Target-Risk Identification and Honest Inference from Weak Labels: The Observed Fiber, Not the Row Space

Anonymous authors
Paper under double-blind review

Abstract

We study target-population evaluation of a fixed predictor when clean target labels are unavailable and source labels are observed only through weak supervision. The standard loss-correction view says that the clean target risk is estimable when the clean loss vector lies in the row space of the weak-label channel, so that an unbiased corrected weak-label loss exists. We show that this row-space condition answers a stronger, uniform question, not the observed-law question faced by an evaluator. The correct population object is the observed weak-label fiber: the set of clean posteriors that reproduce the observed weak conditional distribution. Under exact covariate shift and overlap, the target risk is point identified exactly when the clean-loss functional is constant on this fiber almost surely; otherwise, two pointwise linear programs give the sharp identified interval. The main technical addition is a finite-sample inference layer for the realistic case in which the weak-label law, target covariate weights, and weak-label channel are estimated or sensitivity-modeled. We introduce *confidence fibers*, prove honest coverage of the clean target risk from joint confidence sets for these nuisance objects, give an exact linear-program formulation under polyhedral multinomial confidence sets, and show convergence to the structural fiber interval without a separation condition at the boundary between point and partial identification. The resulting audit output is deliberately conservative: it certifies point identification when weak labels justify a number, and otherwise reports an honest interval rather than a pseudo-corrected point estimate. Public WRENCH audits illustrate the warning that coarse weak labels can severely understate clean risk while confidence-fiber intervals expose the missing information.

1 Introduction

Weak labels are common in modern evaluation pipelines. A deployed classifier may be assessed using noisy annotations, complementary labels, crowd labels, heuristic labels, or programmatic weak supervision, while the target population may differ from the source population in its covariate distribution. The operational question is simple: can we evaluate the clean target risk of a fixed predictor without clean target labels?

Let $Y \in \{1, \dots, K\}$ be the unobserved clean label, $Z \in \{1, \dots, J\}$ the weak label, and f a fixed predictor. For a loss ℓ, define the clean-loss vector

$$L_f(x) = (\ell(f(x), 1), \dots, \ell(f(x), K))^\top. \tag{1.1}$$

If the weak-label channel $K_x(z, y) = P(Z = z \mid X = x, Y = y)$ is known and there is a vector $g_x \in \mathbb{R}^J$ satisfying

$$K_x^\top g_x = L_f(x), \tag{1.2}$$

then $g_x(Z)$ is an unbiased corrected loss for the clean loss at covariate value x. This is the algebra behind loss correction for corrupted labels (Natarajan et al., 2013; Patrini et al., 2017; Menon et al., 2015; Scott et al., 2013); with a density ratio $w = dQ_X/dP_X$, it yields a covariate-shift risk estimator (Shimodaira, 2000; Sugiyama et al., 2007; Bickel et al., 2009). Because (1.2) is solvable exactly when $L_f(x)$ lies in the row space of K_x, row-space membership is often treated as the identification criterion.

This paper shows that row-space membership is not the right first question. It asks for a corrected loss that works for every clean posterior consistent with the channel. Evaluation at the actually observed weak-label law is weaker and has a different identification object. The observable conditional weak law is

$$r_x = P(Z = \cdot \mid X = x), \tag{1.3}$$

and it restricts the latent clean posterior to the *observed fiber*

$$\Gamma_x = \{p \in \Delta_K : K_x p = r_x\}. \tag{1.4}$$

Only the values of $p \mapsto L_f(x)^\top p$ on this fiber matter. A risk can therefore be point identified even when no full corrected loss exists. Conversely, if the functional varies on the observed fiber, weak labels alone justify an identified interval, not a single clean-risk number. This changes the evaluation target: the method is not a new way to force weak labels into a point estimate, but an audit that decides whether a point estimate is identified at all.

Contributions. The main contribution is the identification target: observed fibers, not row spaces, determine whether a fixed black-box risk is identified at the observed weak law. We make this precise in five steps. First, under overlap and exact covariate shift, we prove that $R_Q(f)$ is point identified if and only if $L_f(x)^\top p$ is constant over Γ_x for Q_X-almost every x. Second, when point identification fails, we give the sharp target-risk interval by two pointwise linear programs over Γ_x, together with dual certificates that can be checked from the observed weak law. Third, and crucially for practice, we replace oracle nuisances by *confidence fibers*: finite-stratum confidence or sensitivity sets for the channel K, weak law r, and target covariate weights q. The resulting random interval has honest finite-sample coverage whenever the nuisance confidence system has joint coverage; with polyhedral multinomial sets, its endpoints are still ordinary LPs. Fourth, we prove convergence back to the structural fiber interval and show that the coverage statement is uniform at the boundary between point identification and partial identification. Fifth, we prove that the usual row-space condition is exactly the uniform identification criterion, derive an active-face refinement equivalent to observed-specific point identification, and give corrected-loss estimators only after identification has been diagnosed. Controlled simulations and public WRENCH audits then instantiate the protocol and show when weak-label evaluation is informative, uninformative, or overconfident.

Scope. The paper is an evaluation paper, not a training algorithm. The predictor is fixed before the audit sample is used, and the clean audit labels in the public case studies are held out from the interval construction and revealed only to check the protocol. The intended output is a defensible risk statement under stated channel, shift, and overlap assumptions: either a certified point-identified risk or an interval that displays the information weak labels do not contain.

2 Setup

Let $\mathcal{X}$ be a standard Borel space, $\mathcal{Y} = \{1, \dots, K\}$, and $\mathcal{Z} = \{1, \dots, J\}$. Source clean data follow a law P on (X, Y), but the evaluator observes only (X, Z). At the population level, the weak label is generated by a column-stochastic matrix

$$K_x(z, y) = P(Z = z \mid X = x, Y = y). \tag{2.1}$$

The channel may be known from design, estimated from a clean calibration split, or specified through a sensitivity set. Sections 3–5 first describe the population identification target; Section 4 gives honest finite-sample inference when K_x, r_x, and the target covariate law are uncertain. Let

$$p_x = P(Y = \cdot \mid X = x) \in \Delta_K, \qquad r_x = P(Z = \cdot \mid X = x) \in \Delta_J. \tag{2.2}$$

The observable relation is

$$r_x = K_x p_x. \tag{2.3}$$

The target clean law is Q. We assume exact covariate shift,

$$Q(Y = \cdot \mid X = x) = P(Y = \cdot \mid X = x) \quad \text{for } Q_X\text{-almost every } x, \tag{2.4}$$

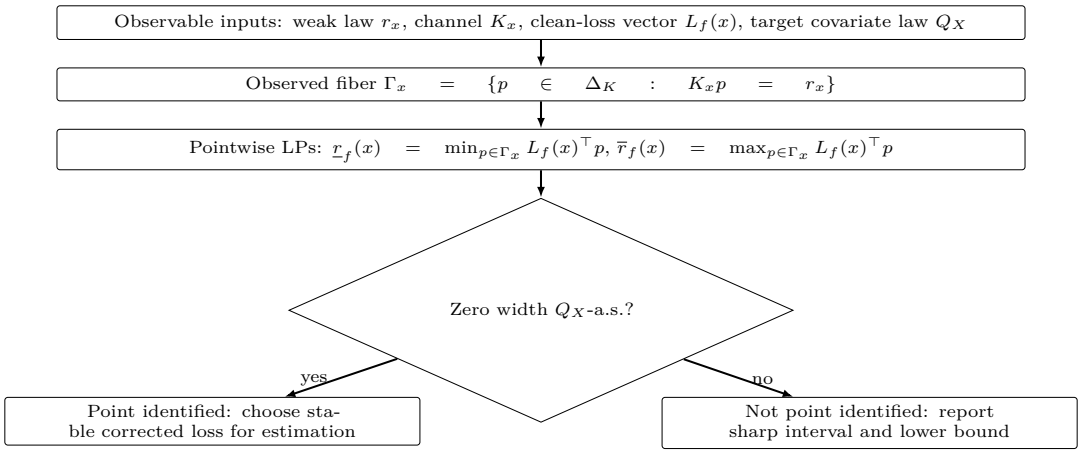

Figure 1: Observed-fiber workflow. Corrected-loss estimation enters only after the fiber interval has degenerated.

overlap $Q_X \ll P_X$, and integrability

$$\int \|L_f(x)\|_\infty \, dQ_X(x) < \infty. \tag{2.5}$$

The target risk is

$$R_Q(f) = \mathbb{E}_Q[\ell(f(X), Y)] = \int L_f(x)^\top p_x \, dQ_X(x). \tag{2.6}$$

The predictor is fixed before the weak-label evaluation sample is used. If f is selected using the same sample, sample splitting or post-selection analysis is required.

Assumption 2.1 (Measurability and nonempty fibers)**.** *The maps $x \mapsto K_x$, $x \mapsto r_x$, and $x \mapsto L_f(x)$ are measurable. For $(P_X + Q_X)$-almost every x, the fiber*

$$\Gamma_x = \{p \in \Delta_K : K_x p = r_x\} \tag{2.7}$$

is nonempty. The pointwise minima and maxima over Γ_x below are measurable and admit measurable minimizer and maximizer selections.

For finite K and J, these are standard measurable-selection conditions; they are automatic on finite covariate strata.

3 Observed-Fiber Identification

For each x, define the pointwise lower and upper clean conditional risks

$$\underline{r}_f(x) = \min_{p\in\Gamma_x} L_f(x)^\top p, \tag{3.1}$$

$$\overline{r}_f(x) = \max_{p\in\Gamma_x} L_f(x)^\top p. \tag{3.2}$$

These are linear programs over compact polytopes.

Theorem 3.1 (Sharp identified interval)**.** *Under Assumption 2.1, integrability (2.5), covariate shift (2.4), and overlap $Q_X \ll P_X$, the sharp identified set for $R_Q(f)$ given (P_{XZ}, Q_X, K) is*

$$\mathcal{I}(f) = \left[\int \underline{r}_f(x) \, dQ_X(x), \ \int \overline{r}_f(x) \, dQ_X(x)\right]. \tag{3.3}$$

Consequently, $R_Q(f)$ is point identified if and only if

$$p \mapsto L_f(x)^\top p \quad \text{is constant on } \Gamma_x \quad Q_X\text{-almost surely.} \tag{3.4}$$

The theorem says that the identifiable object is the image of the observed fiber under the clean-loss functional. It is neither the full clean posterior p_x nor a full corrected-loss representation. If the image is not a singleton, the interval in (3.3) is the narrowest target-risk interval justified by the observable law and assumptions.

Corollary 3.2 (Finite-stratum diagnostic)**.** *Suppose $\mathcal{X}$ is partitioned into finitely many strata $s \in \mathcal{S}$ such that K_x, r_x, and $L_f(x)$ are constant within each stratum. Write these values as K_s, r_s, and L_s, and let $q_s = Q_X(s)$. Then*

$$\mathcal{I}(f) = \left[\sum_{s\in\mathcal{S}} q_s \underline{r}_s,\ \sum_{s\in\mathcal{S}} q_s \overline{r}_s\right], \tag{3.5}$$

$$\underline{r}_s = \min_{p\in\Delta_K : K_s p = r_s} L_s^\top p, \qquad \overline{r}_s = \max_{p\in\Delta_K : K_s p = r_s} L_s^\top p. \tag{3.6}$$

The active label set in stratum s is computed by

$$A_s = \left\{ y : \max_{p\in\Delta_K : K_s p = r_s} p_y > 0 \right\}. \tag{3.7}$$

Thus the same LP family computes the sharp interval, diagnoses point identification, and constructs the active-face row-space test.

Corollary 3.3 (Dual certificates)**.** *For every x with nonempty Γ_x, the endpoints have the dual forms*

$$\underline{r}_f(x) = \sup_{a\in\mathbb{R}^J} \{a^\top r_x : K_x^\top a \le L_f(x)\}, \tag{3.8}$$

$$\overline{r}_f(x) = \inf_{b\in\mathbb{R}^J} \{b^\top r_x : K_x^\top b \ge L_f(x)\}. \tag{3.9}$$

A pointwise certificate of identification is therefore a pair (a, b) such that $K_x^\top a \le L_f(x) \le K_x^\top b$ and $a^\top r_x = b^\top r_x$. If $K_x^\top g = L_f(x)$ is solvable, then $a = b = g$ is a certificate, but certificates can exist even when full row-space membership fails.

Operational diagnostic. Given an estimate or model for r_x, solve the two LPs in (3.1)–(3.2) on each stratum or covariate cell. If the integrated width is positive, report the sharp interval. If it degenerates, the risk is point identified and one can choose a stable corrected loss for estimation. This sequence deliberately separates identification from finite-sample nuisance estimation.

Corollary 3.4 (Observational equivalence)**.** *If*

$$\int \{\overline{r}_f(x) - \underline{r}_f(x)\}\, dQ_X(x) > 0, \tag{3.10}$$

then there exist two clean conditional laws that induce the same observable source law P_{XZ} and the same target covariate law Q_X, satisfy exact covariate shift, and have different target risks.

Proposition 3.5 (Width lower bound)**.** *Fix any model class in which two clean laws induce the same observable law but have target risks R_1 and R_2. Let $\Delta = |R_1 - R_2|$. Every point estimator $\widehat{R}$ based only on observable data satisfies*

$$\max_{j=1,2} \mathbb{E}_j[(\widehat{R} - R_j)^2] \ge \Delta^2/4. \tag{3.11}$$

If a random interval $C \subset \mathbb{R}$ satisfies $\mathbb{P}_j(R_j \in C) \ge 1 - \delta$ for $j = 1, 2$, then under the common observable law,

$$\mathbb{P}(\text{length}(C) \ge \Delta) \ge 1 - 2\delta. \tag{3.12}$$

A positive LP width is therefore a minimax obstruction, not a conservative estimator artifact.

4 Confidence Fibers Under Estimated Channels

The sharp interval in Theorem 3.1 is a population identified set: it assumes that the weak law r_x, the channel K_x, and the target covariate law Q_X are the objects supplied to the identification problem. In weak-supervision audits these objects are usually estimated from a clean calibration split, a weak-label audit split, and a target-covariate sample. This section gives the missing inference layer. We state it for finite observable strata because unrestricted nonparametric estimation of $x \mapsto (K_x, r_x)$ is a separate smoothing problem and because the audit protocol in Section 7 is stratum-based.

Let $S \in \mathcal{S} = \{1, \dots, m\}$ be an observable stratum. This section uses the stratified covariate-shift model

$$Q(Y = \cdot \mid S = s) = P(Y = \cdot \mid S = s) \qquad \text{for every stratum with } q_s > 0, \tag{4.1}$$

which is the finite-stratum analogue of (2.4). It holds, for example, when S is the full covariate used for evaluation, when the clean conditional law and the loss are constant within strata, or when the target law is formed by reweighting source strata. Within stratum s, write

$$p_s = P(Y = \cdot \mid S = s), \quad r_s = P(Z = \cdot \mid S = s), \quad K_s(z, y) = P(Z = z \mid Y = y, S = s), \quad q_s = Q_X(S = s), \tag{4.2}$$

and let $L_s \in \mathbb{R}^K$ be the clean-loss vector of the fixed predictor on that stratum. Under (4.1), the structural finite-stratum interval is the interval in Corollary 3.2.

Definition 4.1 (Nuisance confidence system)**.** *For each stratum s and clean label y, let $\mathcal{K}_{sy} \subseteq \Delta_J$ be a random set for the channel column $K_s(\cdot, y)$. Let $\mathcal{R}_s \subseteq \Delta_J$ be a random set for the weak conditional law r_s, and let $\mathcal{Q} \subseteq \Delta_m$ be a random set for the target stratum weights $q = (q_1, \dots, q_m)$. We say these sets form a level-$1 - \delta$ nuisance confidence system if*

$$\mathbb{P}\left(q \in \mathcal{Q},\ r_s \in \mathcal{R}_s,\ K_s(\cdot, y) \in \mathcal{K}_{sy} \text{ for all } s \in \mathcal{S}, y \in \mathcal{Y}\right) \geq 1 - \delta. \tag{4.3}$$

The same notation also covers deterministic sensitivity sets, in which case (4.3) is interpreted as the modeling event that the true nuisances lie in the specified sets.

Definition 4.2 (Confidence fiber and confidence-fiber interval)**.** *Given a nuisance confidence system, the confidence fiber in stratum s is*

$$\Gamma_s^{\mathcal{C}} = \left\{ p \in \Delta_K : \exists k_{sy} \in \mathcal{K}_{sy}\ (y = 1, \dots, K),\ \exists r \in \mathcal{R}_s \text{ such that } \sum_{y=1}^{K} p_y k_{sy} = r \right\}. \tag{4.4}$$

The confidence-fiber interval for the target risk is

$$\underline{C}(f) = \inf_{\substack{q \in \mathcal{Q} \\ p_s \in \Gamma_s^{\mathcal{C}} \text{ whenever } q_s > 0}} \sum_{s \in \mathcal{S}} q_s L_s^\top p_s, \tag{4.5}$$

$$\overline{C}(f) = \sup_{\substack{q \in \mathcal{Q} \\ p_s \in \Gamma_s^{\mathcal{C}} \text{ whenever } q_s > 0}} \sum_{s \in \mathcal{S}} q_s L_s^\top p_s, \tag{4.6}$$

$$\widehat{\mathcal{C}}_{1-\delta}(f) = [\underline{C}(f), \overline{C}(f)]. \tag{4.7}$$

A stratum with $q_s = 0$ contributes zero to the target risk, so its posterior is immaterial. When all realized confidence fibers are nonempty, this convention is equivalent to imposing $p_s \in \Gamma_s^{\mathcal{C}}$ for every s. If $\mathcal{K}_{sy} = \{K_s(\cdot, y)\}$, $\mathcal{R}_s = \{r_s\}$, and $\mathcal{Q} = \{q\}$, then $\widehat{\mathcal{C}}_{1-\delta}(f)$ reduces to the structural observed-fiber interval.

Theorem 4.3 (Honest confidence-fiber coverage)**.** *Assume stratified covariate shift (4.1), nonempty true fibers, and finite losses. If the nuisance sets form a level-$1 - \delta$ confidence system in the sense of (4.3), then*

$$\mathbb{P}\{R_Q(f) \in \widehat{\mathcal{C}}_{1-\delta}(f)\} \geq 1 - \delta. \tag{4.8}$$

Moreover, conditional on any realized nuisance sets for which the compatible class is nonempty, the endpoints in (4.5)–(4.6) are the smallest closed interval containing all target risks generated by some $q \in \mathcal{Q}$, some channel columns $k_{sy} \in \mathcal{K}_{sy}$, some weak laws $r_s \in \mathcal{R}_s$, and, for each stratum with $q_s > 0$, some clean posterior $p_s \in \Delta_K$ satisfying $\sum_y p_{sy} k_{sy} = r_s$.

The theorem deliberately separates identification from nuisance learning. It does not require a point estimate of the channel to be correct. If the calibration data are weak, the interval widens; if the channel is only sensitivity-modeled, the interval is a sensitivity result; if the nuisance confidence system is valid, the final interval is honest.

Proposition 4.4 (The confidence-fiber endpoints are LPs)**.** *Suppose the nuisance sets are nonempty polyhedra:*

$$\mathcal{K}_{sy} = \{k \in \mathbb{R}_+^J : \mathbf{1}^\top k = 1,\ A_{sy}^K k \le b_{sy}^K\}, \tag{4.9}$$

$$\mathcal{R}_s = \{r \in \mathbb{R}_+^J : \mathbf{1}^\top r = 1,\ A_s^r r \le b_s^r\}, \tag{4.10}$$

$$\mathcal{Q} = \{q \in \mathbb{R}_+^m : \mathbf{1}^\top q = 1,\ A^q q \le b^q\}. \tag{4.11}$$

Then $\underline{C}(f)$ and $\overline{C}(f)$ are the optimal values of the following linear program, with "min" and "max" objectives respectively:

$$\textit{optimize} \quad \sum_{s=1}^{m}\sum_{y=1}^{K} L_{sy}\mu_{sy} \tag{4.12}$$

$$\textit{over} \quad q_s \ge 0,\ \mu_{sy} \ge 0,\ u_{sy} \in \mathbb{R}_+^J,\ v_s \in \mathbb{R}_+^J,$$

$$\textit{subject to} \quad \mathbf{1}^\top q = 1, \quad A^q q \le b^q, \tag{4.13}$$

$$\sum_{y=1}^{K} \mu_{sy} = q_s, \qquad s = 1, \dots, m, \tag{4.14}$$

$$\mathbf{1}^\top u_{sy} = \mu_{sy}, \quad A_{sy}^K u_{sy} \le \mu_{sy} b_{sy}^K, \qquad s = 1, \dots, m, \tag{4.15}$$

$$y = 1, \dots, K, \tag{4.16}$$

$$v_s = \sum_{y=1}^{K} u_{sy}, \quad \mathbf{1}^\top v_s = q_s, \quad A_s^r v_s \le q_s b_s^r, \qquad s = 1, \dots, m. \tag{4.17}$$

Here $\mu_{sy} = q_s p_{sy}$ and $u_{sy} = q_s p_{sy} k_{sy}$ are perspective variables. Thus allowing estimated channels does not destroy the LP audit protocol. If the sets in (4.9)–(4.11) are replaced by ellipsoids or likelihood-ratio regions, the same perspective construction gives a conic or convex program whenever the column sets are convex and conic-representable.

A concrete multinomial construction. The preceding theorem needs only joint coverage, so the following is one convenient finite-sample choice. Let $\widehat{q}$ be the empirical target-stratum distribution from n_Q target covariates, let $\widehat{r}_s$ be the empirical weak-label distribution from n_s^r audit observations in stratum s, and let $\widehat{K}_{sy}$ be the empirical distribution of Z among n_{sy}^K clean-calibration observations with $(S, Y) = (s, y)$. If a count is zero, take the corresponding empirical vector to be any fixed element of the simplex; the radius below then makes the confidence set the whole simplex. When the cell counts are random, the bounds are applied conditionally on the realized counts and then averaged over those counts. For $d = 1$, set $\rho(1, n, \alpha) = 0$. For $d \ge 2$, $n \ge 1$, and $0 < \alpha < 1$, define

$$\rho(d, n, \alpha) = \min\left\{2, \sqrt{\frac{2\log(c_d/\alpha)}{n}}\right\}, \qquad c_d = 2^d - 2, \tag{4.18}$$

and set $\rho(d, 0, \alpha) = 2$ for $d \ge 2$ and $\rho(d, n, 0) = 2$ for $d \ge 2$. Choose nonnegative error budgets satisfying

$$\delta_Q + \sum_s \delta_s^r + \sum_{s,y} \delta_{sy}^K \le \delta. \tag{4.19}$$

Then define

$$\mathcal{Q} = \{q \in \Delta_m : \|q - \widehat{q}\|_1 \le \rho(m, n_Q, \delta_Q)\}, \tag{4.20}$$

$$\mathcal{R}_s = \{r \in \Delta_J : \|r - \widehat{r}_s\|_1 \le \rho(J, n_s^r, \delta_s^r)\}, \tag{4.21}$$

$$\mathcal{K}_{sy} = \{k \in \Delta_J : \|k - \widehat{K}_{sy}\|_1 \le \rho(J, n_{sy}^K, \delta_{sy}^K)\}. \tag{4.22}$$

The ℓ_1 balls are polytopes after the usual absolute-value linearization, so Proposition 4.4 applies. By the multinomial empirical-distribution inequality of Weissman et al. (2003) and a union bound, these sets satisfy (4.3). More aggressive exact multinomial or simultaneous-interval constructions, such as Goodman-type intervals (Goodman, 1965), can be substituted without changing Theorem 4.3.

Assumption 4.5 (Local fiber stability)**.** *For each stratum s, there are a neighborhood $\mathcal{N}_s$ of (K_s, r_s) and a finite constant H_s such that for every $(\widetilde{K}_s, \widetilde{r}_s) \in \mathcal{N}_s$ with nonempty fiber,*

$$d_H\big(\{p \in \Delta_K : \widetilde{K}_s p = \widetilde{r}_s\}, \{p \in \Delta_K : K_s p = r_s\}\big) \le H_s \left(\|\widetilde{r}_s - r_s\|_1 + \max_y \|\widetilde{K}_s(\cdot, y) - K_s(\cdot, y)\|_1 \right), \quad (4.23)$$

where d_H is Hausdorff distance in ℓ_1. This is a local metric-regularity condition, obtainable from a uniform Hoffman error bound for a locally stable polyhedral system (Hoffman, 1952). It can fail at ill-conditioned strata where arbitrarily small channel or weak-law perturbations cause a large jump in the feasible clean posterior set.

Theorem 4.6 (Excess width and consistency)**.** *Let $B = \max_s \|L_s\|_\infty < \infty$, and let $\mathcal{I}(f)$ be the structural interval computed from the true (q, K, r). Suppose Assumption 4.5 holds. On any event on which $\mathcal{Q}$ and the realized confidence fibers are nonempty, the nuisance sets are contained in the neighborhoods from Assumption 4.5, and the following bounds hold,*

$$\sup_{\widetilde{q} \in \mathcal{Q}} \|\widetilde{q} - q\|_1 \le \eta_q, \quad (4.24)$$

$$\sup_{\widetilde{r} \in \mathcal{R}_s} \|\widetilde{r} - r_s\|_1 \le \eta_s^r, \quad (4.25)$$

$$\sup_{\widetilde{k} \in \mathcal{K}_{sy}} \|\widetilde{k} - K_s(\cdot, y)\|_1 \le \eta_{sy}^K, \quad (4.26)$$

define

$$\Delta_s = B H_s \left(\eta_s^r + \max_y \eta_{sy}^K \right), \qquad \overline{\Delta} = \max_s \Delta_s. \quad (4.27)$$

Then

$$d_H\big(\widehat{\mathcal{C}}_{1-\delta}(f), \mathcal{I}(f)\big) \le B\eta_q + \sum_{s=1}^m q_s \Delta_s + \eta_q \overline{\Delta}, \quad (4.28)$$

and hence

$$\mathrm{length}\{\widehat{\mathcal{C}}_{1-\delta}(f)\} \le \mathrm{length}\{\mathcal{I}(f)\} + 2\left(B\eta_q + \sum_s q_s \Delta_s + \eta_q \overline{\Delta} \right). \quad (4.29)$$

For the multinomial construction (4.20)–(4.22) with fixed m, J, K and fixed positive error-budget proportions, the excess term is $O_p\{n_Q^{-1/2} + \max_{s,y}(n_{sy}^K)^{-1/2} + \max_s (n_s^r)^{-1/2}\}$ whenever all relevant stratum counts diverge and the stability constants remain bounded.

Corollary 4.7 (Boundary-uniform validity)**.** *Consider any sequence of data-generating laws, possibly depending on the sample size, for which the nuisance confidence systems satisfy (4.3) uniformly over the sequence. Then*

$$\inf_P \mathbb{P}_P\{R_{Q,P}(f) \in \widehat{\mathcal{C}}_{1-\delta}(f)\} \ge 1 - \delta. \quad (4.30)$$

No lower bound on the structural width $\mathrm{length}\{\mathcal{I}_P(f)\}$ is required. Consequently the same construction is valid when the structural identified set has positive width, when it degenerates to a point, and along local sequences where the width tends to zero. If the radii in Theorem 4.6 vanish, the confidence-fiber interval shrinks to the structural interval; in the point-identified case, its length shrinks at the nuisance-estimation rate.

This corollary is the reason we use nuisance-set enlargement rather than a plug-in estimate of the two structural endpoints followed by endpoint standard errors. It is conservative relative to asymptotic confidence intervals for partially identified parameters (Imbens and Manski, 2004; Chernozhukov et al., 2013), but it gives a finite-sample statement and does not require a separate case distinction near point identification.

5 Why Row Space Is a Uniform Criterion

Definition 5.1 (Uniform identification). *Fix a column-stochastic matrix $K \in \mathbb{R}^{J\times K}$ and a loss vector $L \in \mathbb{R}^K$. We say L is uniformly identified through K if, for all $p_1, p_2 \in \Delta_K$,*

$$Kp_1 = Kp_2 \quad \Longrightarrow \quad L^\top p_1 = L^\top p_2. \tag{5.1}$$

Theorem 5.2 (Row space equals uniform identification). *For column-stochastic K and $L \in \mathbb{R}^K$, the following are equivalent: (i) L is uniformly identified through K; (ii) $L \in \mathrm{Row}(K)$; (iii) there exists $g \in \mathbb{R}^J$ such that $K^\top g = L$.*

Thus the row-space test is not wrong; it answers a stronger question. The following example shows that the stronger question can be unnecessarily pessimistic.

Example 5.3 (Same channel, different observed fibers). *Let*

$$K = \begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 1 & 1/2 \end{pmatrix}, \qquad L = (0,0,1)^\top. \tag{5.2}$$

The row space of K consists of vectors $(a, b, (a+b)/2)^\top$, so $L \notin \mathrm{Row}(K)$. If the observed weak law is $r^{(a)} = (1,0)^\top$, then $\Gamma^{(a)} = \{(1,0,0)^\top\}$ and $L^\top p = 0$ is point identified. If instead $r^{(b)} = (1/2, 1/2)^\top$, then

$$\Gamma^{(b)} = \{((1-t)/2, (1-t)/2, t)^\top : 0 \le t \le 1\}, \tag{5.3}$$

so the sharp interval for $L^\top p$ is $[0,1]$. The channel and loss are the same; only the observed fiber changes.

The exact row-space refinement restricts attention to labels that remain possible on the observed fiber.

Definition 5.4 (Active label set). *For a nonempty fiber Γ_x, define*

$$A_x = \{y : \exists p \in \Gamma_x \text{ with } p_y > 0\}. \tag{5.4}$$

Let K_{x,A_x} be the submatrix of K_x with columns in A_x, and let L_{x,A_x} be the corresponding subvector of $L_f(x)$.

Theorem 5.5 (Active-face row-space criterion). *For a fixed observed fiber Γ_x, the following are equivalent: (i) $p \mapsto L_f(x)^\top p$ is constant on Γ_x; (ii) $L_{x,A_x} \in \mathrm{Row}(K_{x,A_x})$. Under Theorem 3.1's assumptions, $R_Q(f)$ is point identified if and only if*

$$L_{x,A_x} \in \mathrm{Row}(K_{x,A_x}) \qquad Q_X\text{-almost surely.} \tag{5.5}$$

6 Estimation After Identification

Point identification is a population property. If a stable corrected loss exists on the full channel or on the active face, it yields a direct weak-label estimator. Let $w = dQ_X/dP_X$.

Proposition 6.1 (Corrected-loss risk rewrite). *Assume covariate shift and overlap. Suppose there is a measurable $g : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ such that*

$$K_x^\top g_x = L_f(x) \tag{6.1}$$

for Q_X-almost every x, and $w(X)g(X,Z)$ is integrable under P_{XZ}. Then

$$R_Q(f) = \mathbb{E}_{P_{XZ}}[w(X)g(X,Z)]. \tag{6.2}$$

The same identity holds if (6.1) is replaced by the active-face equation $K_{x,A_x}^\top g_x = L_{x,A_x}$ on the active clean-label columns.

Given an evaluation sample $(X_i, Z_i)_{i=1}^n \sim P_{XZ}$ independent of any training or model-selection data, the oracle estimator is

$$\widehat{R}_{\mathrm{or}}(f) = \frac{1}{n}\sum_{i=1}^n w(X_i)g(X_i, Z_i). \tag{6.3}$$

Proposition 6.2 (Oracle concentration). *Under Proposition 6.1,* $\mathbb{E}[\widehat{R}_{\mathrm{or}}(f)] = R_Q(f)$ *and*

$$\mathrm{Var}(\widehat{R}_{\mathrm{or}}(f)) = \frac{1}{n}\,\mathrm{Var}_{P_{XZ}}(w(X)g(X,Z)). \tag{6.4}$$

If $|w(X)| \le W$ *and* $|g(X,Z)| \le G$ *almost surely, then with probability at least* $1-\delta$,

$$|\widehat{R}_{\mathrm{or}}(f) - R_Q(f)| \le WG\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{6.5}$$

When corrected losses are nonunique, the following choice minimizes the conditional second moment.

Proposition 6.3 (Least-second-moment corrected loss). *Fix* x *such that* $L_f(x) \in \mathrm{Row}(K_x)$ *and* $r_x(z) > 0$ *for all* z. *Let* $D_x = \mathrm{diag}(r_x)$. *Among all* $g_x \in \mathbb{R}^J$ *satisfying* $K_x^\top g_x = L_f(x)$, *the minimum of* $g_x^\top D_x g_x$ *is attained at*

$$g_x^\star = D_x^{-1}K_x\big(K_x^\top D_x^{-1}K_x\big)^\dagger L_f(x), \tag{6.6}$$

and the minimum value is $L_f(x)^\top(K_x^\top D_x^{-1}K_x)^\dagger L_f(x)$.

This quantity is a stability diagnostic: it can be large even when the risk is identified, for instance when the channel is nearly singular or some weak-label probabilities are small.

For estimated nuisance functions, let $\widehat{w}$ and $\widehat{g}$ be trained on data independent of the evaluation sample and define

$$\widehat{R}_{\mathrm{plug}}(f) = \frac{1}{n}\sum_{i=1}^{n}\widehat{w}(X_i)\widehat{g}(X_i,Z_i). \tag{6.7}$$

Theorem 6.4 (Sample-split plug-in bound). *Condition on the nuisance estimates. Suppose* $|w| \le W$, $|g| \le G$, $|\widehat{w}| \le \overline{W}$, $|\widehat{g}| \le \overline{G}$, *and*

$$\|\widehat{w} - w\|_{L_2(P_X)} \le \varepsilon_w, \qquad \|\widehat{g} - g\|_{L_2(P_{XZ})} \le \varepsilon_g. \tag{6.8}$$

Then with conditional probability at least $1-\delta$,

$$\begin{aligned}|\widehat{R}_{\mathrm{plug}}(f) - R_Q(f)| \le \overline{WG}\sqrt{\frac{2\log(2/\delta)}{n}} \\ + G\varepsilon_w + W\varepsilon_g + \varepsilon_w\varepsilon_g.\end{aligned} \tag{6.9}$$

7 Numerical Illustration and Evaluation Audit

We instantiate Example 5.3. For $r = (r_1, 1-r_1)$, the fiber restricts the loss coordinate p_3 to

$$0 \le p_3 \le \min\{2r_1, 2(1-r_1), 1\}. \tag{7.1}$$

Figure 2 plots this sharp interval. At the boundary weak laws, the interval degenerates even though $L \notin \mathrm{Row}(K)$. At $r = (1/2, 1/2)$, the same channel and loss produce the interval $[0,1]$. The second panel places target mass α on the ambiguous stratum and the remaining mass on a point-identified zero-loss stratum, giving the target interval $[0,\alpha]$.

Coverage and boundary validation. We next check the finite-sample inference layer in a controlled finite-stratum model where the true target risk and structural fiber interval are known. We use the multinomial ℓ_1 construction in (4.20)–(4.22) at nominal level 90%, solve the lifted LP in Proposition 4.4, and repeat each design 1000 times. The three designs are point identified, near the boundary between point and partial identification, and partially identified. Figure 3 shows that empirical coverage is 1000/1000 in all three designs and all three sample sizes. The mean interval width shrinks toward the structural width: from 0.611 to 0.132 in the point-identified design, from 0.624 to 0.160 near the boundary, and from 0.854 to 0.586 in the partially identified design whose structural width is 0.500. In the near-boundary design, the naive plug-in structural interval covers the truth in only 79.9% of repetitions at the smallest sample size, illustrating why the confidence-fiber enlargement is useful near the boundary.

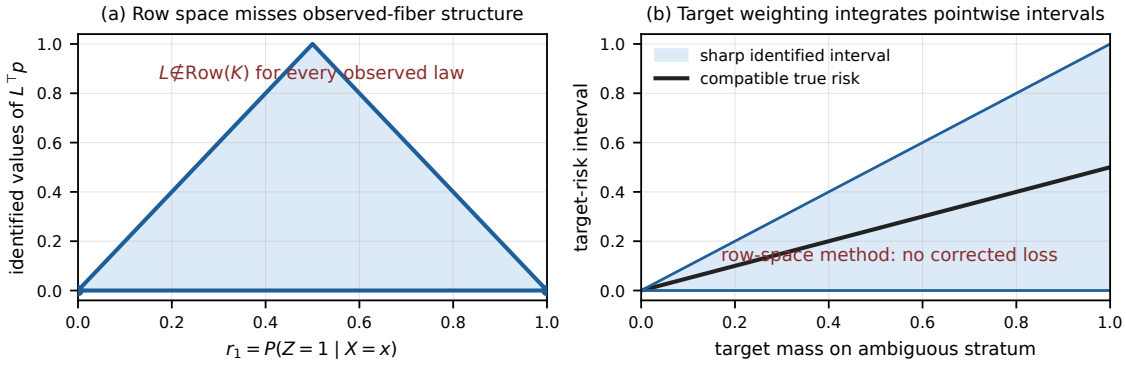

Figure 2: Observed-fiber identification can succeed or fail for the same channel and loss depending on the observed weak law. Left: the sharp pointwise interval for Example 5.3 as $r_1 = P(Z = 1 \mid X = x)$ varies. Right: target weighting integrates pointwise intervals across strata; the compatible true risk lies inside the sharp interval.

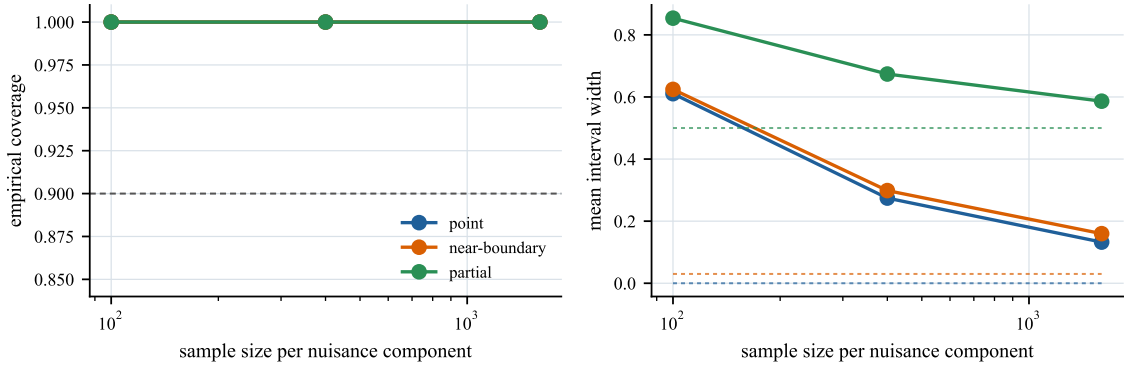

Figure 3: Controlled confidence-fiber validation. Left: empirical coverage at nominal level 90%. Right: mean confidence-fiber interval width, with dotted horizontal lines marking the structural fiber widths. Coverage is stable in point-identified, near-boundary, and partially identified regimes, while widths shrink toward the structural interval.

Audit protocol. The public-data experiments use the same four-step audit. First, fix the predictor, loss, and observable strata before looking at audit clean labels. Second, use a calibration split or sensitivity model to form confidence sets for the weak channel and target-stratum weights. Third, solve the confidence-fiber LP and report the interval together with naive weak-label risk and any row-space residual. Fourth, when clean labels are available only for retrospective validation, reveal them after the interval is formed and report the oracle risk separately. This protocol is meant to test whether weak labels support a risk claim, not to tune the predictor until the interval looks favorable.

Public weak-supervision audit. We next turn the identification result into a small evaluation audit on WRENCH AGNews (Zhang et al., 2021). We train a TF–IDF logistic classifier on the WRENCH training split. The weak source is a coarse majority vote over the provided programmatic weak labels, with weak categories World, Sports, and Business/Tech; abstentions and ties are excluded, giving 63.9% weak-label coverage on the audit half. A disjoint calibration half of the AGNews test split estimates the coarse weak channel K. The audit half supplies only weak labels and base-model strata for the interval calculation; clean audit labels are held out and used only after the fact as an oracle check. The deployed predictor is fixed by collapsing base Sci/Tech predictions to Business, creating a realistic failure mode that coarse weak labels may miss.

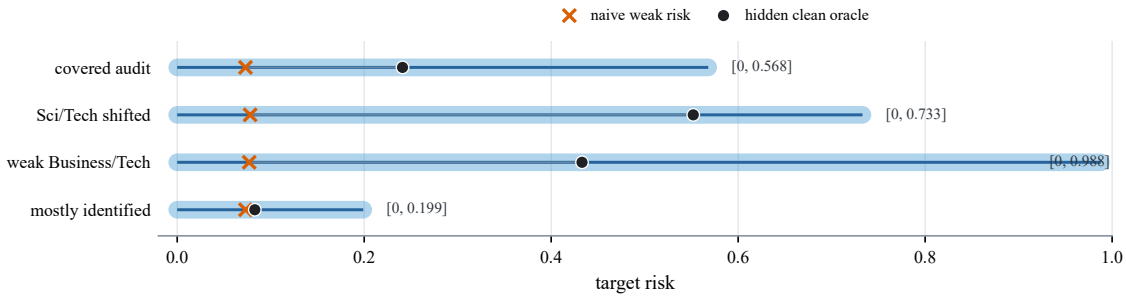

Figure 4: Interval summary for the WRENCH AGNews audit. Blue bars are calibrated 90% confidence-fiber intervals, orange crosses are naive weak-label risks, and black dots are hidden clean-oracle risks used only after the audit calculation. The figure makes the qualitative pattern in Table 1 visible: naive weak-label evaluation can substantially understate risk, while the interval contains the oracle and shrinks in the mostly identified target law.

Table 1: Public weak-supervision audit on WRENCH AGNews. Target weights are over base-predicted strata World, Sports, Business, and Sci/Tech. The weak Business/Tech row is a natural observable subgroup, not a hand-designed target shift. Coarse programmatic weak labels can substantially understate the clean risk of the fixed deployed predictor; the calibrated confidence-fiber interval contains the hidden clean oracle and narrows when most target mass lies on nearly identified strata.

scenario	target weights q	naive	90% conf. fiber	oracle	width	covered
covered audit	$(0.236, 0.281, 0.257, 0.226)$	0.073	$[0.000, 0.568]$	0.241	0.568	yes
Sci/Tech shifted	$(0.150, 0.150, 0.100, 0.600)$	0.078	$[0.000, 0.733]$	0.552	0.733	yes
weak Business/Tech	$(0.036, 0.023, 0.503, 0.438)$	0.077	$[0.000, 0.988]$	0.433	0.988	yes
mostly identified	$(0.450, 0.450, 0.050, 0.050)$	0.073	$[0.000, 0.199]$	0.083	0.199	yes

Figure 4 and Table 1 report calibrated 90% confidence-fiber intervals obtained by replacing the earlier sensitivity tolerance with the multinomial radii in (4.20)–(4.22) and solving the lifted LP in Proposition 4.4. For empirical target laws, the target-weight set $\mathcal{Q}$ is also enlarged; for hand-specified target shifts, q is treated as fixed and only the weak law and channel are enlarged. Appendix A gives a 50-split AGNews stability check, a controlled digits check, and a calibrated WRENCH TREC audit. Together, Figure 4 and Table 1 answer the audit question in three ways. First, naive weak-label evaluation can be badly overconfident: in the Sci/Tech-shifted law, it reports risk 0.078 while the hidden clean oracle is 0.552, and the confidence-fiber interval is $[0.000, 0.733]$. Second, the same failure appears in a natural observable subgroup: conditioning on coarse weak label Business/Tech gives naive risk 0.077, oracle risk 0.433, and interval $[0.000, 0.988]$. Third, the interval narrows when target mass concentrates on nearly identified strata: the mostly identified law has width 0.199 and oracle 0.083. The clean labels used for the oracle column are never used to form the intervals.

8 Related Work

Noisy and weak labels. Loss-correction methods for label noise construct unbiased or consistent losses from corrupted labels when the corruption process is known or estimable (Natarajan et al., 2013; Patrini et al., 2017; Menon et al., 2015; Scott et al., 2013). Classical crowd-label models (Dawid and Skene, 1979) and modern weak-supervision systems (Ratner et al., 2020) provide motivations for known or modeled weak-label channels. These literatures usually ask for a corrected loss, a learned classifier, or a latent-label model. Our question is narrower and more diagnostic: for a fixed deployed predictor and an actually observed weak-label law, what clean target-risk statements are identified? The answer shows that the corrected-loss equation is sufficient and necessary for uniform correction, but can be strictly stronger than observed-specific risk identification.

Covariate shift and off-distribution evaluation. Importance weighting under covariate shift is a standard route to target-risk estimation (Shimodaira, 2000; Sugiyama et al., 2007; Bickel et al., 2009). We combine the covariate-shift identity with weak-label fibers: the target covariate law determines how pointwise identified quantities are weighted, while the observed weak-label fiber determines whether the clean conditional risk is identified at each x.

Partial identification and inference. Partial identification asks what is implied by observations and assumptions when point identification fails (Manski, 2003; Imbens and Wooldridge, 2009). Inference for interval-identified parameters must account for endpoint sampling variation and for the boundary between point and partial identification (Imbens and Manski, 2004; Chernozhukov et al., 2013). Discrete measurement-error models also optimize latent distributions subject to linear constraints (Molinari, 2008; Finkelstein et al., 2021). Our results can be viewed as a weak-label, covariate-shifted target-risk specialization of this broad viewpoint, but the specialization changes the practical object: the loss vector, channel, observed weak law, and target covariate law together determine whether a black-box evaluation is a point, a sharp interval, or a nonidentification warning. The row-space comparison further separates uniform corrected-loss estimability from observed-law identification.

Our inference layer follows a set-enlargement route: if the nuisance sets cover the weak law, channel, and target weights, the resulting risk interval covers the clean target risk without estimating which endpoint regime applies. This is why the guarantee remains valid at the boundary between point and partial identification. Recent work on weak-supervision performance evaluation also uses partial identification and convex optimization to bound metrics without ground-truth labels (Polo et al., 2024); our contribution is to isolate the observed fiber for covariate-shifted clean target risk, prove the exact relation to row-space correction, and provide finite-sample confidence fibers under estimated or sensitivity-modeled channels.

9 Limitations and Broader Impact

The main assumptions are finite labels, a fixed predictor, exact covariate shift, overlap, and a valid source of channel information. The confidence-fiber results no longer require the weak-label channel to be known exactly, but they do require either calibration data that deliver joint confidence sets or a defensible sensitivity set. They do not claim that $x \mapsto (K_x, r_x)$ can be estimated uniformly over unrestricted continuous covariate spaces without smoothing or stratification. They also do not justify reusing the same weak-label audit sample to select the predictor, target strata, or weak source without sample splitting or post-selection analysis.

Approximate covariate shift, support mismatch, adaptive data-dependent predictor selection, multiple dependent weak sources without a validated label model, and fully nonparametric continuous-X confidence bands remain important extensions. Some intervals will be wide; this is a feature of the audit rather than a failure of the LP, because a wide interval records information that weak labels do not identify. The practical benefit is conservative: the method can prevent overconfident clean-risk claims from weak labels. The practical risk is misuse: if the channel confidence set, shift assumption, or overlap diagnostics are invalid, the interval can be misleading. We recommend reporting the channel source, overlap diagnostics, nuisance-estimation protocol, confidence level or sensitivity radius, stratum counts, row-space residuals, and fiber widths alongside any point estimate.

10 Reproducibility Statement

All theoretical claims are stated with their assumptions and proved in Appendix B. The main numerical illustration is population-level and generated by closed-form or linear-program endpoints. The source bundle accompanying the submission contains scripts that regenerate the confidence-fiber coverage simulation, calibrated WRENCH AGNews audit, AGNews split-stability summaries, controlled digits check, and calibrated WRENCH TREC audit. The public case studies use the sklearn digits toy dataset and the public WRENCH AGNews and TREC tasks. Clean audit labels in the WRENCH experiments are used only for retrospective oracle checks after intervals have been computed from weak labels, calibration data, and target covariate information.

References

Steffen Bickel, Michael Brueckner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10:2137–2155, 2009.

Victor Chernozhukov, Sokbae Lee, and Adam M. Rosen. Intersection bounds: Estimation and inference. *Econometrica*, 81(2):667–737, 2013.

Alexander P. Dawid and Allan M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C*, 28(1):20–28, 1979.

Noam Finkelstein, Roy Adams, Suchi Saria, and Ilya Shpitser. Partial identifiability in discrete data with measurement error. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence*, pages 1798–1808, 2021.

Leo A. Goodman. On simultaneous confidence intervals for multinomial proportions. *Technometrics*, 7(2):247–254, 1965.

Alan J. Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49(4):263–265, 1952.

Guido W. Imbens and Charles F. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

Guido W. Imbens and Jeffrey M. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86, 2009.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

Aditya Krishna Menon, Brendan van Rooyen, Cheng Soon Ong, and Robert C. Williamson. Learning from corrupted binary labels via class-probability estimation. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 125–134, 2015.

Francesca Molinari. Partial identification of probability distributions with misclassified data. *Journal of Econometrics*, 144(1):81–117, 2008.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep K. Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems 26*, pages 1196–1204, 2013.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1944–1952, 2017.

Felipe Maia Polo, Subha Maity, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. Weak supervision performance evaluation via partial identification. In *Advances in Neural Information Processing Systems 37*, 2024.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Re. Snorkel: rapid training data creation with weak supervision. *The VLDB Journal*, 29:709–730, 2020.

Clayton Scott, Gilles Blanchard, and Gregory Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In *Proceedings of the 26th Annual Conference on Learning Theory*, pages 489–511, 2013.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Mueller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdú, and Marcelo J. Weinberger. Inequalities for the L_1 deviation of the empirical distribution. Hewlett–Packard Laboratories Technical Report HPL-2003-97R1, 2003.

Jieyu Zhang, Yue Yu, Yinghao Li, Yujing Wang, Yaming Yang, Mao Yang, and Alexander Ratner. WRENCH: A comprehensive benchmark for weak supervision. *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.

Table 2: Calibrated AGNews split stability over 50 stratified random calibration/audit splits with calibration fraction 0.5. Coverage is the fraction of splits in which the 90% confidence-fiber interval contains the hidden clean oracle. The naive weak risk is stable but can substantially understate the clean oracle risk; interval widths remain narrow only in the mostly identified target law.

scenario	coverage	naive	oracle	mean width	sd width
covered audit	1.000	0.071	0.241	0.564	0.006
Sci/Tech shifted	1.000	0.076	0.551	0.731	0.003
weak Business/Tech	1.000	0.074	0.436	0.993	0.003
mostly identified	1.000	0.070	0.083	0.193	0.008

A Additional Evaluation Details

Confidence-fiber simulation. The coverage experiment in Figure 3 uses three strata, three clean labels, and two weak labels. In every stratum the weak channel has columns

$$K(\cdot,1) = (1,0)^\top, \qquad K(\cdot,2) = (0,1)^\top, \qquad K(\cdot,3) = (1/2,1/2)^\top. \tag{A.1}$$

The clean posteriors are $p_1 = (1,0,0)$, $p_2 = (0,1,0)$, and $p_3 = (0.25,0.25,0.50)$, and the loss vector in every stratum is $L = (0,0,1)$. The point-identified, near-boundary, and partially identified target weights are respectively

$$q = (0.50,0.50,0), \qquad q = (0.485,0.485,0.030), \qquad q = (0.25,0.25,0.50). \tag{A.2}$$

For each sample size $n \in \{100,400,1600\}$, we draw independent multinomial samples for the target weights, each stratum weak law, and each channel column, build the level-90% nuisance confidence system using the radii in (4.18) with equal error-budget allocation, and solve the lifted LP in Proposition 4.4. The reported coverage and widths average over 1000 repetitions. The plug-in comparison uses the empirical $(\widehat{q},\widehat{K},\widehat{r})$ as if known and then solves the structural fiber LP.

AGNews split stability. To check that the public audit is not an artifact of one calibration/audit split, we repeat the calibrated AGNews audit over 50 stratified random splits and three calibration fractions $0.3, 0.5, 0.7$. Table 2 reports the middle fraction; the other two fractions have the same 50/50 oracle coverage in every scenario, with mean widths within 0.02 of the displayed values. The plug-in structural fiber is infeasible in all runs because the empirical weak law and empirical channel need not be exactly compatible, which is precisely the finite-sample problem addressed by confidence fibers.

Controlled digits check. We include a small semi-real check using the sklearn handwritten digits data restricted to clean digits $0,1,2$. A three-class logistic model is trained on clean source labels and then fixed. The deployed predictor intentionally maps base prediction 2 to class 0, so the clean error is concentrated in the stratum where the base model predicts 2. Weak labels are generated through the known two-output channel

$$K = \begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 1 & 1/2 \end{pmatrix}, \tag{A.3}$$

which labels clean digits 0 and 1 correctly and maps clean digit 2 to weak labels 0 and 1 with equal probability. We stratify by the base-model prediction and impose target shifts $q = (q_0, q_1, q_2)$ over these observable strata. Clean labels in the held-out set are used only as a hidden oracle to check whether the interval contains the true clean risk.

Table 3 isolates the basic mechanism in a setting with real covariates and an exact known channel. The full row-space test is too pessimistic for target laws supported on active singleton fibers: it rejects although the observed-fiber interval is $[0,0]$. For target laws with nonzero ambiguous mass, the interval widens exactly on the nonidentified portion while retaining the point-identified mass in the remaining strata.

Second public weak-supervision task. We also run the same calibrated audit logic on WRENCH TREC, a six-class question classification task. We train a TF–IDF logistic classifier on the WRENCH training split, aggregate programmatic weak labels into three coarse weak categories, and fix a deployed predictor by

Table 3: Controlled weak-label case study on digits 0/1/2. The full row-space correction test rejects in every row. Observed-fiber analysis is nevertheless point identified when the target puts no mass on the ambiguous stratum, and otherwise returns a sharp interval that contains the hidden clean oracle.

scenario	target weights q	naive	full row	fiber interval	oracle	width	id. mass
identified target	$(0.500, 0.500, 0.000)$	0.000	reject	$[0.000, 0.000]$	0.000	0.000	1.000
source-like	$(0.331, 0.338, 0.331)$	0.165	reject	$[0.165, 0.331]$	0.331	0.165	0.669
class-2 shifted	$(0.200, 0.200, 0.600)$	0.300	reject	$[0.300, 0.600]$	0.600	0.300	0.400
mostly identified	$(0.450, 0.450, 0.100)$	0.050	reject	$[0.050, 0.100]$	0.100	0.050	0.900

Table 4: Additional calibrated WRENCH TREC audit. The hidden clean oracle lies inside every confidence-fiber interval, but the wide intervals show that the available weak labels do not identify a precise clean risk. The HUM-focused target also shows why forcing a pseudo-corrected point estimate is unsafe when the row-space equation has nonzero residual: the projected estimate is negative.

scenario	naive	90% conf. fiber	oracle	pseudo	row resid.
test audit	0.253	$[0.000, 0.993]$	0.278	0.575	0.753
NUM shifted	0.220	$[0.000, 0.975]$	0.636	0.589	0.799
HUM focused	0.129	$[0.000, 0.769]$	0.110	-0.084	0.255

collapsing base NUM predictions to LOC. The validation split estimates the coarse channel, and the test split is used for the weak-label audit; clean test labels again only report the oracle. Table 4 reports calibrated 90% confidence-fiber intervals. Here the intervals are wide, especially for the test and NUM-shifted target laws, which is an informative negative audit: the coarse weak labels and calibration split do not support a precise clean-risk claim. The pseudo-corrected column is obtained by least-squares projection onto the row space and is not a valid corrected-loss estimator when the displayed row residual is positive.

B Proofs

B.1 Proof of Theorem 3.1

Let p_x be any clean conditional law compatible with the observations. Then $p_x \in \Gamma_x$ for P_X-almost every x, and by covariate shift the same conditional is used under the target law. Hence

$$\underline{r}_f(x) \le L_f(x)^\top p_x \le \overline{r}_f(x) \tag{B.1}$$

for Q_X-almost every x, and integration gives containment in (3.3).

For sharpness, let p_x^- and p_x^+ be measurable minimizer and maximizer selections from Γ_x on the $(P_X + Q_X)$-support. For $\lambda \in [0,1]$, set $p_x^\lambda = (1-\lambda)p_x^- + \lambda p_x^+$. Convexity of Γ_x gives $p_x^\lambda \in \Gamma_x$, so $K_x p_x^\lambda = r_x$ and the observable source weak-label law is unchanged. Use P_X as the source covariate law, Q_X as the target covariate law, and the same clean conditional p_x^λ in source and target. Exact covariate shift holds by construction, and the risk is

$$(1-\lambda)\int \underline{r}_f(x)\, dQ_X(x) + \lambda \int \overline{r}_f(x)\, dQ_X(x). \tag{B.2}$$

Thus every point in the interval is attainable by an observationally equivalent clean law. Since the integrand $\overline{r}_f - \underline{r}_f$ is nonnegative, the interval degenerates exactly when it is zero Q_X-almost surely, which is equivalent to constancy of the loss functional on Γ_x.

B.2 Proof of Corollary 3.2

On each stratum, the pointwise lower and upper values in Theorem 3.1 are constant and equal to the displayed LP optima. Integrating over Q_X reduces to finite weighted sums. The formula for A_s is exactly the definition of activity: a label is active if and only if some feasible posterior assigns it positive mass.

B.3 Proof of Corollary 3.3

Because K_x is column stochastic and $r_x \in \Delta_J$, the constraints $p \in \Delta_K$ and $K_x p = r_x$ are equivalent, for feasible points, to $p \ge 0$ and $K_x p = r_x$. The lower endpoint is

$$\min_{p \ge 0} L_f(x)^\top p \quad \text{subject to} \quad K_x p = r_x, \tag{B.3}$$

whose dual is (3.8). The upper endpoint follows by applying the same argument to $-L_f(x)$. Strong LP duality applies because the fiber is nonempty and compact. If a and b satisfy the displayed inequalities, weak duality gives $a^\top r_x \le L_f(x)^\top p \le b^\top r_x$ for all $p \in \Gamma_x$; equality of the bounds certifies constancy.

B.4 Proof of Corollary 3.4

Take measurable minimizer and maximizer selections p_x^- and p_x^+. They both satisfy $K_x p_x^\pm = r_x$ for P_X-almost every x, hence induce the same observable source law. In both models use the same P_X and Q_X and set the target conditional equal to the selected source conditional. Their target risks differ by the positive integrated width.

B.5 Proof of Proposition 3.5

The observable data distribution is identical under the two clean laws. Hence $\widehat{R}$ has the same distribution in the two models. For every realized value a,

$$(a-R_1)^2 + (a-R_2)^2 = 2\left(a - \frac{R_1+R_2}{2}\right)^2 + \frac{(R_1-R_2)^2}{2} \ge \frac{\Delta^2}{2}. \tag{B.4}$$

Taking expectations gives the mean-squared-error bound. For the interval claim, the two coverage events are evaluated under the same observable distribution. A union bound gives $\mathbb{P}(R_1 \in C, R_2 \in C) \ge 1 - 2\delta$. On this event, C has length at least $|R_1 - R_2|$.

B.6 Proof of Theorem 4.3

On the joint nuisance coverage event (4.3), the true target weights satisfy $q \in \mathcal{Q}$, the true weak laws satisfy $r_s \in \mathcal{R}_s$, and the true channel columns satisfy $K_s(\cdot, y) \in \mathcal{K}_{sy}$. The true clean posterior p_s obeys $\sum_y p_{sy} K_s(\cdot, y) = r_s$, so $p_s \in \Gamma_s^{\mathcal{C}}$ for every stratum with $q_s > 0$. Therefore the true risk

$$R_Q(f) = \sum_s q_s L_s^\top p_s \tag{B.5}$$

is feasible for the optimization defining (4.5)–(4.6), which implies $R_Q(f) \in \widehat{\mathcal{C}}_{1-\delta}(f)$ on that event. Taking probabilities gives (4.8). For the sharpness statement relative to realized nuisance sets with a nonempty compatible class, every compatible model supplies feasible variables in (4.5)–(4.6), so its risk lies between the endpoints. Conversely, any feasible tuple (q, p_s, k_{sy}, r_s), with p_s required only when $q_s > 0$, defines a clean conditional law on the target-relevant strata and a weak-label law generated by $\sum_y p_{sy} k_{sy} = r_s$, hence realizes the displayed objective value. The smallest closed interval containing all such risks is therefore the interval between the infimum and supremum.

B.7 Proof of Proposition 4.4

First suppose $q_s > 0$ and $p_s \in \Gamma_s^{\mathcal{C}}$ is generated by some $k_{sy} \in \mathcal{K}_{sy}$ and $r_s \in \mathcal{R}_s$. Set $\mu_{sy} = q_s p_{sy}$, $u_{sy} = q_s p_{sy} k_{sy}$, and $v_s = q_s r_s$. The polyhedral constraints for k_{sy} and r_s imply the perspective constraints (4.16) and (4.17); if $q_s = 0$, set $\mu_{sy} = u_{sy} = v_s = 0$. Thus every feasible point of (4.5)–(4.6) maps to a feasible point of the lifted LP with the same objective value.

Conversely, take any feasible solution of the lifted LP. If $q_s > 0$, define $p_{sy} = \mu_{sy}/q_s$ and $r_s = v_s/q_s$. For each y with $\mu_{sy} > 0$, define $k_{sy} = u_{sy}/\mu_{sy}$; if $\mu_{sy} = 0$, choose any element of the nonempty polytope $\mathcal{K}_{sy}$. The perspective inequalities imply $k_{sy} \in \mathcal{K}_{sy}$ and $r_s \in \mathcal{R}_s$, and $v_s = \sum_y u_{sy}$ gives $\sum_y p_{sy} k_{sy} = r_s$. Thus $p_s \in \Gamma_s^{\mathcal{C}}$. If $q_s = 0$, the stratum contributes zero to the objective and no posterior is needed under the convention in Definition 4.2. The objective equals $\sum_{s,y} L_{sy} \mu_{sy} = \sum_s q_s L_s^\top p_s$, so the lifted LP has exactly the same optimal value.

B.8 Proof of the multinomial construction

For a multinomial distribution on $d \geq 2$ categories, the inequality of Weissman et al. (2003) gives

$$\mathbb{P}\{\|\widehat{p} - p\|_1 > \epsilon\} \leq (2^d - 2) \exp(-n\epsilon^2/2) \tag{B.6}$$

for $n \geq 1$. For $d = 1$ the simplex is a singleton, so the deviation is zero. The definition (4.18) makes the conditional failure probability at most the allocated error level whenever the realized count is positive; if a relevant cell count is zero, or if an error budget is set to zero, the radius convention makes the confidence set the whole simplex. Applying these conditional bounds to the target-stratum empirical distribution, to each stratum weak-label empirical distribution, and to each calibration channel column, then averaging over random counts and using (4.19) with a union bound, yields the joint coverage event (4.3). The ℓ_1 constraints are polyhedral because $\|a - b\|_1 \leq \rho$ is equivalent to the existence of nonnegative slack variables t_j with $-t_j \leq a_j - b_j \leq t_j$ and $\sum_j t_j \leq \rho$.

B.9 Proof of Theorem 4.6

Fix a stratum s and let

$$J_s = \{L_s^\top p : p \in \Delta_K,\ K_s p = r_s\}, \qquad J_s^{\mathcal{C}} = \{L_s^\top p : p \in \Gamma_s^{\mathcal{C}}\}.$$

Every $p \in \Gamma_s^{\mathcal{C}}$ belongs to a nonempty fiber generated by some $(\widetilde{K}_s, \widetilde{r}_s)$ selected from the realized nuisance sets. Assumption 4.5 and (4.25)–(4.26) give a true-fiber point p' with

$$\|p - p'\|_1 \leq H_s \left(\eta_s^r + \max_y \eta_{sy}^K \right).$$

Conversely, because the realized confidence fiber is nonempty, choose one nonempty perturbed fiber represented in $\Gamma_s^{\mathcal{C}}$; the symmetric part of the Hausdorff bound in Assumption 4.5 gives, for every true-fiber point p', a point $p \in \Gamma_s^{\mathcal{C}}$ obeying the same display. Since $|L_s^\top(p - p')| \le B\|p - p'\|_1$, we have

$$d_H(J_s^{\mathcal{C}}, J_s) \le \Delta_s.$$

Thus each confidence-fiber stratum value $a_s \in J_s^{\mathcal{C}}$ can be paired with a structural value $b_s \in J_s$ satisfying $|a_s - b_s| \le \Delta_s$, and conversely each $b_s \in J_s$ can be paired with such an a_s.

For any $\widetilde{q} \in \mathcal{Q}$ and any paired stratum values a_s, b_s,

$$\left|\sum_s \widetilde{q}_s a_s - \sum_s q_s b_s\right| \le B\|\widetilde{q} - q\|_1 + \sum_s \widetilde{q}_s \Delta_s \le B\eta_q + \sum_s q_s \Delta_s + \eta_q \overline{\Delta}. \tag{B.7}$$

The same inequality, using any element of the nonempty set $\mathcal{Q}$, pairs structural aggregate values with confidence aggregate values. Taking infima and suprema over feasible selections gives the Hausdorff bound (4.28); the width bound (4.29) follows because expanding each endpoint by at most the right-hand side expands length by at most twice that amount. The stochastic rate follows by substituting the multinomial radii (4.18) when the relevant counts diverge.

B.10 Proof of Corollary 4.7

The coverage statement in Theorem 4.3 uses only the joint nuisance coverage event and the feasibility of the true posterior; it never assumes that the structural lower and upper endpoints are separated. Therefore the same proof applies uniformly over any class for which (4.3) holds uniformly. The convergence and point-identified length statements are immediate consequences of Theorem 4.6 when the radii vanish.

B.11 Proof of Theorem 5.2

The equivalence between $L \in \mathrm{Row}(K)$ and existence of g with $K^\top g = L$ is the definition of the row space. If $L = K^\top g$ and $Kp_1 = Kp_2$, then $L^\top p_1 = g^\top K p_1 = g^\top K p_2 = L^\top p_2$.

Conversely, suppose uniform identification holds. Let $h \in \mathrm{Null}(K)$. Since K is column stochastic, $\mathbf{1}_K^\top h = \mathbf{1}_J^\top K h = 0$. Choose an interior point $p_0 \in \Delta_K$ and $\epsilon > 0$ small enough that $p_0 \pm \epsilon h \in \Delta_K$. Then $K(p_0 + \epsilon h) = K(p_0 - \epsilon h)$, so uniform identification implies $L^\top h = 0$. Thus L is orthogonal to $\mathrm{Null}(K)$, and finite-dimensional linear algebra gives $L \in \mathrm{Null}(K)^\perp = \mathrm{Row}(K)$.

B.12 Proof of Theorem 5.5

Fix x and suppress it from notation. By definition, every $p \in \Gamma$ has support contained in A, and for every $y \in A$ some feasible point has positive mass on y. Averaging these feasible points yields $p^\star \in \Gamma$ with $p^\star_y > 0$ for all $y \in A$.

If $L_A = K_A^\top g$ for some g, then for every $p \in \Gamma$,

$$L^\top p = L_A^\top p_A = g^\top K_A p_A = g^\top r, \tag{B.8}$$

which is constant. Conversely, suppose $L^\top p$ is constant on Γ. Let $h \in \mathrm{Null}(K_A)$. Since K_A is column stochastic, $\mathbf{1}_A^\top h = 0$. For sufficiently small $\epsilon > 0$, $p_A^\star \pm \epsilon h$ remain in the active simplex, and the corresponding full vectors remain in Γ. Constancy gives $L_A^\top h = 0$ for every $h \in \mathrm{Null}(K_A)$, hence $L_A \in \mathrm{Null}(K_A)^\perp = \mathrm{Row}(K_A)$.

B.13 Proof of Proposition 6.1

By integrability and finiteness of $\mathcal{Z}$,

$$\mathbb{E}_{P_{XZ}}[w(X)g(X,Z)] = \int w(x) g_x^\top r_x \, dP_X(x) = \int g_x^\top r_x \, dQ_X(x). \tag{B.9}$$

For Q_X-almost every x, $r_x = K_x p_x$, so $g_x^\top r_x = g_x^\top K_x p_x = L_f(x)^\top p_x$. Covariate shift gives (6.2). If only the active-face equation holds, all feasible posteriors place zero mass outside A_x, so the same calculation holds with K_{x,A_x} and L_{x,A_x}.

B.14 Proof of Proposition 6.2

Unbiasedness follows from Proposition 6.1. The variance formula follows from independence. If $|w(X)g(X,Z)| \le WG$, Hoeffding's inequality applied to variables in an interval of length at most $2WG$ gives the displayed concentration bound.

B.15 Proof of Proposition 6.3

Suppress x and write $D = \mathrm{diag}(r)$. Since $r_z > 0$, D is positive definite. The problem is

$$\min_g g^\top D g \quad \text{subject to} \quad K^\top g = L. \tag{B.10}$$

With $u = D^{1/2} g$, this becomes $\min_u \|u\|_2^2$ subject to $K^\top D^{-1/2} u = L$. The minimum-norm feasible solution is $u^\star = D^{-1/2} K (K^\top D^{-1} K)^\dagger L$, so $g^\star = D^{-1} K (K^\top D^{-1} K)^\dagger L$. The squared norm of $u^\star$ gives the displayed minimum value.

B.16 Proof of Theorem 6.4

Let P_n be the empirical measure of the evaluation sample and P the law of (X, Z). Conditional on nuisance estimates, the evaluation sample remains i.i.d. from P and

$$\widehat{R}_{\mathrm{plug}} - R_Q(f) = (P_n - P)(\widehat{w}\widehat{g}) + P(\widehat{w}\widehat{g} - wg). \tag{B.11}$$

The empirical term is bounded by Hoeffding because $|\widehat{w}\widehat{g}| \le \overline{WG}$. For the bias term,

$$\widehat{w}\widehat{g} - wg = (\widehat{w} - w)g + w(\widehat{g} - g) + (\widehat{w} - w)(\widehat{g} - g). \tag{B.12}$$

Cauchy–Schwarz gives the three bounds $G\varepsilon_w$, $W\varepsilon_g$, and $\varepsilon_w \varepsilon_g$. Combining them proves the theorem.