'_{\mathrm{H},\lambda}(u) = \mathbb{I}\{|u| \leq \lambda\}, \tag{217}$$

$$\psi_{\mathrm{T},\lambda}(u) = u \left(1 - \frac{u^2}{\lambda^2} \right)_+^2 \mathbb{I}\{|u| < \lambda\}, \quad \psi'_{\mathrm{T},\lambda}(u) = \left(1 - \frac{u^2}{\lambda^2} \right)_+ \left(1 - 5\frac{u^2}{\lambda^2} \right) \mathbb{I}\{|u| < \lambda\}. \tag{218}$$

The Huber score derivative at its two breakpoints is assigned the value one for calibration, matching the implementation. The oracle Huber results assume no noise atom at $\pm\lambda$. The Tukey derivative is negative when $\lambda/\sqrt{5} < |u| < \lambda$. The signed Tukey derivative is required by the variance criterion.

Use the fixed multiplier grids

$$\mathcal{C}_{\mathrm{H}} = \{0.5, 0.75, 1, 1.345, 2, 3, 5, 10\}, \qquad \mathcal{C}_{\mathrm{T}} = \{2, 3, 4.685, 6, 10\}. \tag{219}$$

For each pair (f, c) with $f \in \{\mathrm{H}, \mathrm{T}\}$ and $c \in \mathcal{C}_f$, let $\lambda = c\widehat{s}$ and compute

$$\widehat{A}_{f,c} = \frac{1}{m} \sum_{i \in J} \psi'_{f,\lambda}(a_i), \qquad \widehat{B}_{f,c} = \frac{1}{m} \sum_{i \in J} \psi_{f,\lambda}(a_i)^2, \qquad \widehat{V}_{f,c} = \begin{cases} \widehat{B}_{f,c}/\widehat{A}_{f,c}^2, & \widehat{A}_{f,c} > 0, \\ +\infty, & \widehat{A}_{f,c} \leq 0. \end{cases} \tag{220}$$

Select the smallest value of $\widehat{V}_{f,c}$, resolving ties by the displayed family and grid order. If all values are infinite, return $\boldsymbol{c}$. Otherwise retain the selected family and cutoff throughout the proposal iterations. Appendix E explains the variance criterion under known labels. The recovery theorem does not assume that the calibration criterion consistently estimates the inlier variance in the two-component MLR model.

D.3 Two precisely specified majorization objectives

Huber polishing uses soft component weights, whereas Tukey polishing uses nearest-response assignments. The objectives below make each update explicit and permit a deterministic descent proof for both branches.

For a pair $\boldsymbol{v} = (\boldsymbol{v}_1^\top, \boldsymbol{v}_2^\top)^\top$, write $r_{ij}(\boldsymbol{v}) = y_i - \boldsymbol{x}_i^\top \boldsymbol{v}_j$. For Huber polishing, define

$$\ell_i^{\mathrm{H}}(\boldsymbol{v}) = -\widehat{s}^2 \log \left\{ \frac{1}{2} \sum_{j=1}^{2} \exp\left(-\frac{H_\lambda(r_{ij}(\boldsymbol{v}))}{\widehat{s}^2} \right) \right\}, \tag{221}$$

$$\mathcal{L}_{\mathrm{H}}(\boldsymbol{v}) = \min_{I \subseteq [n],\ |I|=m} \frac{1}{n} \sum_{i \in I} \ell_i^{\mathrm{H}}(\boldsymbol{v}). \tag{222}$$

The factor $1/2$ assigns equal working prior probabilities to the two regressions. The bounded-noise guarantee below does not require $\mathcal{L}_{\mathrm{H}}$ to be a correctly specified likelihood or require the true mixing probabilities to be equal. Multiplication by $n/\widehat{s}^2$ and subtraction of the constant $m \log 2$ preserve the minimizers and descent inequalities.

At a Huber proposal iterate $\boldsymbol{v}$, choose the m smallest row losses and call their index set $I(\boldsymbol{v})$. Define

$$\omega_{ij}(\boldsymbol{v}) = \frac{\exp\{-H_\lambda(r_{ij}(\boldsymbol{v}))/\widehat{s}^2\}}{\sum_{k=1}^{2} \exp\{-H_\lambda(r_{ik}(\boldsymbol{v}))/\widehat{s}^2\}}, \qquad w_{ij}^{\mathrm{H}}(\boldsymbol{v}) = \mathbb{I}\{i \in I(\boldsymbol{v})\}\, \omega_{ij}(\boldsymbol{v}) \min\left\{1, \frac{\lambda}{|r_{ij}(\boldsymbol{v})|}\right\}. \tag{223}$$

The minimum in the last expression is defined to equal one when the residual is zero.

For Tukey polishing, define the hard-assignment objective

$$\mathcal{L}_{\mathrm{T}}(\boldsymbol{v}) = \min_{I \subseteq [n],\ |I|=m} \frac{1}{n} \sum_{i \in I} \min_{j=1,2} T_\lambda(r_{ij}(\boldsymbol{v})). \tag{224}$$

Set $a_i(\boldsymbol{v}) = \arg\min_{j \in \{1,2\}} |r_{ij}(\boldsymbol{v})|$. Choose $I(\boldsymbol{v})$ as the indices of the m smallest $\min_{j \in \{1,2\}} |r_{ij}(\boldsymbol{v})|$. The retained set also minimizes the trimmed Tukey loss since T_λ is nondecreasing in absolute residual. The flat tail of T_λ permits additional minimizers but preserves the nearest-residual choice. The weights are

$$w_{ij}^{\mathrm{T}}(\boldsymbol{v}) = \mathbb{I}\{i \in I(\boldsymbol{v}),\ a_i(\boldsymbol{v}) = j\} \left(1 - \frac{r_{ij}(\boldsymbol{v})^2}{\lambda^2}\right)_+^2. \tag{225}$$

For either selected family, one proposal update solves

$$\boldsymbol{v}_j^+ \in \arg\min_{\boldsymbol{b} \in \mathbb{R}^d} \sum_{i=1}^{n} w_{ij}(\boldsymbol{v})(y_i - \boldsymbol{x}_i^\top \boldsymbol{b})^2, \qquad j = 1, 2. \tag{226}$$

Weighted least squares admits a minimizer even when its matrix is singular. A minimum-norm solution is

$$\boldsymbol{v}_j^+ = (\boldsymbol{X}^\top \boldsymbol{W}_j \boldsymbol{X})^\dagger \boldsymbol{X}^\top \boldsymbol{W}_j \boldsymbol{y}, \qquad \boldsymbol{W}_j = \mathrm{diag}(w_{1j}, \ldots, w_{nj}), \tag{227}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ and $\dagger$ denotes the Moore–Penrose inverse. Keeping the previous block is also allowed in the mathematical descent statement, since its contribution to the weighted quadratic does not change. The implemented Tukey polishing abandons the proposal and returns the preceding composite estimate if either component has fewer than d positive weights. Returning $\boldsymbol{c}$ satisfies the final acceptance test.

Starting at $\boldsymbol{v}^{(0)} = \boldsymbol{c}$, perform a specified finite number $L \geq 0$ of weighted least-squares updates and let $\boldsymbol{z} = \boldsymbol{v}^{(L)}$. A finite stopping rule based on $\|\boldsymbol{v}^{(k+1)} - \boldsymbol{v}^{(k)}\| < \epsilon_{\mathrm{r}}$, with specified $\epsilon_{\mathrm{r}} \geq 0$, may stop earlier. The final acceptance rule below is the same for either stopping convention.

Lemma D.2 (Quadratic upper models for the two response losses)**.** *For $R \in \{H_\lambda, T_\lambda\}$ and all real u, u_0,*

$$R(u) \leq R(u_0) + \frac{w_R(u_0)}{2}(u^2 - u_0^2), \qquad w_{H_\lambda}(u_0) = \min\{1, \lambda/|u_0|\}, \quad w_{T_\lambda}(u_0) = (1 - u_0^2/\lambda^2)_+^2, \tag{228}$$

with $w_{H_\lambda}(0) = 1$. Equality holds at $u = u_0$.

Lemma D.2 gives a global quadratic upper bound for both response losses, including the nonconvex Tukey loss. The bound justifies the exact weights in the response-polishing updates.

Proof. Write $R(u) = b_R(u^2)$ for $u^2 \geq 0$. Direct differentiation gives

$$b'_{H_\lambda}(v) = \frac{1}{2}\min\{1, \lambda/\sqrt{v}\}, \qquad b'_{T_\lambda}(v) = \frac{1}{2}(1 - v/\lambda^2)_+^2. \tag{229}$$

Use the right derivative $1/2$ at $v = 0$. Both derivatives are nonincreasing on $[0, \infty)$. For $v \geq v_0$,

$$b_R(v) - b_R(v_0) = \int_{v_0}^{v} b'_R(s)\,\mathrm{d}s \leq b'_R(v_0)(v - v_0). \tag{230}$$

For $v < v_0$ the same conclusion follows from

$$b_R(v) - b_R(v_0) = -\int_{v}^{v_0} b'_R(s)\,\mathrm{d}s \leq -b'_R(v_0)(v_0 - v). \tag{231}$$

Substitute $v = u^2$, $v_0 = u_0^2$, and $2b'_R(u_0^2) = w_R(u_0)$. □

Proposition D.3 (Exact descent of the response-polishing proposal)**.** *Keep λ and $\widehat{s} > 0$ fixed during response polishing. Every update* (226) *satisfies*

$$\mathcal{L}_f(\boldsymbol{v}^+) \leq \mathcal{L}_f(\boldsymbol{v}), \qquad f \in \{\mathrm{H}, \mathrm{T}\}, \tag{232}$$

for the objective of the selected family. More generally, the conclusion holds whenever the update does not increase its frozen weighted quadratic.

Proposition D.3 establishes descent of $\mathcal{L}_\mathrm{H}$ or $\mathcal{L}_\mathrm{T}$. Descent alone does not imply a global optimum, correct labels, or lower coefficient error. The acceptance test supplies the separate recovery guarantee.

Proof. Fix the current $\boldsymbol{v}$. For the Huber loss, abbreviate $\omega_{ij} = \omega_{ij}(\boldsymbol{v})$. For an arbitrary candidate pair $\boldsymbol{u}$, concavity of the logarithm gives

$$\log\left\{\frac{1}{2}\sum_{j=1}^{2} e^{-H_\lambda(r_{ij}(\boldsymbol{u}))/\widehat{s}^2}\right\} = \log\left\{\sum_{j=1}^{2}\omega_{ij}\frac{e^{-H_\lambda(r_{ij}(\boldsymbol{u}))/\widehat{s}^2}}{2\omega_{ij}}\right\} \tag{233}$$

$$\geq \sum_{j=1}^{2}\omega_{ij}\left\{-\frac{H_\lambda(r_{ij}(\boldsymbol{u}))}{\widehat{s}^2} - \log(2\omega_{ij})\right\}. \tag{234}$$

Thus

$$\ell_i^\mathrm{H}(\boldsymbol{u}) \leq \sum_{j=1}^{2}\omega_{ij}H_\lambda(r_{ij}(\boldsymbol{u})) + \widehat{s}^2\sum_{j=1}^{2}\omega_{ij}\log(2\omega_{ij}). \tag{235}$$

At $\boldsymbol{u} = \boldsymbol{v}$, every ratio inside the logarithm equals the same positive sum. The Jensen inequality is therefore an equality. Apply Lemma D.2 to each Huber term and sum over the fixed $I(\boldsymbol{v})$. Since the trimmed objective at $\boldsymbol{u}$ is no larger than the value on $I(\boldsymbol{v})$,

$$\mathcal{L}_\mathrm{H}(\boldsymbol{u}) - \mathcal{L}_\mathrm{H}(\boldsymbol{v}) \leq \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{2} w_{ij}^\mathrm{H}(\boldsymbol{v})\{r_{ij}(\boldsymbol{u})^2 - r_{ij}(\boldsymbol{v})^2\}. \tag{236}$$

The right side is nonpositive at the weighted least-squares update.

For the Tukey loss, retain both the current set and the current hard labels. The definition of the two minima gives

$$\mathcal{L}_\mathrm{T}(\boldsymbol{u}) \leq \frac{1}{n}\sum_{i\in I(\boldsymbol{v})} T_\lambda(r_{i,a_i(\boldsymbol{v})}(\boldsymbol{u})), \tag{237}$$

with equality at $\boldsymbol{u} = \boldsymbol{v}$. Lemma D.2 now yields

$$\mathcal{L}_\mathrm{T}(\boldsymbol{u}) - \mathcal{L}_\mathrm{T}(\boldsymbol{v}) \leq \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{2} w_{ij}^\mathrm{T}(\boldsymbol{v})\{r_{ij}(\boldsymbol{u})^2 - r_{ij}(\boldsymbol{v})^2\}. \tag{238}$$

Weighted least squares again makes the right side nonpositive. A retained previous block contributes zero to the corresponding change, proving the more general assertion. □

D.4 A COMPUTABLE GUARD AND ITS COMPLETE RECOVERY BOUND

The acceptance slack must be controlled in coefficient-error units. We first bound the observed residual scale, then combine the final composite model comparison with uniform local growth. The proof permits any measurable polishing proposal, including proposals selected from the same observations.

Use the last frozen absolute-composite objective, denoted here by $F = F_{T-1}$. Set

$$\boldsymbol{\theta}_{\mathrm{out}} = \begin{cases} \boldsymbol{z}, & \boldsymbol{z} \in \Omega \text{ and } F(\boldsymbol{z}) \le F(\boldsymbol{c}) + \widehat{s}, \\ \boldsymbol{c}, & \text{otherwise.} \end{cases} \tag{239}$$

The acceptance test uses observed responses and the stored final model. Neither true labels nor the bound σ is required. Recomputing the weights at $\boldsymbol{c}$ would define a different objective and would not give the same comparison below.

Lemma D.4 (An observable residual-scale upper bound)**.** *Suppose $\bar{\epsilon} \le 1/4$ and $\Lambda > 0$. For any pair $\boldsymbol{c}$ with aligned error $\mathcal{E}_c = \|\boldsymbol{c} - \boldsymbol{\theta}^*\|$, the scale in* (214) *satisfies*

$$\widehat{s} \le \frac{\sigma + 2\sqrt{\Lambda}\mathcal{E}_c}{q} + \upsilon. \tag{240}$$

On the clean-sample event of Lemma C.2, the bound implies

$$\widehat{s} \le \frac{\sigma + 4\mathcal{E}_c}{q} + \upsilon. \tag{241}$$

Lemma D.4 bounds the acceptance slack without a supplied noise scale. The result controls a residual statistic from above and does not assert noise-scale consistency.

Proof. Let $\boldsymbol{d}_j = \boldsymbol{c}_j - \boldsymbol{\beta}_j^*$. On an uncorrupted row,

$$a_i \le |\xi_i - \boldsymbol{x}_i^\top \boldsymbol{d}_{z_i}| \le \sigma + |\boldsymbol{x}_i^\top \boldsymbol{d}_{z_i}|. \tag{242}$$

First assume $\mathcal{E}_c > 0$. The sum of the squared projections is bounded by

$$\sum_{i=1}^{n} |\boldsymbol{x}_i^\top \boldsymbol{d}_{z_i}|^2 \le \sum_{i=1}^{n}\sum_{j=1}^{2} |\boldsymbol{x}_i^\top \boldsymbol{d}_j|^2 \le n\Lambda(\|\boldsymbol{d}_1\|^2 + \|\boldsymbol{d}_2\|^2) = n\Lambda\mathcal{E}_c^2. \tag{243}$$

Every index with $|\boldsymbol{x}_i^\top \boldsymbol{d}_{z_i}| > 2\sqrt{\Lambda}\mathcal{E}_c$ contributes strictly more than $4\Lambda\mathcal{E}_c^2$. Consequently the number of such indices is at most $\lceil n/4 \rceil - 1$. There are at most $\lfloor n/4 \rfloor$ replaced rows. The number of observed residuals strictly above $\sigma + 2\sqrt{\Lambda}\mathcal{E}_c$ is therefore at most

$$\lceil n/4 \rceil - 1 + \lfloor n/4 \rfloor \le n - 1 - \lceil (n-1)/2 \rceil. \tag{244}$$

Inequality (244) holds for each of $n = 4k, 4k+1, 4k+2, 4k+3$. The right side counts the entries above the upper-median rank $1 + \lceil (n-1)/2 \rceil$. Thus the full-sample upper median is at most $\sigma + 2\sqrt{\Lambda}\mathcal{E}_c$. If $\mathcal{E}_c = 0$, every uncorrupted row has residual at most σ, and the same conclusion follows since $n - \lfloor n/4 \rfloor$ is at least the upper-median rank.

The retained residuals are the first m order statistics of the full sample. Their upper-median rank is $1 + \lceil (m-1)/2 \rceil \le 1 + \lceil (n-1)/2 \rceil$. Their median in the stated averaging convention cannot be larger. Finally, $\max(a, b) \le a + b$ for $a, b \ge 0$ proves (240). On the event $\Lambda \le 2$, use $2\sqrt{\Lambda} \le 2\sqrt{2} \le 4$ to obtain (241). □

Proof of Corollary D.1. Equation (196), with $t = T - 1$, $\boldsymbol{c} = \boldsymbol{\theta}_T$, and $\kappa = 0$, gives

$$F(\boldsymbol{c}) - F(\boldsymbol{\theta}^*) \le \ell_{T-1}\mathcal{E}_{T-1}^2 + \eta_{T-1}. \tag{245}$$

Equation (145), at $\boldsymbol{\theta} = \boldsymbol{c}$ and $\kappa = 0$, gives

$$\mu\mathcal{E}_c - 6\sigma \le F(\boldsymbol{c}) - F(\boldsymbol{\theta}^*) \le \ell_{T-1}\mathcal{E}_{T-1}^2 + \eta_{T-1}, \qquad \mathcal{E}_c \le P_T/\mu. \tag{246}$$

The accepted and rejected cases of the guard both satisfy

$$\boldsymbol{\theta}_{\mathrm{out}} \in \Omega, \qquad F(\boldsymbol{\theta}_{\mathrm{out}}) \le F(\boldsymbol{c}) + \widehat{s}. \tag{247}$$

Applying (145) at $\boldsymbol{\theta}_{\mathrm{out}}$, followed by (245), (241), and (246), yields

$$\begin{aligned}
\mu\|\boldsymbol{\theta}_{\mathrm{out}} - \boldsymbol{\theta}^*\| &\le 6\sigma + F(\boldsymbol{\theta}_{\mathrm{out}}) - F(\boldsymbol{\theta}^*) && (248)\\
&\le P_T + \widehat{s} && (249)\\
&\le P_T + \frac{\sigma + 4\mathcal{E}_c}{q} + \upsilon && (250)\\
&\le \left(1 + \frac{4}{q\mu}\right) P_T + \frac{\sigma}{q} + \upsilon. && (251)
\end{aligned}$$

Divide by μ. Every inequality holds on the original uniform event, hence no independence assumption about calibration or proposal selection is needed. □

Corollary D.5 (Noiseless preservation and a bounded-noise floor)**.** *Under Corollary D.1,*

$$\|\boldsymbol{\theta}_{\mathrm{out}} - \boldsymbol{\theta}^*\| \le \frac{C_{\mathrm{g}}\overline{L}}{\mu}\mathcal{E}_{T-1}^2 + \frac{6C_{\mathrm{g}} + q^{-1}}{\mu}\sigma + \frac{C_{\mathrm{g}}}{\mu}\eta_{T-1} + \frac{\upsilon}{\mu}. \tag{252}$$

If $\sigma = \upsilon = 0$ and the composite objective gaps follow the schedule in Theorem C.1, the right side tends to zero at its doubly exponential composite error rate. Under the same gap schedule with nonzero noise, the bound is $O(\sigma + \upsilon)$ after the optimization term becomes negligible, with constants depending on p_0 and a.

Corollary D.5 separates the zero-noise limit from the positive numerical floor. Exact recovery also requires objective gaps tending to zero according to the stated schedule.

Proof. Substitute (211) into (212) and use $\ell_{T-1} \le \overline{L}$. For $\sigma = 0$, Theorem C.1 gives $\mathcal{E}_{T-1} \le r2^{-2^{T-1}}$ and $\eta_{T-1} \le (\mu r/2)2^{-2^T}$. Squaring the first bound makes both optimization contributions proportional to 2^{-2^T}. For $\sigma > 0$, the composite recovery theorem gives $\limsup_T \mathcal{E}_{T-1} \le 12\sigma/\mu$ and $6\sigma/\mu \le r/16$. Since $\overline{L}/\mu = 1/(16r)$,

$$\limsup_T \frac{\overline{L}}{\mu}\mathcal{E}_{T-1}^2 \le \frac{9\sigma^2}{r\mu^2} \le \frac{3\sigma}{32\mu}. \tag{253}$$

Thus (252) yields the explicit bound

$$\limsup_T \|\boldsymbol{\theta}_{\mathrm{out}} - \boldsymbol{\theta}^*\| \le \frac{(195/32)C_{\mathrm{g}} + q^{-1}}{\mu}\sigma + \frac{\upsilon}{\mu}, \tag{254}$$

which proves the claimed floor. □

Proof of Theorem 4.2. Work on the event in Theorem 4.1. Its proof represents the observed data as bounded perturbations of size u plus at most $\lfloor n\bar{\epsilon} \rfloor$ response outliers. Put $\boldsymbol{c} = \boldsymbol{\theta}_T$ and retain the final frozen objective $F = F_{T-1}$. Equation (196), applied at $t = T - 1$, gives

$$F(\boldsymbol{c}) - F(\boldsymbol{\theta}^*) \le \ell_{T-1}\mathcal{E}_{T-1}^2 + \eta_{T-1}. \tag{255}$$

Uniform local growth with effective noise bound u then gives

$$\mu\mathcal{E}(\boldsymbol{c}) \le 6u + F(\boldsymbol{c}) - F(\boldsymbol{\theta}^*) \le P_T. \tag{256}$$

Every polishing-stage early return outputs $\boldsymbol{c}$. The bound $\mathcal{E}(\boldsymbol{c}) \le P_T/\mu$ already implies (27), since $C_{\mathrm{g}} \ge 1$. For the remaining branch, Lemma D.4 applies to the same effective replacement set and yields

$$\widehat{s} \le \frac{u + 4\mathcal{E}(\boldsymbol{c})}{q} + \upsilon. \tag{257}$$

The final acceptance test gives

$$\boldsymbol{\theta}_{\mathrm{out}} \in \Omega, \qquad F(\boldsymbol{\theta}_{\mathrm{out}}) \le F(\boldsymbol{c}) + \widehat{s}. \tag{258}$$

Apply local growth and Lemma C.11 to obtain

$$\mu\mathcal{E}(\boldsymbol{\theta}_{\mathrm{out}}) \le 6u + F(\boldsymbol{\theta}_{\mathrm{out}}) - F(\boldsymbol{\theta}^*) \tag{259}$$
$$\le P_T + \widehat{s} \tag{260}$$
$$\le P_T + \frac{u + 4\mathcal{E}(\boldsymbol{c})}{q} + \upsilon \tag{261}$$
$$\le \left(1 + \frac{4}{q\mu}\right) P_T + \frac{u}{q} + \upsilon. \tag{262}$$

Division by μ proves (27). The argument is deterministic on the event already used for Theorem 4.1. Response polishing and proposal selection require no independent sample. □

An explicit finite-iteration bound. Under the objective-gap schedule in Theorem 4.1, the complete estimator also satisfies

$$\mathcal{E}(\boldsymbol{\theta}_{\mathrm{out}}) \le \frac{9C_{\mathrm{g}}}{16} r 2^{-2^T} + \frac{(219/32)C_{\mathrm{g}} + q^{-1}}{\mu} u + \frac{\upsilon}{\mu}. \tag{263}$$

The first term gives a doubly exponential optimization envelope. The remaining terms record the independent-error threshold and the numerical scale floor.

Derivation of (263). To prove (263) under the stated objective-gap schedule, put

$$b = \frac{6u}{\mu}, \qquad v_t = r2^{-2^t}. \tag{264}$$

The construction gives $b \le r/16$, $v_t \le r/2$, $v_T = v_{T-1}^2/r$, and $\overline{L}/\mu = 1/(16r)$. The composite error bound and objective-gap schedule imply

$$\frac{P_T}{\mu} \le \frac{(v_{T-1} + 2b)^2}{16r} + b + \frac{v_T}{2} \tag{265}$$
$$= \frac{9v_T}{16} + \frac{v_{T-1}b}{4r} + \frac{b^2}{4r} + b \tag{266}$$
$$\le \frac{9v_T}{16} + \frac{b}{8} + \frac{b}{64} + b \tag{267}$$
$$= \frac{9v_T}{16} + \frac{73b}{64} = \frac{9v_T}{16} + \frac{219u}{32\mu}. \tag{268}$$

Substitute (268) into (27) to obtain (263). If $u = \upsilon = 0$, only the term proportional to 2^{-2^T} remains. □

D.5 Explicit Gaussian and Laplace specializations

The next corollary states the noise scale, available trimming budget, and final recovery bound in the same units. The Gaussian conclusion also applies to independent errors with the stated sub-Gaussian tail bound. The Laplace conclusion uses its exact tail probability, avoiding the unnecessary factor two in a generic sub-exponential envelope.

Corollary D.6 (Explicit tail thresholds and full-estimator bounds)**.** *Use the algorithmic parameters and constants of Theorem 4.1, including its grid-spacing restriction. Let $T \ge 1$, let $\delta \in (0, 1/2)$, let $0 \le \epsilon \le \bar{\epsilon} \le \epsilon_0$, and suppose*

$$n \ge C\{d + \log(1/\delta)\}\log(en/\delta), \qquad \eta_t \le \frac{\mu r}{2} 2^{-2^{t+1}} \quad (0 \le t < T). \tag{269}$$

Assume the independent-error model of Section 2 and either

$$\begin{array}{ll} \text{(G)} & \mathbb{P}(|\xi_i| > t) \le 2e^{-t^2/(2v^2)} \quad (t \ge 0, v > 0), \\ \text{(L)} & \xi_i \sim \mathrm{Laplace}(0, b) \qquad\qquad\quad (b > 0), \end{array} \tag{270}$$

for every $i \in [n]$. In particular, (G) *includes $\xi_i \sim N(0, v^2)$. Define*

$$C_{\mathrm{opt}} = \frac{9C_{\mathrm{g}}}{16}, \qquad K_{\mathrm{out}} = \frac{(219/32)C_{\mathrm{g}} + q^{-1}}{\mu}, \qquad C_{\mathrm{g}} = 1 + \frac{4}{q\mu}, \qquad q = \Phi^{-1}(3/4). \tag{271}$$

The following two alternatives each hold with probability at least $1-\delta$, simultaneously for all allowed response outliers.

(i) Additional outlier budget. *Let $\omega = \bar{\epsilon} - \epsilon > 0$ and suppose*

$$n \geq \frac{2\log(2/\delta)}{\omega^2}. \tag{272}$$

Choose the analysis threshold and impose the corresponding separation condition:

$$\begin{aligned} &\text{(G)} \quad u_{\mathrm{G}} = v\sqrt{2\log(4/\omega)}, && \frac{v}{\Delta} \leq \frac{s_0}{\sqrt{2\log(4/\omega)}}, \\ &\text{(L)} \quad u_{\mathrm{L}} = b\log(2/\omega), && \frac{b}{\Delta} \leq \frac{s_0}{\log(2/\omega)}. \end{aligned} \tag{273}$$

The output of the complete SARP-P *procedure satisfies, respectively,*

$$\mathcal{E}(\boldsymbol{\theta}_{\mathrm{out}}) \leq C_{\mathrm{opt}} r 2^{-2^T} + \frac{v}{\mu} + \begin{cases} K_{\mathrm{out}} v\sqrt{2\log(4/\omega)}, & \text{(G)}, \\ K_{\mathrm{out}} b\log(2/\omega), & \text{(L)}. \end{cases} \tag{274}$$

(ii) No additional outlier budget. *The choice $\bar{\epsilon} = \epsilon$ is allowed. Choose*

$$\begin{aligned} &\text{(G)} \quad \widetilde{u}_{\mathrm{G}} = v\sqrt{2\log(4n/\delta)}, && \frac{v}{\Delta} \leq \frac{s_0}{\sqrt{2\log(4n/\delta)}}, \\ &\text{(L)} \quad \widetilde{u}_{\mathrm{L}} = b\log(2n/\delta), && \frac{b}{\Delta} \leq \frac{s_0}{\log(2n/\delta)}. \end{aligned} \tag{275}$$

The full output satisfies

$$\mathcal{E}(\boldsymbol{\theta}_{\mathrm{out}}) \leq C_{\mathrm{opt}} r 2^{-2^T} + \frac{v}{\mu} + \begin{cases} K_{\mathrm{out}} v\sqrt{2\log(4n/\delta)}, & \text{(G)}, \\ K_{\mathrm{out}} b\log(2n/\delta), & \text{(L)}. \end{cases} \tag{276}$$

Corollary D.6 quantifies the tradeoff between the noise threshold and the additional trimming budget. The Gaussian parameter v is its standard deviation, while Laplace scale b corresponds to standard deviation $\sqrt{2}b$. With fixed positive ω, the noise contribution in (274) does not vanish as n increases. Without spare budget, the maximum-noise threshold increases with n and requires a correspondingly stronger separation-to-noise ratio. Neither alternative establishes a conventional stochastic estimation rate such as $v\sqrt{d/n}$.

Proof. The Gaussian tail bound follows from Lemma A.4. A centered Gaussian with standard deviation v therefore satisfies (G). The Laplace density is $e^{-|x|/b}/(2b)$, giving for every $t \geq 0$

$$\mathbb{P}(|\xi_i| > t) = 2\int_t^\infty \frac{e^{-x/b}}{2b}\,\mathrm{d}x = e^{-t/b}. \tag{277}$$

Thus admissible envelopes in Theorem 4.1 are $p_{\mathrm{tail}}(t) = \min\{1, 2e^{-t^2/(2v^2)}\}$ in (G) and $p_{\mathrm{tail}}(t) = e^{-t/b}$ in (L).

For part (i), (272) gives

$$h_{n,\delta} := \sqrt{\frac{\log(2/\delta)}{2n}} \leq \frac{\omega}{2}. \tag{278}$$

The thresholds in (273) obey

$$\begin{aligned} &\text{(G)} \quad p_{\mathrm{tail}}(u_{\mathrm{G}}) \leq 2e^{-\log(4/\omega)} = \frac{\omega}{2}, \\ &\text{(L)} \quad p_{\mathrm{tail}}(u_{\mathrm{L}}) = e^{-\log(2/\omega)} = \frac{\omega}{2}. \end{aligned} \tag{279}$$

For the applicable threshold u, let $A_u = \{i : |\xi_i| > u\}$. The independent indicators of A_u satisfy the one-sided Hoeffding bound

$$\mathbb{P}\{|A_u|/n > p_{\mathrm{tail}}(u) + h_{n,\delta}\} \le e^{-2nh_{n,\delta}^2} = \delta/2. \tag{280}$$

On the complementary event, every allowed original outlier set O satisfies

$$|O \cup A_u| \le n\{\epsilon + p_{\mathrm{tail}}(u) + h_{n,\delta}\} \le n(\epsilon + \omega) = n\bar{\epsilon}. \tag{281}$$

The cardinality is an integer and is therefore at most $\lfloor n\bar{\epsilon} \rfloor$. Outside $O \cup A_u$ the error magnitude is at most u. The scale inequalities in (273) are exactly $u \le s_0\Delta$.

For part (ii), the thresholds in (275) instead give

$$\begin{aligned} &\text{(G)} \quad np_{\mathrm{tail}}(\widetilde{u}_{\mathrm{G}}) \le 2ne^{-\log(4n/\delta)} = \delta/2, \\ &\text{(L)} \quad np_{\mathrm{tail}}(\widetilde{u}_{\mathrm{L}}) = ne^{-\log(2n/\delta)} = \delta/2. \end{aligned} \tag{282}$$

The union bound implies that all original errors are bounded by the applicable threshold with failure probability at most $\delta/2$. Only the original replacement set is then needed. The separation conditions again ensure $u \le s_0\Delta$.

For either part, intersect the noise event with the uniform clean-design event at failure probability $\delta/2$ used in Theorem 4.1. The total failure probability is at most δ. The sample constant in (269) already incorporates this allocation. Theorem 4.2 and (263) give on this intersection

$$\mathcal{E}(\boldsymbol{\theta}_{\mathrm{out}}) \le C_{\mathrm{opt}} r 2^{-2^T} + K_{\mathrm{out}} u + \frac{\upsilon}{\mu}. \tag{283}$$

Substituting (273) into (283) proves (274). Substituting (275) into (283) proves (276). Uniformity over all replacement sets and response values follows from the same clean-design event. □

Using the entire available tail budget. The half-budget choice in part (i) gives a simple sufficient sample condition. A sharper direct specialization is available whenever $\omega > h_{n,\delta}$. Define

$$\alpha_{\mathrm{tail}} = \omega - h_{n,\delta} \in (0,1), \qquad u_{\mathrm{G}}^{\mathrm{sharp}} = v\sqrt{2\log(2/\alpha_{\mathrm{tail}})}, \qquad u_{\mathrm{L}}^{\mathrm{sharp}} = b\log(1/\alpha_{\mathrm{tail}}). \tag{284}$$

The corresponding envelope is at most α_{tail}, and hence $\epsilon + p_{\mathrm{tail}}(u) + h_{n,\delta} \le \bar{\epsilon}$. Provided the chosen threshold is at most $s_0\Delta$, the output bound (283) applies with this threshold. This refinement requires no new probability argument.

E Oracle Efficiency and the Limits of Calibration

Response polishing selects a loss family and cutoff through the empirical ratio in (220). The calculations below derive the corresponding population variance ratio. Known component labels reduce two-component MLR to separate regressions, allowing an explicit covariance calculation and comparison of noise distributions. A final concentration result concerns a fixed score family evaluated on independent true errors. None of the oracle results establishes optimal tuning of fitted nearest-component residuals, adaptive trimming, or response outliers. The recovery guarantee for the full SARP-P procedure comes from Theorem 4.2.

E.1 A general score calculation with an explicit local-root condition

To isolate the effect of a response score, suppose that the component labels are known and consider the resulting single regression. Theorem E.1 derives the covariance of a consistent estimating-equation root for a fixed cutoff. The derivation explains the ratio used for calibration without assuming that the actual polishing iterates satisfy the oracle conditions.

Theorem E.1 (Oracle-label covariance for Huber and Tukey scores)**.** *Fix d and $\lambda > 0$. Observe N independent correctly labeled pairs*

$$\boldsymbol{x}_i \sim N(\mathbf{0}, \boldsymbol{I}_d), \qquad y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}^* + \xi_i, \qquad \boldsymbol{x}_i \perp \xi_i, \tag{285}$$

where the noises are independent identically distributed and symmetric. Choose one of the two scores $\psi = \psi_{f,\lambda}$ in (218)*. Define*

$$A = \mathbb{E}\psi'(\xi) > 0, \qquad B = \mathbb{E}\psi(\xi)^2. \tag{286}$$

For the Huber loss, assume $\mathbb{P}(|\xi| = \lambda) = 0$. Suppose $\widehat{\boldsymbol{\beta}}_N$ is a sequence of consistent score roots, meaning

$$\widehat{\boldsymbol{\beta}}_N \xrightarrow{\mathrm{p}} \boldsymbol{\beta}^*, \qquad \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \psi(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_N) = \mathbf{0}. \tag{287}$$

Then

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^*) \Rightarrow N\left(\mathbf{0}, \frac{B}{A^2}\boldsymbol{I}_d\right). \tag{288}$$

For the Huber loss, consistent roots are supplied by every sequence of global empirical minimizers if additionally $\mathbb{E}|\xi| < \infty$. For the Tukey loss, consistency is an explicit assumption about the chosen local root and is not implied by Proposition D.3.

The ratio B/A^2 motivates comparing score families and cutoffs. The covariance formula in Theorem E.1 requires correct labels and a consistent estimating-equation root.

Proof. Put $\widehat{\boldsymbol{d}} = \widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^*$. Both scores are absolutely continuous and satisfy

$$\psi(u - v) - \psi(u) = -v \int_0^1 \psi'(u - tv)\,\mathrm{d}t. \tag{289}$$

Consequently the root equation is exactly

$$\begin{aligned} &\boldsymbol{D}_N(\widehat{\boldsymbol{d}})\sqrt{N}\widehat{\boldsymbol{d}} = \boldsymbol{S}_N, && \boldsymbol{S}_N = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \boldsymbol{x}_i \psi(\xi_i), \\ &\boldsymbol{D}_N(\boldsymbol{d}) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top \int_0^1 \psi'(\xi_i - t\boldsymbol{x}_i^\top \boldsymbol{d})\,\mathrm{d}t. \end{aligned} \tag{290}$$

The scores and their derivatives are bounded. Independence, finite Gaussian fourth moments, and the entrywise weak law give $\boldsymbol{D}_N(\mathbf{0}) \xrightarrow{\mathrm{p}} A\boldsymbol{I}_d$.

For the Huber loss, for every $\|\boldsymbol{d}\| \le r$,

$$\|\boldsymbol{D}_N(\boldsymbol{d}) - \boldsymbol{D}_N(\mathbf{0})\|_{\mathrm{op}} \le \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{x}_i\|^2 \mathbb{I}\{\big||\xi_i| - \lambda\big| \le r\|\boldsymbol{x}_i\|\}. \tag{291}$$

As $r \downarrow 0$, the expectation of the right side of (291) tends to zero by dominated convergence (Folland, 1999, Theorem 2.24) and the no-atom assumption. For the Tukey loss, its derivative is continuous and bounded, and the corresponding bound is

$$\sup_{\|\boldsymbol{d}\| \le r} \|\boldsymbol{D}_N(\boldsymbol{d}) - \boldsymbol{D}_N(\mathbf{0})\|_{\mathrm{op}} \le \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{x}_i\|^2 \sup_{|v| \le r\|\boldsymbol{x}_i\|} |\psi'(\xi_i - v) - \psi'(\xi_i)|. \tag{292}$$

The expectation of the envelope in (292) also tends to zero as $r \downarrow 0$, with integrable envelope $2\sup_u |\psi'(u)|\,\|\boldsymbol{x}_i\|^2$. The Markov inequality in either case gives, for every $\epsilon > 0$,

$$\lim_{r \downarrow 0} \sup_N \mathbb{P}\left\{\sup_{\|\boldsymbol{d}\| \le r} \|\boldsymbol{D}_N(\boldsymbol{d}) - \boldsymbol{D}_N(\mathbf{0})\|_{\mathrm{op}} > \epsilon\right\} = 0. \tag{293}$$

Let W_r denote the nonnegative single-observation envelope in (291) or (292). Equations (291) and (292), with dominated convergence, give $\mathbb{E}W_r \to 0$. For every $r, \epsilon > 0$,

$$\mathbb{P}\{\|\boldsymbol{D}_N(\widehat{\boldsymbol{d}}) - A\boldsymbol{I}_d\|_{\mathrm{op}} > 2\epsilon\} \tag{294}$$

$$\le \mathbb{P}\{\|\widehat{\boldsymbol{d}}\| > r\} + \mathbb{P}\{\|\boldsymbol{D}_N(\mathbf{0}) - A\boldsymbol{I}_d\|_{\mathrm{op}} > \epsilon\} + \frac{\mathbb{E}W_r}{\epsilon}. \tag{295}$$

Take $N \to \infty$ for fixed r, then $r \downarrow 0$. Consistency and the weak law make the first two terms vanish, and the envelope bound makes the last term vanish. Hence

$$\boldsymbol{D}_N(\widehat{\boldsymbol{d}}) \xrightarrow{\mathrm{p}} A\boldsymbol{I}_d. \tag{296}$$

Conditionally on $\xi_1, \ldots, \xi_N$, $\boldsymbol{S}_N$ is exactly centered Gaussian with covariance $B_N \boldsymbol{I}_d$, where $B_N = N^{-1} \sum_{i=1}^N \psi(\xi_i)^2$. The weak law gives $B_N \xrightarrow{\mathrm{p}} B$. Equivalently, the marginal distribution of $\boldsymbol{S}_N$ equals that of $\sqrt{B_N}\boldsymbol{Z}$ on an auxiliary space with an independent $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I}_d)$. The Slutsky lemma (Van der Vaart, 2000, Lemma 2.8) implies $\boldsymbol{S}_N \Rightarrow N(\boldsymbol{0}, B\boldsymbol{I}_d)$. Equations (290) and (296), with $A > 0$, give (288) by the same matrix Slutsky lemma.

We next verify consistency of global empirical Huber minimizers. Define the empirical contrast

$$R_N(\boldsymbol{d}) = \frac{1}{N} \sum_{i=1}^N \{H_\lambda(\xi_i - \boldsymbol{x}_i^\top \boldsymbol{d}) - H_\lambda(\xi_i)\}, \qquad R(\boldsymbol{d}) = \mathbb{E} R_N(\boldsymbol{d}). \tag{297}$$

Each contrast is bounded in absolute value by $\lambda |\boldsymbol{x}_i^\top \boldsymbol{d}|$. The empirical and population Lipschitz constants of R_N and R on a compact ball are at most $\lambda N^{-1} \sum_{i=1}^N \|\boldsymbol{x}_i\|$ and $\lambda \mathbb{E}\|\boldsymbol{x}\|$. Finite ball nets and the weak law therefore give

$$\sup_{\|\boldsymbol{d}\| \le r} |R_N(\boldsymbol{d}) - R(\boldsymbol{d})| \xrightarrow{\mathrm{p}} 0 \qquad (r < \infty). \tag{298}$$

More explicitly, an η-net bounds the supremum by the maximum deviation on its finite points plus

$$\lambda \eta \left(\frac{1}{N} \sum_{i=1}^N \|\boldsymbol{x}_i\| + \mathbb{E}\|\boldsymbol{x}\| \right). \tag{299}$$

Let $N \to \infty$ at fixed η, then let $\eta \downarrow 0$.

The population gradient at zero vanishes since $\mathbb{E}[\boldsymbol{x}\psi(\xi)] = \boldsymbol{0}$. The same dominated-convergence bound as (291) gives a radius $r_0 > 0$ such that

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top \mathbb{I}\{|\xi - \boldsymbol{x}^\top \boldsymbol{d}| < \lambda\}] \succeq (A/2)\boldsymbol{I}_d \quad (\|\boldsymbol{d}\| \le r_0). \tag{300}$$

Integrating the population second derivative along the line segment from zero to $\boldsymbol{d}$ yields

$$R(\boldsymbol{d}) = \int_0^1 (1-t)\, \boldsymbol{d}^\top \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top \mathbb{I}\{|\xi - t\boldsymbol{x}^\top \boldsymbol{d}| < \lambda\}] \boldsymbol{d}\, \mathrm{d}t \ge \frac{A}{4} \|\boldsymbol{d}\|^2. \tag{301}$$

For any fixed $0 < r \le r_0$, (298) implies that, with probability tending to one, $R_N(\boldsymbol{d}) > 0$ for all $\|\boldsymbol{d}\| = r$. The empirical Huber objective is convex. If a minimizer lay outside the radius-r sphere, the line segment from the minimizer to zero would meet the sphere at a point with contrast at most zero, a contradiction.

Such a minimizer exists almost surely for $N \ge d$. The Gaussian design has full column rank, hence compactness of the unit sphere gives $c_N = \inf_{\|\boldsymbol{u}\|=1} \sum_{i=1}^N |\boldsymbol{x}_i^\top \boldsymbol{u}| > 0$. Also

$$\sum_{i=1}^N H_\lambda(y_i - \boldsymbol{x}_i^\top \boldsymbol{b}) \ge \lambda c_N \|\boldsymbol{b}\| - \lambda \sum_{i=1}^N |y_i| - N\lambda^2/2 \longrightarrow \infty \quad (\|\boldsymbol{b}\| \to \infty). \tag{302}$$

The continuous objective is therefore coercive and attains its minimum. Letting r decrease proves consistency of every sequence of global minimizers. The differentiable Huber objective has zero gradient at every such minimizer, giving the required score root. □

E.2 Noise families with different efficiency tradeoffs

The oracle variance can favor different cutoffs for different error distributions. Gaussian, Laplace, and Gaussian scale-mixture calculations quantify the distribution-dependent variance ratios for Huber scores and motivate the noise distributions used in the experiments.

Corollary E.2 (Huber covariance for Gaussian and Laplace noise)**.** *In the Huber case of Theorem E.1,*

$$V_\lambda = \frac{\mathbb{E}\min(\xi^2, \lambda^2)}{\mathbb{P}(|\xi| < \lambda)^2}. \tag{303}$$

For $\xi \sim N(0, s^2)$, let $t = \lambda/s$, $a(t) = 2\Phi(t) - 1$, and let φ be the standard normal density. Then

$$\frac{V_\lambda}{s^2} = \frac{a(t) - 2t\varphi(t) + t^2\{1 - a(t)\}}{a(t)^2}, \qquad \lim_{t\to\infty} \frac{V_\lambda}{s^2} = 1, \qquad \lim_{t \downarrow 0} \frac{V_\lambda}{s^2} = \frac{\pi}{2}. \tag{304}$$

For Laplace noise with density $e^{-|u|/b}/(2b)$, let $t = \lambda/b$. Then

$$\frac{V_\lambda}{b^2} = \frac{2\{1-(1+t)e^{-t}\}}{(1-e^{-t})^2}, \qquad \lim_{t\downarrow 0} V_\lambda = b^2, \qquad \mathbb{E}\xi^2 = 2b^2. \tag{305}$$

Small cutoffs sacrifice oracle efficiency under Gaussian noise but can approach half the least-squares variance under Laplace noise. The Gaussian–Laplace contrast motivates adaptive calibration and noise-family experiments, rather than a claim that one fixed robust cutoff is best in every setting.

Proof. For the Huber loss, $\psi(\xi)^2 = \min(\xi^2, \lambda^2)$ and $\psi'(\xi) = \mathbb{I}\{|\xi| < \lambda\}$, proving (303). For $Z \sim N(0,1)$, integration by parts gives

$$\mathbb{E}[Z^2 \mathbb{I}\{|Z| \le t\}] = 2\int_0^t z^2 \varphi(z)\,\mathrm{d}z = -2t\varphi(t) + 2\Phi(t) - 1. \tag{306}$$

Adding $t^2\mathbb{P}(|Z| > t)$ proves the Gaussian formula. As $t \to \infty$, the truncated second moment tends to one and $t^2\mathbb{P}(|Z| > t) \le 2t^2 e^{-t^2/2} \to 0$. As $t \downarrow 0$, $a(t)/t \to 2\varphi(0)$ and $\mathbb{E}\min(Z^2, t^2)/t^2 \to 1$ by bounded convergence. Their ratio tends to $1/(4\varphi(0)^2) = \pi/2$.

For Laplace noise, $\mathbb{P}(|\xi| > u) = e^{-u/b}$ for $u \ge 0$. The tail-integral identity gives

$$\mathbb{E}\min(\xi^2, \lambda^2) = 2\int_0^\lambda u e^{-u/b}\,\mathrm{d}u = 2b^2\{1-(1+t)e^{-t}\}, \qquad \mathbb{P}(|\xi| < \lambda) = 1 - e^{-t}. \tag{307}$$

Since $1 - e^{-t} = t + O(t^2)$ and $1-(1+t)e^{-t} = t^2/2 + O(t^3)$, the ratio tends to b^2. Letting $\lambda \to \infty$ in the second-moment integral gives $\mathbb{E}\xi^2 = 2b^2$. □

Corollary E.3 (A Gaussian scale-mixture comparison)**.** *For noise*

$$\xi \sim (1-\omega)N(0, s^2) + \omega N(0, M^2 s^2), \qquad 0 < \omega < 1, \quad M > 1, \tag{308}$$

the Huber covariance in (303) *satisfies*

$$\lim_{\lambda\downarrow 0} V_\lambda = \frac{\pi s^2}{2(1-\omega+\omega/M)^2}, \qquad \mathrm{Var}(\xi) = (1-\omega+\omega M^2)s^2. \tag{309}$$

The symmetric Gaussian scale mixture can have substantially different robust and least-squares variances. The calculation motivates studying heterogeneous error scales within the independent-tail model.

Proof. The Gaussian scale-mixture density at zero is $f(0) = (1-\omega+\omega/M)/(\sqrt{2\pi}s)$. Continuity gives

$$\frac{\mathbb{P}(|\xi| < \lambda)}{2\lambda} = \frac{1}{2\lambda}\int_{-\lambda}^{\lambda} f(u)\,\mathrm{d}u \longrightarrow f(0). \tag{310}$$

The variables $\min(\xi^2, \lambda^2)/\lambda^2$ belong to $[0,1]$ and tend to one almost surely, since the noise has no atom at zero. Bounded convergence gives $\mathbb{E}\min(\xi^2,\lambda^2)/\lambda^2 \to 1$. Hence the limiting ratio is $1/(4f(0)^2)$, which is the first expression. The second follows by conditioning on the Gaussian scale. □

E.3 A FINITE-GRID ORACLE CALIBRATION BOUND

Proposition E.4 bounds selection error for a fixed finite score family evaluated on independent true errors. The result isolates the concentration argument behind variance calibration. Fitted nearest-component residuals and data-dependent scales require additional analysis.

Proposition E.4 (Concentration of oracle variance calibration)**.** *Let $\xi_1, \dots, \xi_N$ be independent identically distributed true errors. Consider a fixed finite family of K score functions ψ_k and specified derivative functions d_k, with deterministic cutoffs. Suppose*

$$|\psi_k(u)| \le M_k, \qquad d_k(u) \in [a_k, b_k], \qquad A_k = \mathbb{E}d_k(\xi) > 0, \qquad B_k = \mathbb{E}\psi_k(\xi)^2. \tag{311}$$

Write $R_k = b_k - a_k$, $V_k = B_k/A_k^2$, and define the empirical means $\widehat{A}_k, \widehat{B}_k$ and ratios $\widehat{V}_k$ as in (220)*, using $\xi_1, \dots, \xi_N$ and all N indices. For $\delta \in (0,1)$ set*

$$t_N = \sqrt{\frac{\log(4K/\delta)}{2N}}. \tag{312}$$

If $R_k t_N \le A_k/2$ for every k, then with probability at least $1-\delta$ all empirical denominators are positive and

$$|\widehat{V}_k - V_k| \le E_k, \qquad E_k = \frac{4M_k^2 t_N}{A_k^2} + \frac{10M_k^2 R_k t_N}{A_k^3}. \tag{313}$$

Consequently, if $\widehat{k} \in \arg\min_k \widehat{V}_k$ and $E_{\max} = \max_k E_k$, then

$$V_{\widehat{k}} \le \min_k V_k + 2E_{\max}. \tag{314}$$

For the two score families above one may use $M_k = \lambda_k$, $R_k = 1$ for Huber, and $M_k = 16\lambda_k/(25\sqrt{5})$, $R_k = 9/5$ for Tukey.

Proposition E.4 quantifies calibration from independent true errors at fixed cutoffs. Fitted nearest-component residuals and an estimated scale do not meet the independent-error and fixed-cutoff assumptions. The recovery guarantee for SARP-P instead follows from Theorem 4.2.

Proof. Since $0 \le \psi_k(\xi_i)^2 \le M_k^2$, the Hoeffding inequality gives

$$\mathbb{P}\{|\widehat{B}_k - B_k| > M_k^2 t_N\} \le 2e^{-2Nt_N^2} = \frac{\delta}{2K}. \tag{315}$$

The interval width of d_k is R_k, hence

$$\mathbb{P}\{|\widehat{A}_k - A_k| > R_k t_N\} \le \frac{\delta}{2K}. \tag{316}$$

A union bound over the $2K$ deviations has failure probability at most δ. On the resulting event, $A_k/2 \le \widehat{A}_k \le 3A_k/2$. Therefore

$$\left|\frac{\widehat{B}_k}{\widehat{A}_k^2} - \frac{B_k}{A_k^2}\right| \le \frac{|\widehat{B}_k - B_k|}{\widehat{A}_k^2} + \frac{B_k|A_k - \widehat{A}_k|(A_k + \widehat{A}_k)}{A_k^2 \widehat{A}_k^2} \tag{317}$$

$$\le \frac{4M_k^2 t_N}{A_k^2} + \frac{M_k^2 R_k t_N (5A_k/2)}{A_k^2(A_k^2/4)} = E_k. \tag{318}$$

If $k_* \in \arg\min_k V_k$, then

$$V_{\widehat{k}} \le \widehat{V}_{\widehat{k}} + E_{\max} \le \widehat{V}_{k_*} + E_{\max} \le V_{k_*} + 2E_{\max}. \tag{319}$$

For the Huber loss, the score magnitude is at most λ and its derivative lies in $[0,1]$. For the Tukey loss, set $v = |u|/\lambda \in [0,1]$. Differentiating $v(1-v^2)^2$ shows its maximum occurs at $v = 1/\sqrt{5}$ and equals $16/(25\sqrt{5})$. Its derivative equals $1 - 6v^2 + 5v^4$, whose minimum on $[0,1]$ occurs at $v^2 = 3/5$ and equals $-4/5$. Its maximum is one. Outside this interval the derivative is zero. The score maximum and derivative range give the stated constants M_k and R_k. □

Why the oracle distinction matters. At the true pair, the nearest-component absolute residual of an inlier is generally

$$\min\{|\xi_i|, |\xi_i + \boldsymbol{x}_i^\top(\boldsymbol{\beta}^*_{z_i} - \boldsymbol{\beta}^*_{3-z_i})|\}, \tag{320}$$

which need not equal $|\xi_i|$. Trimming further changes the nearest-component residual distribution. Therefore the oracle covariance and grid-selection results motivate the proposal but do not establish its optimal tuning under arbitrary contamination. The acceptance test gives the finite-sample guarantee in Theorem 4.2 for the actual data-dependent SARP-P procedure.

F Experimental Settings and Additional Results

The experiments compare the full SARP-P procedure with four robust baselines and with the estimate before response polishing. We report all evaluated noise distributions, response-outlier mechanisms, numerical conventions, and variability. The implementation choices below remain distinct from the conservative sufficient constants of the recovery theorems.

F.1 STUDY DESIGN AND COMPLETE RESULTS

We focus on simultaneous inlier noise and response outliers. The study tests whether calibrated response polishing improves estimation accuracy over existing robust methods. The ablation also compares the final estimate with the output of the same normalized composite iterations. No generating noise scale, true coefficient, latent label, or replacement index is supplied to SARP-P.

Data and scope of the comparison. Set $d = 20$, $n = 1000$, $p = 1/2$, and $\epsilon = \bar{\epsilon} = 0.05$. Factor a $d \times 3$ matrix of independent standard Gaussian entries as $\boldsymbol{G} = \boldsymbol{Q}\boldsymbol{R}$, where $\boldsymbol{Q}$ has orthonormal columns and $\boldsymbol{R}$ is upper triangular. Take $\boldsymbol{u}, \boldsymbol{m}, \boldsymbol{v}$ to be the three columns of $\boldsymbol{Q}$. The coefficients are $\boldsymbol{\beta}_1^* = 2\boldsymbol{m} + \boldsymbol{u}/2$ and $\boldsymbol{\beta}_2^* = 2\boldsymbol{m} - \boldsymbol{u}/2$, with separation $\Delta = 1$. Draw clean covariates and labels as in (1). The noise standard deviation is s. We use

$$\begin{aligned} \text{Laplace: } & \xi_i \sim \text{Laplace}(0, s/\sqrt{2}), \\ \text{scale mixture: } & \xi_i \sim 0.9N(0, s^2/9) + 0.1N(0, 9s^2), \\ \text{Gaussian: } & \xi_i \sim N(0, s^2), \qquad \text{uniform: } \xi_i \sim \text{Unif}[-\sqrt{3}s, \sqrt{3}s]. \end{aligned} \tag{321}$$

All four distributions have variance s^2 and satisfy the independent-error model in Section 2. Gaussian and Gaussian scale-mixture errors are sub-Gaussian, Laplace errors have sub-exponential tails, and uniform errors are bounded. The scale mixture combines a concentrated central component with a broader secondary component. Theorem 4.1 applies when a threshold u satisfies the stated noise, separation, and contamination conditions.

Gross replacements set $y_i = 2\boldsymbol{x}_i^\top \boldsymbol{m} + 5Z_i$, with independent $Z_i \sim N(0, 1)$, on $\lfloor 0.05n \rfloor$ uniformly selected rows. Nearby replacements instead set $y_i = \boldsymbol{x}_i^\top (\boldsymbol{\beta}_1^* + 4s\boldsymbol{v})$. A leverage-selected response attack uses the rows with largest $|\boldsymbol{x}_i^\top \boldsymbol{v}|$ and sets $y_i = 2\boldsymbol{x}_i^\top \boldsymbol{m} + 5\,\text{sign}(\boldsymbol{x}_i^\top \boldsymbol{v})$. Covariates are never modified. Each of the twelve settings in Table 2 uses all thirty seeds from 10000 through 10029.

The noise mechanisms were selected after developing the algorithm. The selected mechanisms emphasize distributions where a squared-error final fit can lose efficiency. The algorithm, settings, and evaluation seeds were fixed before the reported runs. Gaussian and uniform controls were retained. The study evaluates response polishing across the selected noise distributions and retains all control settings and evaluated trials.

Four robust baseline algorithms. We compare with the published Mix-IRLS implementation and its existing outlier option (Zilber & Nadler, 2023), locally initialized ILTS (Shen & Sanghavi, 2019), Mixregt (Yao et al., 2014), and Laplace EM (Song et al., 2014). Mixregt profiles every integer degrees of freedom from 1 to 15 by observed likelihood, following its original simulation protocol. Laplace EM uses the published common-scale likelihood, with an exact weighted-LAD M step. Weighted LAD directly solves the coefficient update for the published likelihood, without the additional latent-scale IRLS representation. No response trimming or new robustness adaptation is added to Mixregt or Laplace EM.

All baselines except Mix-IRLS receive the same data-dependent segment initialization as the normalized composite iteration in SARP-P. Mix-IRLS retains its published random initialization. ILTS receives the true balanced proportion $p = 1/2$ for its subset size, whereas SARP-P uses only $p_0 = 0.2$. Thus the comparisons do not deprive ILTS or the likelihood baselines of a strong start. The differing published initialization of Mix-IRLS is disclosed, hence end-to-end differences cannot be attributed solely to final updates. The following subsections specify the numerical conventions.

Practical parameters. Use $a = 0.1$, $\kappa = 0$, and $p_0 = 0.2$. The numerical initializer floors τ at 10^{-12}, a convention distinct from the response floor υ. The scalar initializer searches $[-3\tau, 3\tau]$ with initial spacing 0.05τ and local scalar refinement. The feasible radius is 2τ. Response polishing uses the fixed family in (62), with $\upsilon = 10^{-12}$, and estimates its scale from residuals. The practical grid and radius differ from the conservative sufficient constants in Theorem 4.1, and the experiments do not verify the sufficient contamination threshold. The largest solver-reported composite model gap in normalized coordinates is 9.99×10^{-11}. All runs use the finite numerical stopping rules specified below rather than enforcing the theoretical objective-gap schedule.

Table 2: Median coefficient error $\mathcal{E}(\boldsymbol{\theta})$ in units of 10^{-3} over thirty datasets for each noise and replacement mechanism. The component separation is one, s is the inlier standard deviation, and 5% of responses are replaced. Bold indicates the smallest median in each row. All trials are included.

Noise / replacements	s	SARP-P	Mix-IRLS	Mixregt	Laplace EM	ILTS
Laplace / gross	0.01	**2.35**	2.89	2.44	2.64	2.88
Laplace / gross	0.03	**7.18**	8.81	7.31	7.89	8.85
Laplace / gross	0.1	**25.09**	31.09	25.30	28.16	32.06
Laplace / nearby	0.03	**7.22**	8.54	7.36	7.98	9.00
Scale mixture / gross	0.01	**1.06**	2.72	1.27	1.35	1.32
Scale mixture / gross	0.03	**3.23**	7.79	3.88	4.08	4.04
Scale mixture / gross	0.1	**11.06**	24.87	13.11	13.78	15.39
Scale mixture / nearby	0.03	**3.36**	5.53	3.92	4.21	4.18
Gaussian / gross	0.03	**9.34**	9.43	11.06	11.25	14.27
Gaussian / gross	0.1	33.97	**33.18**	40.20	40.19	58.61
Uniform / gross	0.03	9.33	**9.29**	16.49	15.85	23.47
Scale mixture / leverage	0.03	**3.33**	8.65	4.16	6.76	999.59

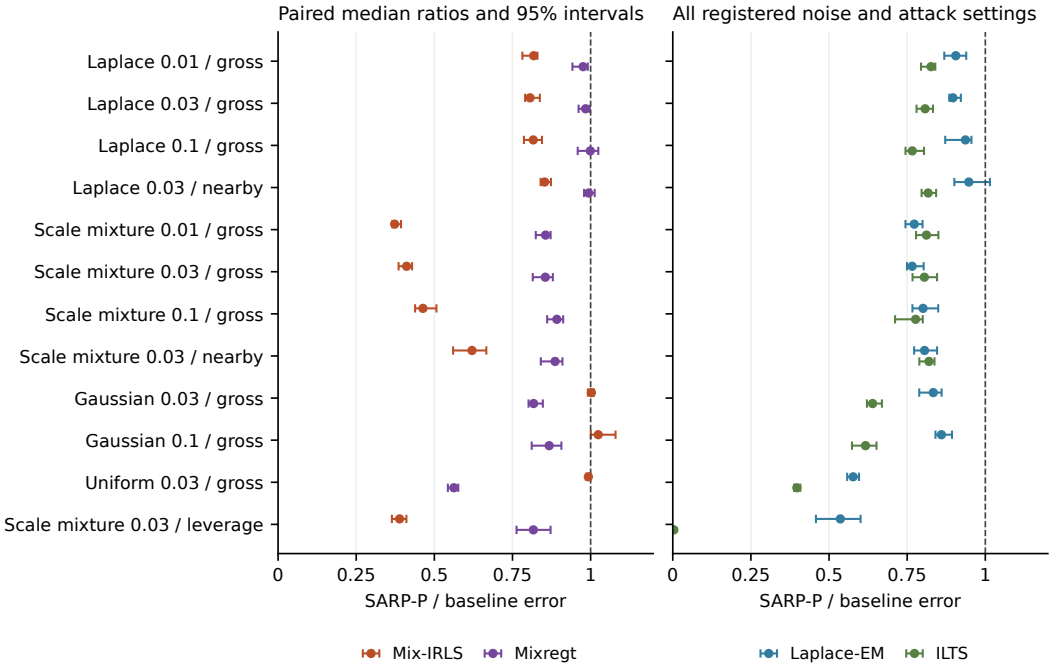

Figure 1: Median paired coefficient-error ratios comparing SARP-P with four robust baselines in all evaluated settings. Ratios below one favor SARP-P. Error bars are descriptive 95% paired-bootstrap intervals. Table 2 reports absolute errors for all four robust baselines.

Accuracy and the strongest comparisons. The results favor SARP-P most clearly for scale-mixture noise. The improvements over Mix-IRLS are also consistent for Laplace noise, while differences from Mixregt are smaller and are not decisive in every setting. Gaussian and uniform controls show that the gain depends on the noise distribution. Mix-IRLS can have lower error under Gaussian or uniform noise. The descriptive intervals in Figure 1 resample paired datasets within each setting and have no multiplicity correction. All evaluated outcomes remain in the summaries.

Ablation and unfavorable outcomes. Response polishing generally improves on the estimate from the same normalized composite iterations, as shown in Figure 2. Both Huber and Tukey families are selected. Every proposal passes the stored-objective guard, and the numerical scale floor is never

active. The acceptance test rejects none of the evaluated proposals. Recovery remains conditional on the initialization and local-growth hypotheses.

The leverage-selected response attack exposes a limitation: Some initial estimates miss the basin and produce large final errors despite a favorable median. All trials are retained in Table 4 and the raw records. A feasible proposal that passes a local guard does not certify that the practical initializer entered the theoretical basin. The limitation matters since the implemented grid and radius differ from the conservative sufficient constants.

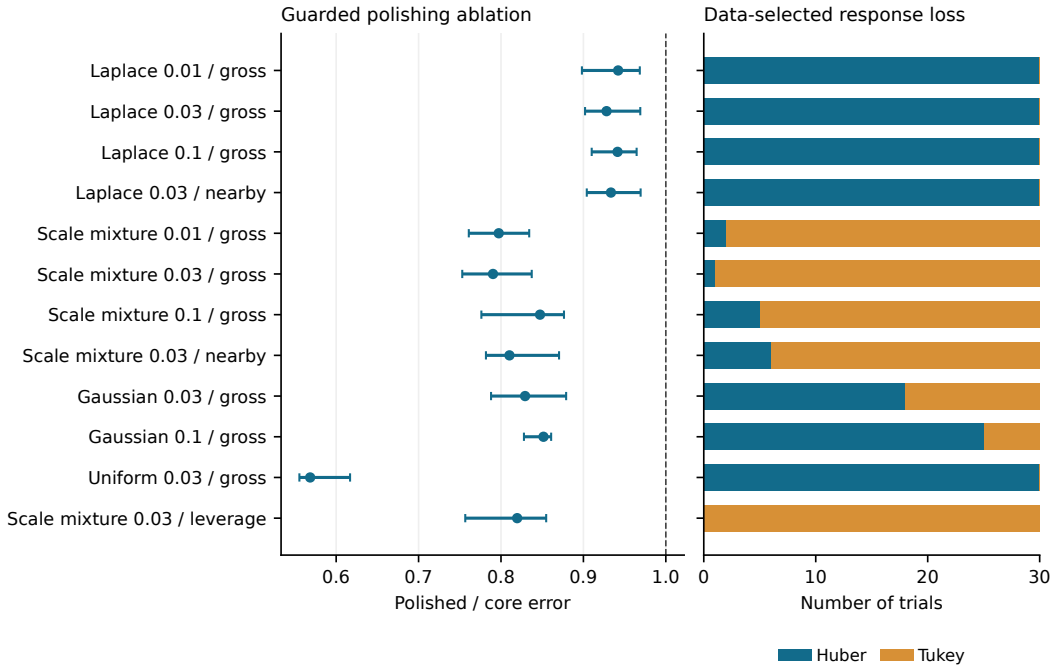

Figure 2: Effect of response polishing over all twelve settings. Left: Median paired ratios of the final coefficient error to the error before response polishing, with descriptive 95% bootstrap intervals. Both estimates share the same initialization and composite iterates. Right: Numbers of datasets selecting the Huber and Tukey losses. Ground truth is used only to evaluate coefficient error.

Computation. The median total runtime is 1.230 seconds for SARP-P and 0.0137 seconds for Mix-IRLS on the evaluation host. Response polishing takes a median 0.0133 seconds. Solving the nonsmooth conic subproblems accounts for most of the additional computation. All reported fits finish with finite coefficient estimates. Large statistical errors remain included in every summary.

F.2 PAIRED DATASETS AND NUMERICAL CONVENTIONS

All twelve settings use the thirty seeds 10000–10029. Equal seeds identify paired datasets across methods and related settings. Saved estimates allow coefficient errors, calibration statistics, and acceptance inequalities to be recomputed independently. True coefficients and labels enter only data generation and evaluation. Fitting uses the observations, the contamination budget, and p_0.

The median in the residual calibration averages the two central retained values when their count is even. The Huber derivative at $|u| = \lambda$ equals one in calibration. Tukey retains its signed derivative $(1 - u^2/\lambda^2)(1 - 5u^2/\lambda^2)$ inside the cutoff. The implementation fixes the selected family, cutoff, and scale during response polishing. If the raw median is exactly zero, the procedure returns $\boldsymbol{c} = \boldsymbol{\theta}_T$ before applying the numerical floor. A numerical failure during response polishing also returns $\boldsymbol{c}$ and is logged.

The normalized composite iteration uses at most 80 outer iterations, with a coefficient-change threshold 10^{-8} in normalized coordinates. Huber and Tukey response polishing use at most 100

Table 3: Initialization and computational budgets for the robust comparisons. Iteration limits refer to the outer iterations of each specified update. Runtimes are medians over the complete primary study, measured in seconds and including initialization and any likelihood profiling.

Method	Initialization	Iteration limit	Seconds
SARP-P	Segment	80 composite, 100/200 polishing	1.230
Mix-IRLS	Gaussian random	1000 per published stage	0.0137
Mixregt	Shared segment	1000 per degree of freedom	0.396
Laplace EM	Shared segment	500	0.668
ILTS	Shared segment	1000	0.0720

and 200 weighted least-squares iterations respectively, with change threshold $\epsilon_{\mathrm{r}} = 10^{-9}$ in original coordinates. The differing maxima allow the flatter redescending objective additional iterations. The recovery theorems require bounds on the composite objective gaps and allow any finite number of response-polishing updates.

F.3 Initialization, tuning, and computational budgets

The protocol below specifies how initialization and computational effort enter the comparisons. Each method is called once per dataset, without an external restart search or a wall-clock budget shared across methods. Mix-IRLS starts from two independent standard Gaussian coefficient vectors. ILTS, Mixregt, and Laplace EM receive the same segment initializer as SARP-P. Mixregt fits all fifteen prescribed degrees of freedom from that initializer and selects the largest observed likelihood. No method uses the coefficient error to select an initialization, parameter, or returned estimate.

Mix-IRLS uses weight threshold 0.1, its parameters $\nu = 0.5$ and $\rho = 1$, and stopping threshold 10^{-6}. The published response-outlier option receives $\bar{\epsilon} = 0.05$. ILTS stops when the coefficient change is below $10^{-9}\tau$. For Mixregt and Laplace EM, write $b_k > 0$ for the fitted common scale and L_k for the observed log likelihood at iteration k. Either of the following inequalities terminates a fit:

$$\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\| + |b_{k+1} - b_k| \le 10^{-9} \max\{b_k, 10^{-8}\}, \qquad |L_{k+1} - L_k| \le 10^{-9} \max\{1, |L_k|\}. \tag{322}$$

The first scale is the Student scale for Mixregt and the Laplace scale for Laplace EM. The Mixregt degrees-of-freedom grid is $\{1, \dots, 15\}$. Laplace EM solves each weighted-LAD subproblem with primal–dual and feasibility thresholds 10^{-9}. The SARP-P stopping rules appear in the preceding subsection. Table 3 documents accuracy at the specified iteration budgets. The results do not establish superiority under equal runtime or a tuned restart budget. In particular, Mix-IRLS is substantially faster in the reported implementation.

F.4 Magnitude of the sufficient parameters

For the experimental choices $p_0 = 0.2$ and $a = 0.1$, the exact formulas in the recovery theorem give

$$\begin{aligned} C_\tau &= 144\sqrt{5} \approx 321.99379, \\ \mu &= \frac{1}{2048(170 + 72\sqrt{5})} \approx 1.4751838 \times 10^{-6}, \\ \delta_{\mathrm{g}}^{\max} &= \frac{p_0 \mu a c_\tau}{8192\sqrt{2} C_\tau} \approx 2.4715745 \times 10^{-16}. \end{aligned} \tag{323}$$

On the clean-design event $1/2 \le \Lambda \le 2$, the certified radius obeys

$$\frac{r}{\tau} = \frac{\mu a}{16\Lambda} \in [\mu/320, \mu/80] \subset [4.6099 \times 10^{-9},\ 1.8440 \times 10^{-8}]. \tag{324}$$

The theoretical grid covers $[-64\tau, 64\tau]$. Even at the largest permitted spacing, its number of points and ordered candidate pairs satisfy

$$\begin{aligned} |\mathcal{G}| &= 1 + \left\lceil \frac{128}{\delta_{\mathrm{g}}} \right\rceil \ge 1 + \left\lceil \frac{128}{\delta_{\mathrm{g}}^{\max}} \right\rceil \approx 5.1788849 \times 10^{17}, \\ |\mathcal{G}|^2 &\ge \left(1 + \left\lceil \frac{128}{\delta_{\mathrm{g}}^{\max}} \right\rceil \right)^2 \approx 2.6820849 \times 10^{35}. \end{aligned} \tag{325}$$

By comparison, the practical initialization searches $[-3\tau, 3\tau]$ with initial spacing 0.05τ and scalar refinement, and the practical composite radius is 2τ. Its initial regular grid has 121 points. The practical spacing is approximately 2.0230×10^{14} times larger than $\delta_{\mathrm{g}}^{\max}\tau$, and the practical radius is approximately 1.0846×10^{8} to 4.3384×10^{8} times larger than the certified radius over $\Lambda \in [1/2, 2]$.

The sufficient construction therefore certifies an impractical parameterization of the method. The reported experiments evaluate a much less conservative parameterization and do not constitute numerical verification of the sufficient basin conditions. Passing the final acceptance test only controls recovery after the uniform local-growth conditions have been established on the same ball. Acceptance by the practical guard does not itself establish those conditions. Sharper constants or an independently justified basin diagnostic remain necessary to certify the evaluated parameterization.

F.5 Baseline objectives and implementation fidelity

The baselines use their published loss functions and existing outlier-handling options. We specify their updates and initialization below such that improvements from response polishing can be assessed without attributing new robustness modifications to the comparison methods.

Mix-IRLS and ILTS. Mix-IRLS uses the published implementation with its existing contamination-fraction option set to 0.05 and stopping threshold 10^{-6} for noisy data (Zilber & Nadler, 2023). Numerical compatibility corrections preserve its residual-based outlier rule. ILTS repeatedly retains the $\lfloor 0.8pn \rfloor$ smallest residuals for each fitted coefficient and solves least squares on that subset (Shen & Sanghavi, 2019). We supply $p = 1/2$ and the shared robust initial pair. The comparison uses the local ILTS update with a supplied data-dependent initializer, not a reproduction of its global coefficient-subspace search.

Student t mixture (Mixregt). We implement Algorithm 2.1 and equations (2.6)–(2.10) of Yao et al. (2014), with common fitted scale and fitted mixing proportions. For completeness, if π_j, $\sigma_t > 0$, and $\nu > 0$ are the current proportion, scale, and degrees of freedom, define $r_{ij} = y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_j$ and

$$\gamma_{ij} = \frac{\pi_j t_\nu(r_{ij}/\sigma_t)}{\sum_{k=1}^{2} \pi_k t_\nu(r_{ik}/\sigma_t)}, \qquad u_{ij} = \frac{\nu + 1}{\nu + r_{ij}^2/\sigma_t^2}, \tag{326}$$

where t_ν is the standard Student density. The coefficient update minimizes $\sum_{i=1}^{n} \gamma_{ij} u_{ij} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_j)^2$, the proportion update is $n^{-1} \sum_{i=1}^{n} \gamma_{ij}$, and the common variance update is $n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{2} \gamma_{ij} u_{ij} r_{ij,\mathrm{new}}^2$. We profile each integer $\nu \in \{1, \ldots, 15\}$ and select the largest observed likelihood, following Section 4 of Yao et al. (2014). Every profile fit starts from the same initial pair. True estimation error is not used to select ν. The model already supplies bounded response influence through its latent precisions. We do not add response trimming.

Laplace mixture. We use equation (4) of Song et al. (2014) and the common-scale option described on page 131. With current common scale $b > 0$, the responsibilities and M steps are

$$\gamma_{ij} = \frac{\pi_j e^{-|r_{ij}|/b}}{\sum_{k=1}^{2} \pi_k e^{-|r_{ik}|/b}}, \qquad \boldsymbol{\beta}_j^+ \in \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \gamma_{ij} |y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}|, \tag{327}$$

$$\pi_j^+ = \frac{1}{n} \sum_{i=1}^{n} \gamma_{ij}, \qquad b^+ = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{2} \gamma_{ij} |y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_j^+|. \tag{328}$$

The weighted-LAD M step is solved as a linear program. The membership-EM implementation maximizes the published Laplace-mixture likelihood, while using a direct coefficient solver instead of its additional latent-scale IRLS representation. Likelihood monotonicity and agreement of the weighted-LAD objective with an independent linear-program solver were verified. The weighted-LAD solver adds no outlier adaptation.

F.6 Composite subproblem conditioning and objective gaps

Centering and rescaling improve the conditioning of the conic subproblem while preserving its objective up to a known scale factor. The formulas below relate numerical solver gaps to the

Table 4: Coefficient error under simultaneous inlier noise and 5% response outliers. Entries give the median and interquartile-range width (in parentheses) of $\mathcal{E}(\boldsymbol{\theta})$ over thirty datasets, both in units of 10^{-3}. The parameter s is the inlier standard deviation. Bold indicates the smallest median in each row. All trials are included.

Noise / replacements	s	SARP-P	Mix-IRLS	Mixregt	Laplace EM	ILTS
Laplace / gross	0.01	**2.35** (0.34)	2.89 (0.38)	2.44 (0.41)	2.64 (0.34)	2.88 (0.45)
Laplace / gross	0.03	**7.18** (1.01)	8.81 (1.16)	7.31 (1.13)	7.89 (1.11)	8.85 (1.03)
Laplace / gross	0.1	**25.09** (3.72)	31.09 (3.59)	25.30 (4.08)	28.16 (4.81)	32.06 (5.83)
Laplace / nearby	0.03	**7.22** (1.23)	8.54 (1.41)	7.36 (1.25)	7.98 (0.94)	9.00 (1.61)
Scale mixture / gross	0.01	**1.06** (0.12)	2.72 (0.32)	1.27 (0.25)	1.35 (0.25)	1.32 (0.21)
Scale mixture / gross	0.03	**3.23** (0.47)	7.79 (1.15)	3.88 (0.76)	4.08 (0.76)	4.04 (0.76)
Scale mixture / gross	0.1	**11.06** (1.52)	24.87 (4.33)	13.11 (2.76)	13.78 (2.69)	15.39 (3.07)
Scale mixture / nearby	0.03	**3.36** (0.44)	5.53 (1.05)	3.92 (0.74)	4.21 (0.83)	4.18 (0.88)
Gaussian / gross	0.03	**9.34** (1.56)	9.43 (1.59)	11.06 (2.92)	11.25 (1.84)	14.27 (2.31)
Gaussian / gross	0.1	33.97 (5.79)	**33.18** (5.72)	40.20 (9.07)	40.19 (6.60)	58.61 (11.97)
Uniform / gross	0.03	9.33 (1.24)	**9.29** (1.34)	16.49 (1.59)	15.85 (1.47)	23.47 (2.20)
Scale mixture / leverage	0.03	**3.33** (0.74)	8.65 (1.26)	4.16 (1.27)	6.76 (2031)	999.59 (0.74)

mathematical gaps used in the recovery analysis. The numerical optimization uses coordinates $\widetilde{y}_i = (y_i - \boldsymbol{x}_i^\top \boldsymbol{b})/\tau$ and $\widetilde{\boldsymbol{\beta}}_j = (\boldsymbol{\beta}_j - \boldsymbol{b})/\tau$. The original absolute-composite objective is τ times its normalized-coordinate counterpart. Thus the reported conic model gap is the original objective gap divided by τ. Rows with

$$|c_{i,t}| > \|\boldsymbol{A}_{i,t}\|(r + \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|) \tag{329}$$

have an affine residual with constant sign throughout the ball. The corresponding absolute values equal affine expressions throughout the ball. Eliminating the associated epigraph variables therefore preserves the optimization problem, up to additive constants. The numerical implementation checks conic primal and dual residuals and records the primal–dual gap of the solver. Proposition C.10 states a mathematical feasible-dual certificate. A floating-point solver report is not an exact-arithmetic proof.

The final guard is evaluated in original response units, using the stored last composite reference and slack $\widehat{s}$. The acceptance ball has the same practical radius 2τ as the normalized composite iteration. All guard tests pass, including the unfavorable large-error trials. Theorem 4.2 also requires the initialization and feasible-ball hypotheses of Theorem 4.1, which guard acceptance alone does not verify.

F.7 Complete variability and reproducibility

Table 4 reports every method and setting, including large finite errors. The retained individual records support independent checks of coefficient errors, calibration choices, and acceptance inequalities.

Means and maxima were also retained for checking the complete trial distribution. Bootstrap intervals use a fixed bootstrap seed and 10000 replicates. Within a setting, trials are resampled by seed. For pooled ratios, the same seed block is selected jointly across all settings to preserve their dependence. The intervals are descriptive and unadjusted for multiple comparisons or development-driven choice of noise mechanisms.

Runtimes were recorded with one numerical-library thread per fit and three concurrent groups of settings. The runtimes are descriptive measurements from a shared machine. An independent numerical check regenerated each dataset, recomputed all coefficient errors, and verified the stored calibration and acceptance inequalities. One full fit in each setting was repeated to reconstruct its last frozen objective independently. The 360 comparisons reported here exclude development trials.

Table 5: Errors across the stages of the adaptive estimator on the four calibration-ablation settings. Entries give median coefficient error (interquartile-range width), in units of 10^{-3}, over all thirty paired datasets per row. The H/T column counts selected Huber and Tukey score families. All 120 adaptive proposals are accepted, hence raw-proposal and final-output errors coincide.

Noise	Initialization $\boldsymbol{\theta}_0$	Composite $\boldsymbol{\theta}_T$	Raw proposal $\boldsymbol{z}$	Output $\boldsymbol{\theta}_{\mathrm{out}}$	H/T
Laplace	311.57 (123.31)	7.71 (0.85)	7.18 (1.01)	7.18 (1.01)	30/0
Scale mixture	301.23 (144.60)	4.00 (0.65)	3.23 (0.47)	3.23 (0.47)	1/29
Gaussian	327.95 (113.55)	11.16 (2.46)	9.34 (1.56)	9.34 (1.56)	18/12
Uniform	312.27 (122.73)	15.94 (1.85)	9.33 (1.24)	9.33 (1.24)	30/0

Table 6: Calibration ablation under simultaneous noise and 5% response outliers. Entries are median coefficient errors with interquartile-range widths in parentheses, both in units of 10^{-3}, over all thirty original datasets per row. The noise standard deviation is 0.03. Every variant uses the same segment initialization, normalized composite estimate, calibrated residual scale, and final acceptance test. Fixed multipliers are chosen in advance. Family-only selection minimizes the same empirical variance criterion over the corresponding frozen cutoff grid.

Noise	Adaptive	Huber 1.345	Tukey 4.685	Huber-only	Tukey-only	No polishing
Laplace	7.18 (1.01)	7.40 (1.15)	7.56 (1.04)	7.18 (1.01)	7.79 (1.42)	7.71 (0.85)
Scale mixture	3.23 (0.47)	3.56 (0.77)	3.23 (0.47)	3.53 (0.72)	3.23 (0.47)	4.00 (0.65)
Gaussian	9.34 (1.56)	9.22 (1.38)	9.56 (1.87)	9.37 (1.43)	9.51 (1.75)	11.16 (2.46)
Uniform	9.33 (1.24)	9.85 (1.48)	10.24 (1.36)	9.33 (1.24)	9.54 (1.25)	15.94 (1.85)

F.8 Calibration ablation and errors across stages

We reuse the four noise distributions, coefficients, Gaussian covariates, and gross response-outlier mechanism of Section 5, with $n = 1000$, $d = 20$, $\Delta = 1$, $s = 0.03$, and $\epsilon = \bar{\epsilon} = 0.05$. Each noise setting uses all thirty original paired datasets. Every variant shares the segment initializer, normalized composite estimate $\boldsymbol{\theta}_T$, residual scale $\widehat{s}$, and frozen-objective acceptance test. The variants differ only in the response score. Fixed Huber and Tukey use multipliers 1.345 and 4.685. Huber-only and Tukey-only minimize (16) within the corresponding grid in (15). The adaptive variant is SARP-P, and no polishing returns $\boldsymbol{\theta}_T$. All response updates use the original iteration limits and stopping rules.

Table 5 separates initialization, composite iteration, the raw response proposal, and the returned estimate. Response polishing reduces the median error in each setting. The raw and returned errors coincide since every adaptive proposal passes the acceptance test. Table 6 isolates score selection from initialization and scale estimation. Adaptive selection attains or ties the lowest median among the evaluated variants for Laplace, scale-mixture, and uniform noise. Fixed Huber has a slightly lower Gaussian median. The selector favors Huber for Laplace and uniform noise and Tukey for scale-mixture noise. The observations support distribution-dependent calibration without establishing optimality of the empirical variance criterion or improvement from every polishing update.

F.9 Sensitivity to supplied nuisance bounds

We keep $n = 1000$, $d = 20$, $\Delta = 1$, scale-mixture standard deviation $s = 0.03$, and true outlier fraction $\epsilon = 0.05$. Coefficients and gross response outliers follow the primary study. Labels are independent with the proportion p in Table 7. The contamination-budget experiment fixes $p = 1/2$ and supplied $p_0 = 0.2$, then varies $\bar{\epsilon} \in \{0.02, 0.05, 0.10\}$. The proportion-bound experiment fixes $p = 0.2$ and $\bar{\epsilon} = 0.05$, then varies $p_0 \in \{0.1, 0.2, 0.3\}$. Within each experiment, all three settings use the same ten datasets generated with seeds 20000–20009. The bounds are prescribed before fitting, and ground truth is used only for data generation and evaluation. Residual calibration estimates the noise scale while p_0 and $\bar{\epsilon}$ remain supplied inputs.

Table 7 reports median errors and acceptance frequencies. Supplied budgets give similar medians in the tested setting, including the underestimated budget 0.02 outside the recovery assumption. Table 8 shows why similar medians for the proportion-bound experiment do not establish reliable recovery. The valid lower bound $p_0 = 0.1$ yields one large initialization error that remains large after an accepted proposal. For $p_0 = 0.3$, which exceeds the true smaller-component proportion, the feasible-ball test rejects a beneficial proposal and retains the less accurate composite estimate. All

Table 7: Sensitivity to supplied nuisance bounds with scale-mixture noise of standard deviation 0.03 and a true response-outlier fraction of 0.05. Entries give median coefficient error (interquartile-range width), in units of 10^{-3}, over ten paired datasets per cell. A dagger marks a supplied bound outside the corresponding model assumption. Acceptance counts include every dataset.

Cell	True p	Supplied p_0	Supplied $\bar{\varepsilon}$	Adaptive	No polishing	Accepted
Budget†	0.5	0.2	0.02	3.33 (0.26)	4.32 (0.39)	10/10
Budget	0.5	0.2	0.05	3.32 (0.44)	4.32 (0.29)	10/10
Budget	0.5	0.2	0.10	3.50 (0.42)	4.31 (0.53)	10/10
Proportion	0.2	0.1	0.05	4.20 (0.58)	5.90 (0.62)	10/10
Proportion	0.2	0.2	0.05	4.19 (0.53)	5.52 (0.76)	10/10
Proportion†	0.2	0.3	0.05	4.19 (0.56)	5.30 (0.78)	9/10

Table 8: Complete sensitivity variability over all ten datasets per cell. Mean and maximum coefficient errors are shown in units of 10^{-3}. Large finite errors are retained. A dagger identifies an input bound outside the corresponding model assumption.

Cell	Adaptive mean	Adaptive maximum	No-polishing mean	No-polishing maximum
$\bar{\varepsilon} = 0.02^\dagger$	3.38	4.16	4.44	5.04
$\bar{\varepsilon} = 0.05$	3.42	4.14	4.29	4.93
$\bar{\varepsilon} = 0.10$	3.57	4.17	4.24	4.72
$p_0 = 0.1$	33.78	299.12	37.08	319.79
$p_0 = 0.2$	4.24	4.94	5.43	5.96
$p_0 = 0.3^\dagger$	8.75	49.50	9.67	49.50

trials remain in the summaries. The supplied proportion bound affects the practical initialization, and guard acceptance alone does not certify entry into the theoretical basin.

Reproducibility of the supplemental comparisons. The supplemental registry fixed data mechanisms, variants, and seeds before fitting. Every registered dataset is retained. Recomputing the adaptive variants reproduces the original coefficient estimates exactly, and independent checks recover every saved error and acceptance decision. The supplemental comparisons were designed after the primary study and reuse its ablation datasets. The resulting summaries are descriptive assessments of calibration and input sensitivity.