Federated Measurement of Demographic Disparities from Quantile Sketches

Anonymous authors
Paper under double-blind review

Abstract

Fairness audits are often defined for a target population, while the score data needed to conduct them are held by separate institutions, jurisdictions, or clients. This fragmentation creates a local–global mismatch: a scoring rule may appear balanced within each silo but display substantial demographic disparity after aggregation, or conversely hide client-level discrepancies in a pooled summary. We study this problem for score-level demographic parity under federated data constraints. We define the population disparity as a Wasserstein–Fréchet variance of sensitive-group score distributions and express the same target through client-level conditional laws and group-specific mixing weights. This representation shows why local audits, and averages of local audits, are generally not valid proxies for the population audit. We propose a one-shot quantile-sketch protocol in which each silo releases only subgroup counts and k empirical score quantiles per sensitive group. The server uses these summaries to reconstruct group-mixture score distributions and estimate the global disparity. For the squared Wasserstein distance, the same summaries yield an ANOVA-style decomposition separating mixture/composition effects, barycentric cross-silo heterogeneity, and their interaction. We prove an $O(1/k)$ deterministic discretization bound and finite-sample guarantees based on quantile concentration. Experiments on synthetic data, `COMPAS`, and `ACSIncome` show that local audits can misrepresent population-level disparity, while moderate sketch sizes closely recover the centralized benchmark, including with natural state-level clients in `ACSIncome`.

1 Introduction

Machine-learning systems are increasingly deployed across institutions, agencies, jurisdictions, and other data-holding entities. Hospitals, insurers, courts, banks, and public agencies often collect and govern data locally, under legal, privacy, or organizational constraints that prevent raw observations from being pooled. Federated learning addresses one version of this problem by enabling collaborative model training without centralizing data McMahan et al. (2017); Kairouz et al. (2021). Fairness auditing raises a related but distinct problem. The scoring rule may already be fixed, but the fairness criterion to be audited is often a population-level property, while the evidence available to the auditor is distributed across silos.

This distinction matters because local audits are not, in general, population audits. A scoring rule may appear demographically balanced within each silo while producing a substantial disparity after aggregation. Conversely, a global audit may hide large client-level discrepancies. The issue is structural, not only statistical. For a sensitive group s, the population score law is a mixture of silo-specific score laws, with weights determined by how that group is distributed across silos. Hence the population-level audit depends both on within-silo score distributions and on group-specific silo composition. Local metrics, or averages of local metrics, need not estimate the population quantity of interest.

This paper studies this measurement problem for score-level demographic parity under federated data constraints. We do not propose a training algorithm, a post-processing rule, or a mitigation strategy. Instead, we ask: given a deployed scoring rule and data held across silos, what population-level distributional disparity can be inferred from limited client-side summaries, and how can one diagnose why local and

global audits disagree? Our answer is a one-shot audit protocol based on quantile sketches. Each silo sends, for each sensitive group, only its subgroup count and a fixed number of empirical score quantiles. From these summaries, the server reconstructs group-mixture score distributions, estimates a population-level demographic-disparity functional, and, for the squared Wasserstein distance, decomposes the global gap into interpretable local–global components.

1.1 Why local audits can fail

Two mechanisms drive the discrepancy between local and population demographic-parity audits.

Selection and composition effects. Sensitive groups need not be distributed across silos in the same way. Access patterns, referral rules, institutional specialization, jurisdiction, geography, or historical sampling processes may induce different client compositions across groups. Even if score distributions are similar within each silo, aggregating heterogeneous silos with group-specific weights can create, amplify, attenuate, or mask a global disparity. This is a distributional analogue of Simpson's paradox Julious & Mullee (1994); Prost et al. (2022): conclusions drawn within strata need not agree with the conclusion obtained after pooling.

Cross-silo score heterogeneity. Score distributions may also vary across silos within the same sensitive group. Such heterogeneity can arise from case mix, measurement conventions, data collection practices, local deployment contexts, or institution-specific versions of a scoring rule. This is related to hidden stratification, where subgroup-specific behavior remains invisible under coarse evaluations Oakden-Rayner et al. (2020); Tosh & Hsu (2022). In medical risk prediction, for example, a model may behave differently across health systems or patient populations Obermeyer et al. (2019); in criminal justice, the `COMPAS` scores illustrate how risk-score distributions may differ substantially across demographic groups Larson et al. (2016); Chouldechova (2017).

A minimal example illustrates the issue. Let $Z = \mathbf{1}\{X > 0\}$ be a binary score or decision, and let

$$X \mid S = 0 \sim \mathcal{N}(-\mu, 1), \qquad X \mid S = 1 \sim \mathcal{N}(\mu, 1), \qquad \mu > 0.$$

Suppose that individuals with $X < 0$ are assigned to one silo and individuals with $X > 0$ to another. Within each silo, Z is constant, so demographic parity holds locally. Globally, however,

$$\mathbb{P}(Z = 1 \mid S = 0) < \frac{1}{2} < \mathbb{P}(Z = 1 \mid S = 1),$$

so demographic disparity appears after aggregation. The example is intentionally simple: it shows that local parity does not imply population parity when silo assignment is informative.

1.2 Distributional auditing from quantiles

Operational fairness audits often focus on decisions, but decisions are frequently obtained by thresholding scores. Auditing only thresholded outcomes can hide distributional shifts in the underlying scores and can be sensitive to downstream thresholds that differ across clients or change over time. We therefore audit demographic parity at the level of score distributions: score-level demographic parity holds when the score law is independent of the sensitive attribute.

Our central object is a population-level disparity functional U_p, defined as a Wasserstein–Fréchet variance of the sensitive-group score laws. In the centralized case, $U_p = 0$ if and only if all sensitive groups have the same score distribution. In the federated case, the same population-level object can be expressed through client-level conditional score laws and group-specific mixing weights. This representation makes the local–global mismatch explicit: the global law for a sensitive group is a mixture over silos, whereas Wasserstein distances are governed by quantile functions, and quantiles do not commute with mixtures.

This observation leads to a simple communication interface. In one dimension, Wasserstein distances are determined by quantile functions. A small number of quantiles per sensitive group and silo can therefore approximate the population audit target. The proposed protocol is a minimal-disclosure audit layer: raw records and individual scores need not leave the clients, but the transmitted summaries are not a formal

privacy guarantee. In privacy-sensitive deployments, they should be combined with safeguards such as minimum cell-size rules, suppression or coarsening of small cells, secure aggregation, or calibrated noise mechanisms Bonawitz et al. (2017).

1.3 Relation to prior work

Our setting is related to federated learning, but the goal is different. Standard federated learning focuses on training models from decentralized data McMahan et al. (2017); Kairouz et al. (2021); fairness-aware variants study how to allocate performance or enforce fairness constraints across clients during training Mohri et al. (2019); Li et al. (2020); Du et al. (2021). We instead treat the scoring rule as fixed and study how to audit a population-level fairness criterion when the score data cannot be centralized.

The distributional target builds on the one-dimensional optimal-transport representation of Wasserstein distances and barycenters Villani (2009); Santambrogio (2015); Agueh & Carlier (2011). Wasserstein barycenters have also been used in fairness, for example in fair regression through distributional post-processing Chzhen et al. (2020). Our contribution is different: we use Wasserstein–Fréchet variance as a population audit target and show how to estimate it from federated quantile summaries.

The protocol is also connected to quantile summaries and streaming sketches. Classical quantile-sketching methods provide compact summaries of empirical distributions Greenwald & Khanna (2001); Karnin et al. (2016); Dunning & Ertl (2019). Here, quantiles are not used to approximate one global distribution directly. Each client communicates group-specific local quantiles; the server reconstructs group-mixture distributions using the corresponding group-specific client weights. This distinction is essential: averaging local quantiles would generally estimate a Wasserstein barycenter over clients, not the actual population mixture law.

1.4 Contributions

This paper makes four contributions.

A population-level distributional fairness target. We formalize score-level demographic-parity auditing through the sensitive-group score laws and define U_p, a Wasserstein–Fréchet variance measuring population-level distributional disparity. The functional targets the global criterion of interest, rather than an average of client-level fairness metrics. The key methodological point is that the protocol estimates the population audit target itself, rather than aggregating local audit values. The same summaries also explain why local and population audits disagree.

A federated formulation and a local–global decomposition. We express the same population target through client-level conditional laws and group-specific selection weights. For $p = 2$, we derive an ANOVA-style identity

$$G_2 = V_{\mathrm{mix}} + V_{\mathrm{bar}} + R,$$

where V_{mix} captures mixture/composition effects, V_{bar} captures barycentric cross-silo heterogeneity, and R is a controlled interaction term. The decomposition explains when local audits can be misleading and provides diagnostics for the source of the local–global mismatch.

A one-shot quantile-sketch audit protocol. We propose a communication-efficient protocol in which each silo sends only k empirical quantiles and one count per sensitive group. The server reconstructs group-mixture distributions, estimates the global Wasserstein disparity, computes a companion Cramér metric, and reports the $p = 2$ diagnostic terms. We separate deterministic finite-grid approximation from empirical quantile error and establish an $O(1/k)$ discretization term together with finite-sample guarantees based on quantile concentration.

Empirical validation, including natural silos. We validate the method on synthetic data, `COMPAS`, and `ACSIncome`. The synthetic and `COMPAS` experiments stress-test controlled local–global mismatch under simulated silo assignments. The `ACSIncome` experiment uses U.S. states and Puerto Rico as natural clients, showing that the method also applies to a large-scale non-simulated federated structure. Across these settings, a few dozen quantiles per group and silo are sufficient to closely recover the centralized benchmark.

Table 1: **Local and population audit targets.** Local audits condition on silo membership, whereas the proposed protocol estimates the population-level group-mixture target.

Quantity	Object compared	Interpretation
Local audit	$\{\nu_{j,s} : s \in \mathcal{S}\}$ for fixed j	Disparity within one silo
Average local audit	average of silo-level metrics	Summary of local disparities; not a population audit
Population audit	$\{\nu_s : s \in \mathcal{S}\}$	Disparity in the pooled target population
Federated estimate	$\{\widehat{\nu}^{\times}_{s,k} : s \in \mathcal{S}\}$	Sketch-based estimate of the population audit

Table 1 summarizes this distinction by separating the local quantities available within silos, naive summaries of local audits, the population-level target, and the federated estimate proposed in this paper.

Organization. Section 2 defines the population distributional fairness target. Section 3 introduces the quantile-sketch approximation, the federated reconstruction, and the statistical guarantees. Section 4 presents the W_2^2 local–global decomposition. Section 5 reports the empirical results. Section 6 discusses scope, privacy limitations, and extensions.

2 Distributional fairness target

This section defines the population-level fairness target estimated by the federated protocol. We first introduce the data-generating structure and the relevant score laws. We then formalize score-level demographic parity as equality of conditional score distributions and define two distributional disparity functionals: a Wasserstein–Fréchet target U_p, used throughout the audit, and a Cramér companion metric H_p, computed from the same reconstructed distributions.

2.1 Data model and score laws

We observe tuples (X, S, Y, A), where $X \in \mathbb{R}^q$ denotes non-sensitive covariates, $S \in \mathcal{S}$ is a sensitive attribute, Y is the prediction target, and $A \in [d] := \{1, \dots, d\}$ denotes the data-holding entity, or silo. In most examples, $\mathcal{S} = \{0, 1\}$, but the definitions apply to any finite set of sensitive groups. Silo j holds the local sample

$$\mathcal{D}_j = \{(x_i, s_i, y_i) : a_i = j\},$$

with size n_j. We write $n_{j,s}$ for the number of observations in silo j and group s,

$$n_s = \sum_{j=1}^{d} n_{j,s}, \qquad n = \sum_{s \in \mathcal{S}} n_s.$$

A deployed system produces a one-dimensional score

$$Z = m(X, S, A).$$

The notation allows either a common scoring rule or client-specific implementations; when useful, we write $m(x, s, a = j) = m_j(x, s)$. The inclusion of S and A in the notation is not a modeling requirement. It only records that, in some applications, the deployed score may depend on the sensitive attribute, on the client, or on variables correlated with either of them. The audit target is defined from the conditional score laws induced by the deployed system, regardless of how the score was constructed.

For each silo j and sensitive group s, let

$$\nu_{j,s} := \mathcal{L}(Z \mid A = j, S = s)$$

be the local conditional score law. The corresponding population-level group law is

$$\nu_s := \mathcal{L}(Z \mid S = s).$$

We write

$$\alpha_s := \mathbb{P}(S = s), \qquad \beta_j := \mathbb{P}(A = j), \qquad \pi_{j|s} := \mathbb{P}(A = j \mid S = s),$$

with empirical counterparts

$$\widehat{\alpha}_s = \frac{n_s}{n}, \qquad \widehat{\beta}_j = \frac{n_j}{n}, \qquad \widehat{\pi}_{j|s} = \frac{n_{j,s}}{n_s}.$$

Population quantities omit groups with $\alpha_s = 0$, and empirical quantities are interpreted only for positive-count groups and cells.

By the law of total probability, the population score law for group s is the silo mixture

$$\nu_s = \sum_{j=1}^{d} \pi_{j|s}\, \nu_{j,s}. \tag{1}$$

This identity is the source of the local–global mismatch studied in the paper. The population law ν_s depends both on the within-silo score laws $\nu_{j,s}$ and on the group-specific allocation weights $\pi_{j|s}$. Thus a local audit and a population audit generally refer to different distributional objects, even before any sampling or sketching approximation is introduced.

For any probability measure ν on $\mathbb{R}$, let F_ν denote its cumulative distribution function and Q_ν its left-continuous quantile function,

$$Q_\nu(u) := \inf\{z \in \mathbb{R} : F_\nu(z) \geq u\}, \qquad u \in (0, 1).$$

When no ambiguity is possible, we write $F_{j,s}, Q_{j,s}$ for $\nu_{j,s}$, and F_s, Q_s for ν_s.

2.2 Score-level demographic parity

At the score level, demographic parity means that the score distribution is independent of the sensitive attribute:

$$Z \perp\!\!\!\perp S.$$

Equivalently,

$$\nu_s = \nu_{s'} \qquad \text{for all } s, s' \in \mathcal{S}. \tag{2}$$

We therefore audit equality of the full conditional score laws $\{\nu_s : s \in \mathcal{S}\}$, rather than only thresholded decisions. This is useful when downstream thresholds are client-specific, unknown to the central auditor, or modified over time, and also when the score itself is experienced by individuals.

We use two one-dimensional discrepancies. For $p \geq 1$, the p-Wasserstein distance has the quantile representation

$$W_p(\nu, \nu')^p = \int_0^1 |Q_\nu(u) - Q_{\nu'}(u)|^p \, du, \tag{3}$$

for probability laws with finite p-th moments. We also use the Cramér p-distance, defined whenever the integral is finite by

$$C_p(\nu, \nu')^p := \int_{\mathbb{R}} |F_\nu(x) - F_{\nu'}(x)|^p \, dx. \tag{4}$$

In one dimension, $C_1 = W_1$, while C_2 is a Cramér–von Mises-type discrepancy. The Wasserstein functional below is naturally expressed through quantile functions, whereas the Cramér companion metric is naturally expressed through cdfs. Both are computable from the reconstructed step distributions used by the federated protocol.

2.3 Wasserstein–Fréchet disparity

Let $\{\nu_s : s \in \mathcal{S}\}$ be the population group score laws, with weights $\{\alpha_s : s \in \mathcal{S}\}$. We define the population-level distributional demographic-parity gap as the Fréchet variance of these laws in Wasserstein geometry.

Definition 2.1 (Population distributional disparity)**.** *For $p \geq 1$, define*

$$U_p := \inf_{\nu \in \mathcal{P}_p(\mathbb{R})} \sum_{s \in \mathcal{S}} \alpha_s W_p(\nu_s, \nu)^p. \tag{5}$$

Any minimizer ν^* of (5) is called a Wasserstein barycenter of the group score laws.

The functional U_p is a property of the population score laws themselves. It is not an average of local fairness metrics. The barycenter is used only as the distributional reference minimizing the weighted within-group Wasserstein dispersion.

For $p = 2$ in one dimension, the barycenter is unique and has quantile function

$$Q_{\nu^*}(u) = \sum_{s \in \mathcal{S}} \alpha_s Q_s(u), \qquad u \in (0, 1). \tag{6}$$

Consequently,

$$U_2 = \sum_{s \in \mathcal{S}} \alpha_s \int_0^1 \left(Q_s(u) - \sum_{r \in \mathcal{S}} \alpha_r Q_r(u) \right)^2 du. \tag{7}$$

This squared case is the basis of the diagnostic decomposition in Section 4. For $p = 1$, barycenters correspond pointwise to weighted medians in quantile space. The objective remains well defined, but the variance identity (7) is no longer available.

The functional U_p vanishes exactly under score-level demographic parity. If $\alpha_s > 0$ for all $s \in \mathcal{S}$, then

$$U_p = 0 \quad \Longleftrightarrow \quad \nu_s = \nu_{s'} \quad \text{for all } s, s' \in \mathcal{S}.$$

The proof is immediate from the metric property of W_p and is given in Appendix C.1. Thus U_p measures deviations from equality of score distributions across sensitive groups.

When $\mathcal{S} = \{0, 1\}$, the expression simplifies further. If $\alpha_s = \mathbb{P}(S = s)$, then

$$U_2 = \alpha_0 \alpha_1 W_2(\nu_0, \nu_1)^2. \tag{8}$$

This two-group form is used in the experiments. It shows that, for a binary sensitive attribute, U_2 is a prevalence-weighted Wasserstein discrepancy between the two conditional score laws.

2.4 A companion Cramér heterogeneity metric

Although U_p is the main audit target, it is useful to report a companion cdf-based metric. Let

$$\nu^{\text{pool}} := \sum_{s \in \mathcal{S}} \alpha_s \nu_s, \qquad F_{\text{pool}}(x) = \sum_{s \in \mathcal{S}} \alpha_s F_s(x), \tag{9}$$

be the population mixture over sensitive groups. Define

$$H_p := \sum_{s \in \mathcal{S}} \alpha_s C_p(\nu_s, \nu^{\text{pool}})^p. \tag{10}$$

The quantity H_p measures the same equality-of-distributions property as U_p, but in Cramér geometry. If all $\alpha_s > 0$, then

$$H_p = 0 \quad \Longleftrightarrow \quad \nu_s = \nu_{s'} \quad \text{for all } s, s' \in \mathcal{S}.$$

For two sensitive groups and $p = 2$,

$$H_2 = \alpha_0 \alpha_1 C_2(\nu_0, \nu_1)^2. \tag{11}$$

The distinction between U_p and H_p is operational. The Wasserstein gap U_p is tied to quantile functions and drives the quantile-sketch estimator. The Cramér gap H_p is tied to cdfs and can be computed exactly from the stepwise reconstructed distributions once the communicated quantiles have been sorted. Reporting both gives complementary views of the same score-level demographic-parity question.

3 Federated audit from quantile sketches

We now describe how to estimate the population-level target defined in Section 2 when raw scores cannot be centralized. The key observation is one-dimensional: Wasserstein distances are determined by quantile functions. A small number of score quantiles per sensitive group and silo therefore provides a natural communication interface for auditing score-level demographic parity.

We distinguish three objects. First, $G_p = U_p$ denotes the population-level target written in federated form. Second, $G_{k,p}$ denotes the finite-grid population target obtained by applying the same reconstruction algorithm to population quantiles. Third, $\widehat{G}_{k,p}$ denotes the empirical federated estimator computed from subgroup counts and empirical local quantiles. This separation makes the error decomposition transparent: one term comes from the finite quantile grid, and the other from estimating local quantiles and mixing weights.

3.1 Quantile-grid approximation

Fix an integer $k \geq 1$, the sketch resolution, and define the midpoint grid

$$u_\ell = \frac{\ell - \frac{1}{2}}{k}, \qquad \ell \in [k]. \tag{12}$$

For a distribution ν with quantile function Q_ν, the associated k-quantile sketch is

$$\big(Q_\nu(u_1), \ldots, Q_\nu(u_k)\big).$$

This sketch induces the atomic reconstruction

$$\nu_k := \frac{1}{k} \sum_{\ell=1}^{k} \delta_{Q_\nu(u_\ell)}, \tag{13}$$

with cdf

$$F_k(x) := \frac{1}{k} \sum_{\ell=1}^{k} \mathbf{1}\{Q_\nu(u_\ell) \leq x\}. \tag{14}$$

The midpoint grid avoids querying the boundary quantiles 0 and 1, which are typically unstable and sensitive to extremes.

Applied to the population group laws $\{\nu_s : s \in \mathcal{S}\}$, this gives

$$\nu_{s,k} := \frac{1}{k} \sum_{\ell=1}^{k} \delta_{Q_s(u_\ell)}.$$

The corresponding finite-grid Wasserstein–Fréchet disparity is

$$U_{k,p} := \inf_{\nu \in \mathcal{P}_p(\mathbb{R})} \sum_{s \in \mathcal{S}} \alpha_s W_p(\nu_{s,k}, \nu)^p. \tag{15}$$

Equivalently, in one dimension,

$$U_{k,p} = \sum_{s \in \mathcal{S}} \alpha_s \frac{1}{k} \sum_{\ell=1}^{k} |q_{s,\ell} - q_\ell^*|^p \,, \tag{16}$$

where $q_{s,\ell} := Q_s(u_\ell)$ and

$$q_\ell^* \in \arg\min_{z \in \mathbb{R}} \sum_{s \in \mathcal{S}} \alpha_s |q_{s,\ell} - z|^p.$$

For $p = 2$,

$$q_\ell^* = \sum_{s \in \mathcal{S}} \alpha_s q_{s,\ell}.$$

The Cramér companion metric is computed from the same reconstructed distributions. Let

$$\nu_k^{\text{pool}} := \sum_{s\in\mathcal{S}} \alpha_s \nu_{s,k}, \qquad F_k^{\text{pool}}(x) := \sum_{s\in\mathcal{S}} \alpha_s F_{s,k}(x).$$

Define

$$H_{k,p} := \sum_{s\in\mathcal{S}} \alpha_s C_p(\nu_{s,k}, \nu_k^{\text{pool}})^p. \tag{17}$$

Since the reconstructed cdfs are step functions, the Cramér integral is computed exactly by sorting their jump locations. If $g_1 < \cdots < g_M$ are the distinct support points of the reconstructed distributions, then

$$C_p(\nu_{s,k}, \nu_k^{\text{pool}})^p = \sum_{m=1}^{M-1} (g_{m+1} - g_m) \left| F_{s,k}(g_m) - F_k^{\text{pool}}(g_m) \right|^p. \tag{18}$$

Proposition 3.1 (Centralized sketch approximation)**.** *Assume that, for each $s \in \mathcal{S}$, ν_s has compact support and admits a continuous density bounded away from 0 and ∞ on its support. Then, for $p \in \{1, 2\}$, there exist constants $C_U, C_H < \infty$ such that*

$$|U_p - U_{k,p}| \le \frac{C_U}{k}, \qquad |H_p - H_{k,p}| \le \frac{C_H}{k}.$$

Moreover, for fixed k, the empirical plug-in estimators obtained by replacing population quantiles and weights by their empirical counterparts are consistent as the relevant sample sizes diverge.

Remark (Discrete or tied scores)**.** *The estimator remains well defined for discrete scores, tied scores, clipped probabilities, and other non-smooth score distributions. What depends on the regularity assumptions is the sharp $O(1/k)$ deterministic approximation rate. In discrete settings, the midpoint-grid estimator can still be computed, but sensitivity to k should be reported as part of the audit. When a continuous approximation is desired, one may add a small independent jitter, as in the* `COMPAS` *experiments; this is a practical device and not a substitute for the regularity assumptions used in the rate statement.*

3.2 Federated population target

In the federated setting, the server does not observe the pooled scores. Each silo j observes only samples from the local laws

$$\nu_{j,s} = \mathcal{L}(Z \mid A = j, S = s).$$

Recall from (1) that the population score law for sensitive group s is

$$\nu_s = \sum_{j=1}^{d} \pi_{j|s} \nu_{j,s}.$$

To emphasize the federated construction, define

$$\nu_s^{\times} := \sum_{j=1}^{d} \pi_{j|s} \nu_{j,s}, \qquad F_s^{\times}(x) = \sum_{j=1}^{d} \pi_{j|s} F_{j,s}(x). \tag{19}$$

The superscript $\times$ indicates a mixture over silos. Mixture and quantile operations do not commute:

$$Q_s^{\times}(u) \neq \sum_{j=1}^{d} \pi_{j|s} Q_{j,s}(u)$$

in general. Thus the server must reconstruct the within-group mixture cdf and then invert it. Averaging local quantiles would estimate a Wasserstein barycenter across silos, not the actual population mixture law.

Let $\nu^{\times *}$ be a Wasserstein barycenter of the group-mixture laws $\{\nu_s^\times : s \in \mathcal{S}\}$. Define

$$G_p := \inf_{\nu \in \mathcal{P}_p(\mathbb{R})} \sum_{s \in \mathcal{S}} \alpha_s W_p(\nu_s^\times, \nu)^p = \sum_{s \in \mathcal{S}} \alpha_s W_p(\nu_s^\times, \nu^{\times *})^p, \tag{20}$$

where the second equality holds for any barycenter $\nu^{\times *}$. Similarly, define the federated Cramér companion metric by

$$\nu^{\mathrm{pool},\times} := \sum_{s \in \mathcal{S}} \alpha_s \nu_s^\times,$$

and

$$H_p^\times := \sum_{s \in \mathcal{S}} \alpha_s C_p(\nu_s^\times, \nu^{\mathrm{pool},\times})^p. \tag{21}$$

Proposition 3.2 (Population target equivalence)**.** *With the notation above, $\nu_s^\times = \nu_s$ for every $s \in \mathcal{S}$. Hence*

$$G_p = U_p, \qquad H_p^\times = H_p.$$

This proposition is a population identity. The federated protocol below does not define a new fairness criterion. It estimates the same target U_p, but from finite-grid reconstructions and empirical summaries communicated by the clients.

3.3 One-shot protocol

Each silo j communicates, for each sensitive group $s \in \mathcal{S}$,

$$n_{j,s} \qquad \text{and} \qquad \widehat{q}_{j,s,\ell} := \widehat{Q}_{j,s}(u_\ell), \quad \ell \in [k],$$

where $\widehat{Q}_{j,s}$ is the empirical quantile function of the local scores

$$\{Z_i : A_i = j, S_i = s\}.$$

Thus each silo sends $|\mathcal{S}|(k+1)$ scalars. With two sensitive groups, this is $2(k+1)$ scalars per silo. For example, with $d = 51$ clients, two sensitive groups, and $k = 50$, the audit communicates $51 \times 2 \times 51 = 5202$ scalars, rather than the full vector of individual scores.

From the counts, the server computes

$$\widehat{\alpha}_s = \frac{n_s}{n}, \qquad \widehat{\pi}_{j|s} = \frac{n_{j,s}}{n_s}, \qquad n_s = \sum_{j=1}^{d} n_{j,s}.$$

For each (j, s), the transmitted quantiles define the local reconstructed cdf

$$\widehat{F}_{j,s,k}(x) := \frac{1}{k} \sum_{\ell=1}^{k} \mathbf{1}\{\widehat{q}_{j,s,\ell} \le x\}.$$

The server then forms the within-group mixture cdf across silos:

$$\widehat{F}_{s,k}^\times(x) := \sum_{j=1}^{d} \widehat{\pi}_{j|s} \widehat{F}_{j,s,k}(x). \tag{22}$$

Let $\widehat{\nu}_{s,k}^\times$ denote the probability law with cdf $\widehat{F}_{s,k}^\times$. By inverting this step cdf on the midpoint grid, the server obtains reconstructed mixture quantiles

$$\widehat{q}_{s,\ell}^\times := \inf\{x : \widehat{F}_{s,k}^\times(x) \ge u_\ell\}, \qquad \ell \in [k]. \tag{23}$$

For $p = 2$, the barycenter quantile across sensitive groups is

$$\widehat{q}_\ell^{\times *} = \sum_{s \in \mathcal{S}} \widehat{\alpha}_s \widehat{q}_{s,\ell}^{\times}.$$

The federated estimate of G_2 is therefore

$$\widehat{G}_{k,2} = \sum_{s \in \mathcal{S}} \widehat{\alpha}_s \frac{1}{k} \sum_{\ell=1}^{k} \left(\widehat{q}_{s,\ell}^{\times} - \widehat{q}_\ell^{\times *} \right)^2. \tag{24}$$

For $p \in \{1, 2\}$, the same construction is written as

$$\widehat{G}_{k,p} = \sum_{s \in \mathcal{S}} \widehat{\alpha}_s \frac{1}{k} \sum_{\ell=1}^{k} \left| \widehat{q}_{s,\ell}^{\times} - \widehat{q}_\ell^{\times *} \right|^p, \tag{25}$$

where

$$\widehat{q}_\ell^{\times *} \in \arg\min_{z \in \mathbb{R}} \sum_{s \in \mathcal{S}} \widehat{\alpha}_s |\widehat{q}_{s,\ell}^{\times} - z|^p.$$

The Cramér estimate is computed from the reconstructed step distributions. Define

$$\widehat{\nu}_k^{\mathrm{pool},\times} := \sum_{s \in \mathcal{S}} \widehat{\alpha}_s \widehat{\nu}_{s,k}^{\times}, \qquad \widehat{F}_k^{\mathrm{pool},\times}(x) := \sum_{s \in \mathcal{S}} \widehat{\alpha}_s \widehat{F}_{s,k}^{\times}(x).$$

Then

$$\widehat{H}_{k,p}^{\times} := \sum_{s \in \mathcal{S}} \widehat{\alpha}_s C_p(\widehat{\nu}_{s,k}^{\times}, \widehat{\nu}_k^{\mathrm{pool},\times})^p. \tag{26}$$

The Cramér integrals are evaluated exactly by sorting the jump locations, using the step-function formula (18). Full pseudocode is given in Appendix F.

3.4 Statistical guarantee

The estimation error has two components. The first is deterministic: the population distributions are replaced by finite-grid reconstructions using k quantile levels. The second is statistical: local quantiles and mixing weights are estimated from finite samples.

Let $G_{k,p}$ denote the finite-grid population counterpart of $\widehat{G}_{k,p}$, obtained by using the population weights $\alpha_s, \pi_{j|s}$ and the population local quantiles $Q_{j,s}(u_\ell)$ in the same reconstruction algorithm. Then

$$|\widehat{G}_{k,p} - G_p| \le |\widehat{G}_{k,p} - G_{k,p}| + |G_{k,p} - G_p|. \tag{27}$$

The second term is a deterministic approximation error. The first term is the empirical error propagated through the step-cdf reconstruction and the barycenter computation.

Proposition 3.3 (Federated consistency and rate)**.** *Assume the conditions of Proposition 3.1 for each group-mixture law $\nu_s^{\times}$. Assume also that the local laws $\nu_{j,s}$ satisfy the trimmed quantile regularity conditions used in Proposition 3.4, and that all positive-probability cells have $n_{j,s} \to \infty$. Let*

$$n_{\min} := \min_{\substack{j \in [d],\, s \in \mathcal{S} \\ n_{j,s} > 0}} n_{j,s}.$$

For fixed $p \in \{1, 2\}$, if $k = k(n) \to \infty$ and $n_{\min} \to \infty$, then

$$\widehat{G}_{k,p} \xrightarrow{\mathbb{P}} G_p, \qquad \widehat{H}_{k,p}^{\times} \xrightarrow{\mathbb{P}} H_p^{\times}.$$

Moreover, for fixed d and fixed $|\mathcal{S}|$,

$$|\widehat{G}_{k,p} - G_p| = O_{\mathbb{P}}\left(\frac{1}{k} + \sqrt{\frac{\log k}{n_{\min}}} \right),$$

and the same rate holds for $\widehat{H}_{k,p}^{\times}$.

The rate in Proposition 3.3 should be read as a communication–accuracy trade-off. Increasing k reduces finite-grid bias, but requires more communication and may increase the finite-sample variability of high-resolution local quantiles. For auditing, k should therefore be treated as part of the audit design, and sensitivity of $\widehat{G}_{k,2}$ to the grid resolution should be reported. Appendix A describes a piecewise-linear reconstruction that can improve the deterministic approximation on smooth interior regions without increasing the number of transmitted quantiles. With the constant boundary extension used there, however, boundary contributions may remain first order unless additional boundary or tail corrections are imposed.

Proposition 3.4 (High-probability control of communicated quantiles)**.** *Fix $\varepsilon \in (0, 1/2)$. Assume that, for every positive-probability cell (j, s), the cdf $F_{j,s}$ is absolutely continuous on*

$$[Q_{j,s}(\varepsilon/2), Q_{j,s}(1-\varepsilon/2)]$$

with density bounded below by $m_\varepsilon > 0$ on this interval. Define

$$\Delta_{n,k,\delta} := \sqrt{\frac{1}{2n_{\min}} \log\left(\frac{2(k+1)d|\mathcal{S}|}{\delta}\right)}.$$

If $\Delta_{n,k,\delta} \le \varepsilon/2$, then, with probability at least $1-\delta$,

$$\max_{\substack{j\in[d],\, s\in\mathcal{S} \\ \ell\in\{\lceil\varepsilon k\rceil,\dots,\lfloor(1-\varepsilon)k\rfloor\}}} \left|\widehat{Q}_{j,s}(u_\ell) - Q_{j,s}(u_\ell)\right| \le \frac{\Delta_{n,k,\delta}}{m_\varepsilon}.$$

Equivalently,

$$\max_{\substack{j\in[d],\, s\in\mathcal{S} \\ \ell\in\{\lceil\varepsilon k\rceil,\dots,\lfloor(1-\varepsilon)k\rfloor\}}} \left|\widehat{Q}_{j,s}(u_\ell) - Q_{j,s}(u_\ell)\right| \le \frac{1}{m_\varepsilon}\sqrt{\frac{1}{2n_{\min}} \log\left(\frac{2(k+1)d|\mathcal{S}|}{\delta}\right)}.$$

This bound follows from the Dvoretzky–Kiefer–Wolfowitz–Massart inequality and Lipschitz inversion of cdfs on a trimmed quantile region Dvoretzky et al. (1956); Massart (1990). Combining it with the stability of the reconstructed step distributions yields high-probability bounds for $\widehat{G}_{k,2}$, $\widehat{H}^{\times}_{k,2}$, and the diagnostic terms introduced in Section 4. The details are given in Appendix D.5.

Role of the assumptions. The assumptions above play different roles. Compact support gives finite constants in the Wasserstein and Cramér approximation bounds. A density bounded away from zero is used to convert cdf error into quantile error; without it, small cdf perturbations can produce large quantile perturbations in flat regions. A density bounded above is useful for controlling how many reconstructed knots can lie in a short interval, which enters the stability of the step-cdf reconstruction. Discrete or tied scores do not invalidate the estimator, but they do invalidate the clean Lipschitz-quantile argument behind the sharp $O(1/k)$ rate. Finally, the dependence on $n_{\min}$ reflects the fact that the most weakly represented silo–group cell controls the worst-case local quantile error.

3.5 Scope of the protocol

The protocol should be read as a minimal-disclosure audit layer, not as a standalone privacy mechanism. It avoids centralizing individual records and individual scores, but subgroup counts and quantiles can still leak information. If $n_{j,s} = 1$, all released quantiles coincide with the individual score; if $n_{j,s}$ is small, extreme or repeated quantiles may reveal partial information about local records. Safe deployment therefore requires additional protections, such as minimum cell-size rules, cell suppression or coarsening, secure aggregation, and, when appropriate, calibrated noise. A minimal deployment rule is to suppress or merge every silo–group cell with $n_{j,s} < m_{\min}$ before releasing quantiles. The threshold $m_{\min}$ should be chosen before the audit and reported with the results. If suppression changes the target population, the audit should report both the retained population share and the affected groups. The stability bounds above can be used to study how perturbations of counts or quantiles affect the audit estimates, but a full differential-privacy analysis is outside the scope of this paper.

The protocol targets score-level demographic parity. This is useful when downstream decision rules are not shared across silos, differ across clients, or include ranking and human review. It is also directly meaningful in settings where the score itself is experienced by individuals, such as insurance premiums, credit scores, clinical risk indices, or priority scores. Decision-level audits, such as demographic parity after thresholding, equalized odds, predictive parity, or calibration, remain complementary. A label-conditional extension is possible in principle by replacing each $\nu_{j,s}$ with conditional laws

$$\nu_{j,s,y} := \mathcal{L}(Z \mid A = j, S = s, Y = y).$$

This would allow label-conditional audits, but at the cost of additional communication and increased small-cell disclosure risk.

4 Diagnosing local–global mismatch

The protocol in Section 3 estimates the population-level disparity $G_p = U_p$. For $p = 2$, the same quantile summaries also provide a diagnostic decomposition of the local–global mismatch. Throughout this section, scores are one-dimensional and $p = 2$. In this case,

$$W_2(\mu, \nu)^2 = \int_0^1 \left(Q_\mu(u) - Q_\nu(u)\right)^2 \, du,$$

so probability laws can be compared through their quantile functions in the Hilbert space $L^2([0, 1])$. This Hilbertian representation is what makes an ANOVA-style decomposition possible.

4.1 Mixtures and transport barycenters within sensitive groups

Recall that the population score law for sensitive group s is the silo mixture

$$\nu_s^\times = \sum_{j=1}^d \pi_{j|s} \nu_{j,s}.$$

This is the law that enters the population audit: it is the actual pooled score distribution of group s. A different object is obtained by averaging the same local laws in Wasserstein geometry. Define the within-group transport barycenter

$$\nu_s^* \in \arg\min_{\nu \in \mathcal{P}_2(\mathbb{R})} \sum_{j=1}^d \pi_{j|s} W_2(\nu_{j,s}, \nu)^2.$$

In one dimension, this barycenter is characterized by the quantile average

$$Q_s^*(u) = \sum_{j=1}^d \pi_{j|s} Q_{j,s}(u), \qquad u \in (0, 1). \tag{28}$$

The distinction between $\nu_s^\times$ and ν_s^* is central. The law $\nu_s^\times$ is a mixture of probability distributions and corresponds to the pooled population within group s. The law ν_s^* is a transport barycenter and corresponds to averaging local quantile functions. In general,

$$Q_{\sum_j \pi_{j|s} \nu_{j,s}}(u) \neq \sum_j \pi_{j|s} Q_{j,s}(u).$$

Thus, the global group law cannot be recovered by simply averaging local quantiles. The non-commutation between mixture and quantile averaging is one mechanism through which silo composition can create a local–global discrepancy.

Let $\nu^{\times *}$ denote the W_2 barycenter of the group-mixture laws $\{\nu_s^\times : s \in \mathcal{S}\}$. Its quantile function is

$$Q^{\times *}(u) = \sum_{s \in \mathcal{S}} \alpha_s Q_s^\times(u).$$

The population disparity can therefore be written as

$$G_2 = \sum_{s\in\mathcal{S}} \alpha_s \int_0^1 \left(Q_s^{\times}(u) - Q^{\times *}(u)\right)^2 du.$$

4.2 ANOVA-style decomposition

Define the mixture non-commutation term

$$V_{\text{mix}} := \sum_{s\in\mathcal{S}} \alpha_s W_2(\nu_s^{\times}, \nu_s^*)^2 = \sum_{s\in\mathcal{S}} \alpha_s \int_0^1 \left(Q_s^{\times}(u) - Q_s^*(u)\right)^2 du, \tag{29}$$

and the barycentric cross-silo term, anchored at the population barycenter $\nu^{\times *}$,

$$V_{\text{bar}} := \sum_{s\in\mathcal{S}} \alpha_s W_2(\nu_s^*, \nu^{\times *})^2 = \sum_{s\in\mathcal{S}} \alpha_s \int_0^1 \left(Q_s^*(u) - Q^{\times *}(u)\right)^2 du. \tag{30}$$

The first term measures how far the actual within-group mixture is from the within-group transport barycenter. The second measures how far the transport-averaged group profiles are from the global barycenter of the actual group mixtures.

Theorem 4.1 (ANOVA-style decomposition for W_2^2)**.** *In one dimension, for $p = 2$,*

$$G_2 = V_{\text{mix}} + V_{\text{bar}} + R, \tag{31}$$

where

$$R := 2\sum_{s\in\mathcal{S}} \alpha_s \left\langle Q_s^{\times} - Q_s^*, Q_s^* - Q^{\times *}\right\rangle_{L^2([0,1])}. \tag{32}$$

Moreover,

$$|R| \le 2\sqrt{V_{\text{mix}} V_{\text{bar}}}. \tag{33}$$

Proof. For each sensitive group s, write

$$Q_s^{\times} - Q^{\times *} = (Q_s^{\times} - Q_s^*) + (Q_s^* - Q^{\times *}).$$

Expanding the squared L^2-norm gives

$$\begin{aligned}\|Q_s^{\times} - Q^{\times *}\|_2^2 = \|Q_s^{\times} - Q_s^*\|_2^2 + \|Q_s^* - Q^{\times *}\|_2^2 \\ + 2\left\langle Q_s^{\times} - Q_s^*, Q_s^* - Q^{\times *}\right\rangle.\end{aligned}$$

Multiplying by α_s and summing over s gives (31). The bound (33) follows from Cauchy–Schwarz:

$$\begin{aligned}|R| &\le 2\sum_s \alpha_s \|Q_s^{\times} - Q_s^*\|_2 \|Q_s^* - Q^{\times *}\|_2 \\ &\le 2\left(\sum_s \alpha_s \|Q_s^{\times} - Q_s^*\|_2^2\right)^{1/2} \left(\sum_s \alpha_s \|Q_s^* - Q^{\times *}\|_2^2\right)^{1/2}.\end{aligned}$$

□

The theorem immediately yields the two-sided bounds

$$\left(\sqrt{V_{\text{mix}}} - \sqrt{V_{\text{bar}}}\right)^2 \le G_2 \le \left(\sqrt{V_{\text{mix}}} + \sqrt{V_{\text{bar}}}\right)^2. \tag{34}$$

Thus V_{mix} and V_{bar} control the possible range of the global disparity. The interaction term R, however, should not be interpreted as a negligible residual. It can be positive or negative and may be of the same order as the two nonnegative components.

4.3 Interpretation

The decomposition separates two sources of local–global mismatch and their interaction.

Mixture/composition effect. The term V_{mix} is large when, within a sensitive group, the actual mixture over silos differs substantially from the transport barycenter across silos. It captures the discrepancy between the population mixture law and the law obtained by averaging local quantiles. This is the component most directly associated with Simpson-style phenomena: the population law is formed by mixture, whereas Wasserstein geometry is governed by quantiles.

Barycentric cross-silo heterogeneity. The term V_{bar} measures how far the within-group transport barycenters are from the global barycenter of the actual group mixtures. It captures a structural component of disparity that remains after each sensitive group has been summarized by its transport-average behavior across silos. It is large when the group-specific barycentric profiles remain separated in Wasserstein geometry.

Interaction. The term R measures the alignment between the mixture discrepancy $Q_s^\times - Q_s^*$ and the barycentric gap $Q_s^* - Q^{\times *}$. If $R > 0$, the two effects reinforce each other; if $R < 0$, they partially offset each other. The Cauchy–Schwarz bound (33) guarantees that the interaction is controlled by V_{mix} and V_{bar}, but it does not imply that R is small.

This interpretation gives a diagnostic reading of the global audit. If V_{mix} dominates, then the population disparity is mainly driven by group-specific silo composition and by the non-commutation between mixture and quantile averaging. If V_{bar} dominates, then disparity persists even after transport-averaging local score laws within each sensitive group. The sign and magnitude of R indicate whether these two mechanisms reinforce or offset each other.

4.4 When do the terms vanish?

The decomposition is especially transparent in several limiting regimes.

Proposition 4.2 (Sufficient conditions for vanishing terms)**.** *Assume one-dimensional scores and $p = 2$.*

1. *If, for each sensitive group s, the local law $\nu_{j,s}$ does not depend on j whenever $\pi_{j|s} > 0$, then*
$$\nu_s^\times = \nu_s^* \qquad \text{for all } s.$$
Consequently, $V_{\mathrm{mix}} = 0$, $R = 0$, and $G_2 = V_{\mathrm{bar}}$. In particular, this holds under $Z \perp\!\!\!\perp A \mid S$, up to the usual positive-cell qualification.

2. *If, for each silo j, the local law $\nu_{j,s}$ does not depend on s whenever $\mathbb{P}(A = j, S = s) > 0$, then local score-level demographic parity holds within each silo. In this case, a population-level disparity may still arise from differences in the group-specific allocation weights $\pi_{j|s}$. If, in addition,*
$$\pi_{j|s} = \beta_j \qquad \text{for all } j, s$$
on the relevant support, then $\nu_s^\times$ does not depend on s, and therefore $G_2 = 0$.

3. *If there exists a common law λ such that*
$$\nu_{j,s} = \lambda \qquad \text{for all positive-probability cells } (j, s),$$
then
$$G_2 = V_{\mathrm{mix}} = V_{\mathrm{bar}} = R = 0.$$
Equivalently, if both $Z \perp\!\!\!\perp A \mid S$ and $Z \perp\!\!\!\perp S \mid A$ hold, together with a positivity condition linking the relevant silos and sensitive groups, then the score law is common across the connected support and the population disparity vanishes.

The second regime is the most relevant for local–global mismatch. It says that local score-level demographic parity within every silo does not, by itself, imply population-level demographic parity. If sensitive groups are allocated differently across heterogeneous silos, the pooled group score laws can differ even though each within-silo comparison is balanced.

4.5 Estimating the decomposition from the same sketches

No additional communication is needed to estimate the terms in Theorem 4.1. The server already has the local quantiles $\widehat{q}_{j,s,\ell}$, the weights $\widehat{\pi}_{j|s}$, the reconstructed mixture quantiles $\widehat{q}^{\times}_{s,\ell}$, and the global barycenter quantiles $\widehat{q}^{\times *}_{\ell}$.

For $p = 2$, define the within-group barycenter quantiles

$$\widehat{q}^{*}_{s,\ell} := \sum_{j=1}^{d} \widehat{\pi}_{j|s} \widehat{q}_{j,s,\ell}. \tag{35}$$

Then the plug-in diagnostic terms are

$$\widehat{V}_{\text{mix}} := \sum_{s\in\mathcal{S}} \widehat{\alpha}_s \frac{1}{k} \sum_{\ell=1}^{k} \left(\widehat{q}^{\times}_{s,\ell} - \widehat{q}^{*}_{s,\ell} \right)^2, \tag{36}$$

$$\widehat{V}_{\text{bar}} := \sum_{s\in\mathcal{S}} \widehat{\alpha}_s \frac{1}{k} \sum_{\ell=1}^{k} \left(\widehat{q}^{*}_{s,\ell} - \widehat{q}^{\times *}_{\ell} \right)^2, \tag{37}$$

and

$$\widehat{R} := 2 \sum_{s\in\mathcal{S}} \widehat{\alpha}_s \frac{1}{k} \sum_{\ell=1}^{k} \left(\widehat{q}^{\times}_{s,\ell} - \widehat{q}^{*}_{s,\ell} \right) \left(\widehat{q}^{*}_{s,\ell} - \widehat{q}^{\times *}_{\ell} \right). \tag{38}$$

By construction,

$$\widehat{G}_{k,2} = \widehat{V}_{\text{mix}} + \widehat{V}_{\text{bar}} + \widehat{R}$$

up to numerical rounding. The same quantile-estimation and finite-grid arguments used for $\widehat{G}_{k,2}$ apply to these diagnostic terms.

Reporting recommendation. In empirical audits, we report both the global estimate $\widehat{G}_{k,2}$ and the triplet

$$(\widehat{V}_{\text{mix}}, \widehat{V}_{\text{bar}}, \widehat{R}).$$

The global estimate answers the audit question: how large is the population-level disparity? The decomposition answers the diagnostic question: is this disparity mainly due to group-specific silo composition, to barycentric cross-silo heterogeneity, or to the interaction between the two?

5 Experiments

The experiments evaluate three claims. First, silo-level audits can misrepresent the population-level demographic disparity when silo assignment is selective or when score distributions vary across silos. Second, the proposed one-shot quantile protocol recovers the centralized population target from a small number of communicated summaries. Third, for $p = 2$, the diagnostic decomposition helps identify whether the local–global mismatch is mainly driven by group-specific silo composition, barycentric cross-silo heterogeneity, or their interaction.

We consider three experimental regimes. The synthetic experiment isolates the mechanism under a controlled selection process. The `COMPAS` experiment uses a standard fairness benchmark with simulated silos, allowing us to stress-test local–global mismatch under known allocation mechanisms. The `ACSIncome` experiment uses natural silos: U.S. states and Puerto Rico are treated as clients. Additional robustness experiments on `Adult` and `Law` are reported in the appendix.

5.1 Protocol and reported quantities

Unless stated otherwise, we focus on $p = 2$. Each silo j sends, for each sensitive group $s \in \mathcal{S}$, the subgroup count $n_{j,s}$ and the empirical quantiles

$$\widehat{q}_{j,s,\ell} = \widehat{Q}_{j,s}(u_\ell), \qquad u_\ell = \frac{\ell - \frac{1}{2}}{k}, \qquad \ell \in [k].$$

Table 2: **Communication budget for representative settings.** The protocol communicates subgroup counts and k quantiles per sensitive group and silo.

Setting	Clients d	Groups $\lvert\mathcal{S}\rvert$	Quantiles k	Communicated scalars
`COMPAS`	5	2	50	510
`ACSIncome`	51	2	50	5202
`ACSIncome`	51	2	100	10302

The server reconstructs the group-mixture cdfs, inverts them on the same grid, and computes the global estimate $\widehat{G}_{k,2}$. It also computes the Cramér companion metric $\widehat{H}^{\times}_{k,2}$ and, when $p = 2$, the diagnostic identity

$$\widehat{G}_{k,2} = \widehat{V}_{\text{mix}} + \widehat{V}_{\text{bar}} + \widehat{R}.$$

We compare $\widehat{G}_{k,2}$ with a centralized benchmark computed from pooled individual-level scores. This benchmark is used only for evaluation; it is not available to the federated audit protocol. Local audits are computed by applying the same distributional metrics within each silo separately.

For two sensitive groups, each silo communicates $2(k + 1)$ scalars: k quantiles and one count for each group. More generally, the total number of communicated scalars is

$$d\,|\mathcal{S}|\,(k + 1).$$

Thus k is both a statistical resolution parameter and a communication parameter. In all experiments, we report sensitivity to k, because finite-grid approximation is part of the audit design rather than an implementation detail. In practice, an auditor can choose k by increasing the sketch resolution until the reported estimate is stable at the desired audit tolerance. For instance, one may select the smallest k such that $|\widehat{G}_{k,2} - \widehat{G}_{k/2,2}|/\max(\widehat{G}_{k,2}, \varepsilon)$ falls below a prespecified threshold, subject to the requirement that all released silo–group cells satisfy the minimum-size rule.

Table 2 gives the scale of the communication budget. The actual number of transmitted bytes depends on numerical encoding and on any additional privacy safeguards, such as secure aggregation or noise addition.

5.2 Synthetic data: controlled selection

We first use a controlled two-group experiment designed to isolate the effect of silo assignment[1]. Scores are generated from two distinct Beta distributions on $[0, 1]$, one for each sensitive group. The pooled group-wise score distributions are then kept fixed throughout the experiment. Only the allocation of individuals to silos is changed. This distinction is important. The population-level disparity U_2 depends only on the two pooled group score distributions. Since these margins are fixed, the global audit target is essentially unchanged as the silo-assignment mechanism varies. By contrast, the local quantities $U_{2,j}$ are computed after conditioning on silo membership. They can therefore change dramatically when silo membership becomes informative about the score.

Silo assignment is controlled by a latent Gaussian pairing parameter ρ. When $\rho \simeq 0$, silo membership is approximately independent of the score and of the sensitive group. In that regime, each silo receives an approximately representative sample of the population, up to Monte Carlo variation. The local score distributions therefore resemble the pooled score distributions, and the local audits are close to the population audit. The situation changes when ρ is positive. In the construction used here, larger positive values of ρ increasingly sort individuals into silos according to their score ranks. This can make the two sensitive groups look locally more similar inside each silo, because each silo contains a narrower and more aligned slice of the two group distributions. As a result, the local values $U_{2,j}$ may decrease, sometimes approaching zero, even though the population-level disparity has not changed. In other words, the silos can appear locally fair precisely because the allocation mechanism has hidden the global difference inside the selection process. Negative values of ρ illustrate the opposite failure mode. The same fixed population-level disparity can

[1]See https://anonymous.4open.science/r/federated-6FE5/ for replication.

be associated with much larger local disparities when the allocation mechanism separates the two groups differently across score ranges. Thus, by changing only the silo-assignment mechanism, one can make the local audits understate, overstate, or approximately match the population audit. This is the local–global mismatch targeted by the paper.

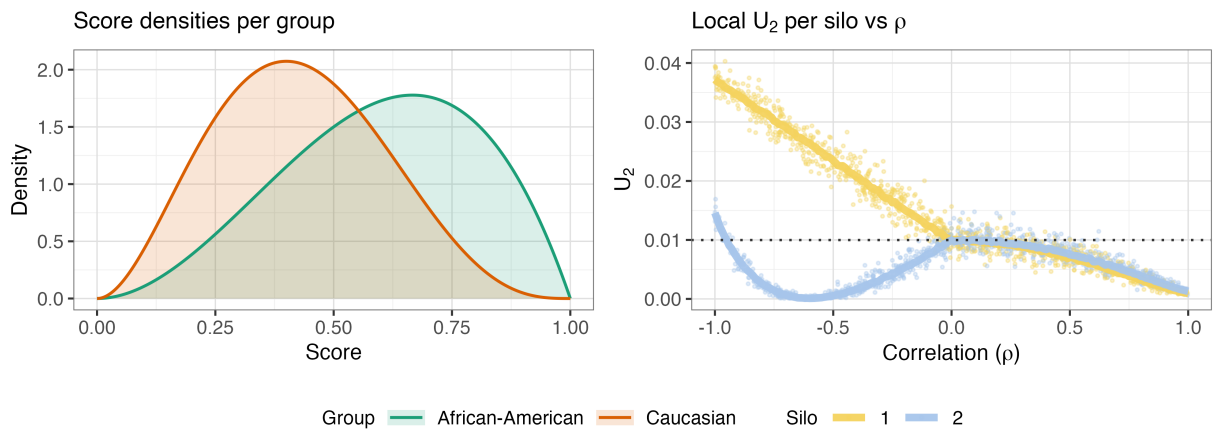


Figure 1: **Synthetic Beta experiment.** Left: marginal score distributions by sensitive group. Right: local U_2 values by silo as a function of the selection parameter ρ. The dotted horizontal line is the population-level U_2, which is essentially fixed because the group-wise score margins are unchanged. Local audits, however, vary strongly with the silo-assignment mechanism. For positive ρ, silos may appear increasingly fair even though the global disparity remains present.

Figure 1 shows this phenomenon directly. The left panel displays the two fixed marginal score distributions. The right panel changes only the way individuals are allocated to silos. Around $\rho = 0$, the local values are close to the global benchmark, as expected under approximately independent allocation. As ρ increases, however, the local values can fall below the population benchmark and may become very small, even though the global disparity remains close to its original level. Local fairness, in this experiment, is not a property of the scoring rule alone; it is also a property of how the population is partitioned. This experiment therefore gives a simple warning. A collection of silo-level audits cannot, by itself, identify the population-level audit. If silo membership is independent of score and sensitive-group composition, local and global audits may be close. But once silo membership is selective, local audits can become arbitrarily misleading: the same global disparity can coexist with apparently small local disparities, large local disparities, or a heterogeneous mix of both. The density plots in Figure 2 make this selection effect visible in the two-silo case: changing ρ does not change the pooled group margins, but it changes which parts of those margins are seen inside each silo.

Figure 3 separates this structural issue from the finite-grid approximation. Under independent allocation, the sketch estimator is already close to the centralized benchmark for moderate values of k. Under selection-biased allocation, small sketches are more variable and may be biased, because the local conditional distributions are more sharply selected. In both cases, the estimates stabilize as k increases. The discrepancy in Figure 1 is therefore not a numerical artifact of using too few quantiles; it is a structural effect of conditioning on selective silos.

Together, Figures 1–3 illustrate the main failure mode motivating the federated protocol. The population-level score-distribution disparity is a property of the pooled group laws. Local audits condition on silo membership and can be close to, far below, or far above that population quantity depending on the allocation mechanism. A valid federated audit should therefore reconstruct the population group-mixture distributions rather than average, or visually inspect, local fairness metrics.

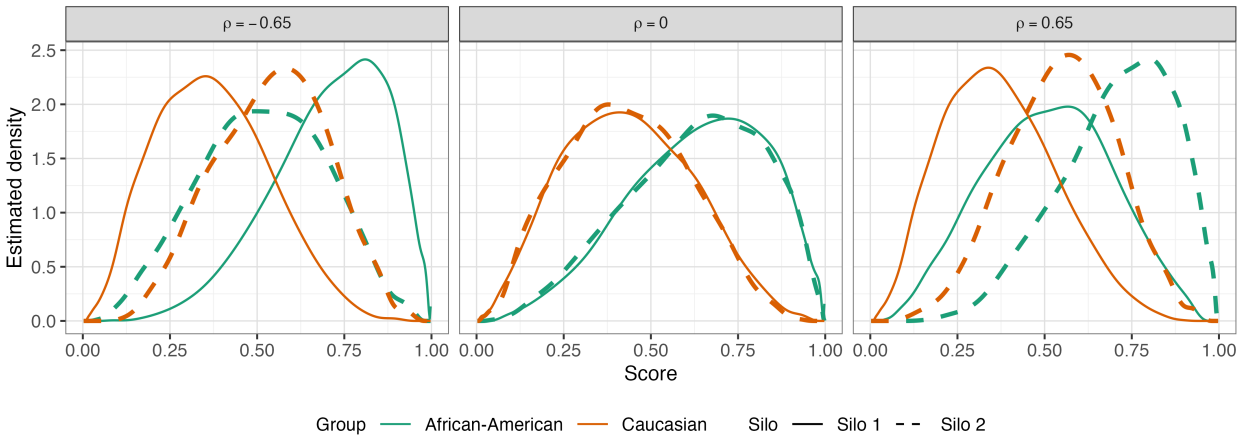


Figure 2: **Synthetic Beta experiment with two silos: score densities under different allocation regimes.** For each value of the selection parameter ρ, the figure shows kernel density estimates of the score distributions within each silo and sensitive group. Colors identify the sensitive groups, while line type identifies the silo. When $\rho = 0$, silo assignment is approximately independent of score, so each silo resembles the pooled group distributions. For $\rho > 0$, the two silos capture different score ranges, which can make the groups appear locally more similar within each silo despite unchanged pooled group distributions.

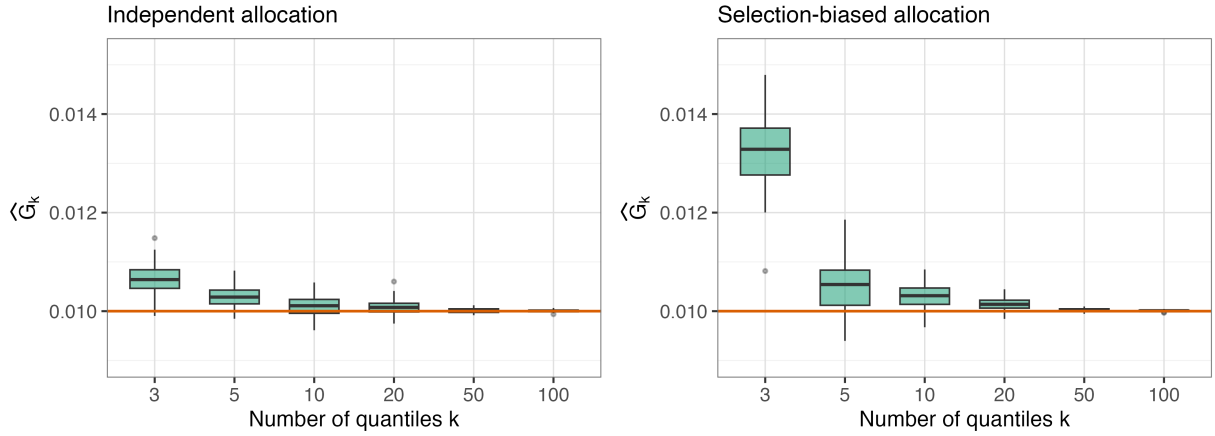


Figure 3: **Sensitivity to k on synthetic data.** Convergence of the quantile-grid estimator under independent allocation, left, and selection-biased allocation, right. The horizontal line is the centralized benchmark. The estimates stabilize as k increases, consistent with the deterministic finite-grid approximation term.

Table 3: **COMPAS, simulated silos with selection bias.** African-American (AA) vs. Caucasian (C) scores, with $d = 5$ silos.

	Global	Silo 1	Silo 2	Silo 3	Silo 4	Silo 5
n_{AA}	3696	535	704	745	814	898
n_{C}	2454	695	536	521	407	295
$\lvert\Delta\bar{z}\rvert$	1.6357	0.4483	0.9784	1.0978	0.9667	0.5654
U_2	0.7609	0.0748	0.2634	0.3073	0.2355	0.0745
H_2	0.0077	0.0017	0.0045	0.0044	0.0035	0.0012
W_2	1.7813	0.5479	1.0324	1.1319	1.0104	0.6326

5.3 COMPAS: simulated silos and local–global mismatch

We next analyze the COMPAS recidivism scores, restricting attention to African-American and Caucasian individuals. The score is the reported decile score, made continuous by adding a small independent jitter within each integer bin. This preserves the decile ordering while avoiding ties in empirical quantiles. After filtering, the centralized sample contains

$$n_{\mathrm{AA}} = 3696, \qquad n_{\mathrm{C}} = 2454.$$

We create $d = 5$ silos and consider allocation mechanisms that induce different degrees of selection bias. These silos are artificial: the goal is not to model the original institutional structure of the dataset, but to stress-test the audit under controlled local–global mismatch. This makes COMPAS complementary to ACSIncome, where silos are natural geographic units.

Table 3 reports one selection-biased allocation. The population disparity is much larger than the disparities observed within individual silos. This confirms that silo-level demographic-parity audits do not, in general, characterize the population-level audit.

The full COMPAS appendix reports additional allocation regimes and decomposition plots. Across these regimes, the quantile estimator stabilizes with a moderate number of grid points, and the diagnostic terms help distinguish whether the population gap is primarily due to composition effects, barycentric cross-silo heterogeneity, or their interaction.

5.4 ACSIncome: natural state silos

The most realistic federated experiment uses ACSIncome (Ding et al., 2021), a large-scale benchmark derived from the American Community Survey Public Use Microdata Sample. In contrast to the synthetic and COMPAS experiments, the silos are not simulated: individuals are partitioned by state, yielding 51 clients corresponding to the 50 U.S. states and Puerto Rico. We use the 2018 ACSIncome data and define the binary prediction target

$$Y = \mathbf{1}\{\texttt{PINCP} > 50{,}000\},$$

where PINCP denotes annual personal income. We train a ridge logistic regression model and use the predicted probability as the score $Z \in (0, 1)$. For each audit, the sensitive attribute being audited is excluded from the score model. We report two binary sensitive-group comparisons: SEX, comparing male and female individuals, and RAC1P, restricted to individuals coded as White alone or Black/African American alone[2].

Each state j releases only $n_{j,s}$ and k score quantiles per sensitive group. The centralized benchmark is computed from pooled scores and used only to evaluate the accuracy of the federated estimate.

Figure 4 shows that the federated estimate closely tracks the centralized reference for both sensitive attributes. This confirms, on a large dataset with natural silos, the empirical pattern observed in the controlled experiments: quantile sketches can recover the population-level score-distribution disparity with modest communication.

[2]See https://anonymous.4open.science/r/federated-6FE5/ for replication.

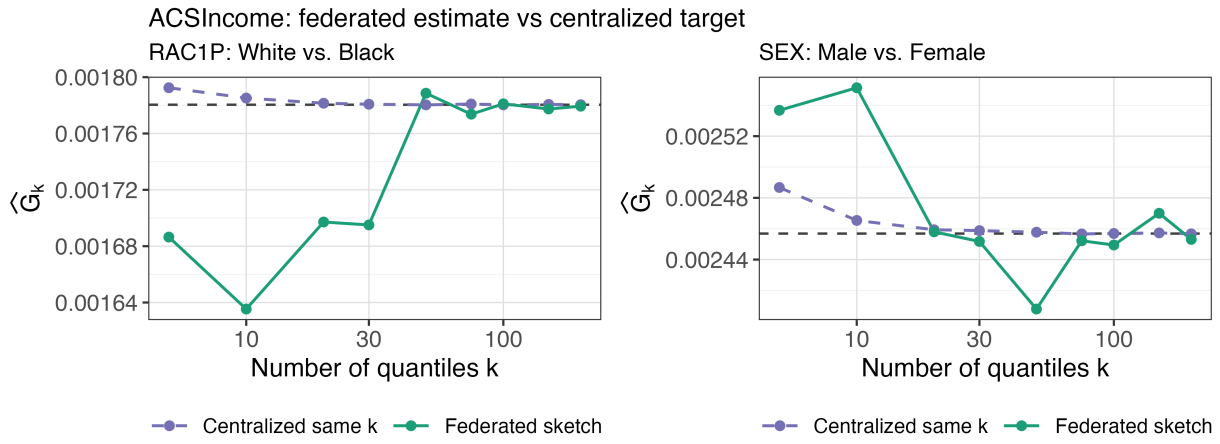


Figure 4: **`ACSIncome`: convergence of the federated quantile estimator.** Each state is treated as one client. The plots compare the federated estimate $\widehat{G}_{k,2}$, computed from k quantiles per state and sensitive group, with the centralized same-grid estimate computed from pooled scores.

Table 4: **`ACSIncome`, natural state silos.** Population-level benchmark and representative local audits. The score is the predicted probability of `PINCP` $> 50{,}000$, trained without the audited sensitive attribute.

Audit	Groups	Global U_2	CA	TX	FL	NY	PA
`SEX`	Male/Female	0.0024	0.0019	0.0030	0.0024	0.0018	0.0025
`RAC1P`	White/Black	0.0018	0.0005	0.0009	0.0027	0.0033	0.0015

Table 4 summarizes the population benchmark and representative local state audits. The key point is that local state values are not interchangeable with the global metric. In the `SEX` audit, some large states have local disparities below the pooled benchmark and others above it. In the `RAC1P` audit, the variation across states is stronger, reflecting both group-composition differences and cross-state differences in score distributions.

The full state-level local–global comparison is reported in Appendix G. It shows that state-level audits can lie both below and above the pooled benchmark. This is precisely the regime targeted by the federated audit: the server estimates the population-level score-distribution gap without accessing individual-level scores.

5.5 Empirical summary

Across datasets, the same qualitative pattern emerges. Local audits can substantially under- or over-estimate the population-level disparity. The proposed quantile protocol recovers the centralized target with modest communication cost. Finally, the $p = 2$ decomposition provides a diagnostic layer: it separates disparities driven by group-specific silo composition from disparities that persist after transport-averaging across silos, and it reports whether these mechanisms reinforce or offset each other through the interaction term.

6 Discussion and limitations

This paper studies how to audit population-level demographic disparity when score data are distributed across silos. The main message is that local fairness audits are not, in general, population audits. For a sensitive group, the population score law is a mixture of silo-specific score laws, with weights that may differ across groups. Aggregation can therefore create, amplify, attenuate, or mask distributional disparities. A valid federated audit must reconstruct the population-level object of interest, rather than average local metrics.

The proposed protocol provides a minimal communication interface for this task. Each silo releases subgroup counts and a fixed number of score quantiles per sensitive group. From these summaries, the server

Table 5: **Summary of experimental regimes.** Synthetic and `COMPAS` use simulated silo assignments to stress-test local–global mismatch. `ACSIncome` uses natural state-level clients.

Dataset	Silo structure	Sensitive comparison	d	Main role
Synthetic	simulated	two groups	variable	controlled selection mechanism
`COMPAS`	simulated	AA/Caucasian	5	fairness benchmark with controlled mismatch
`ACSIncome`	natural states	Male/Female	51	large-scale natural federation
`ACSIncome`	natural states	White/Black	51	large-scale natural federation

reconstructs group-mixture score distributions and estimates a Wasserstein–Fréchet disparity. For $p = 2$, the same summaries also yield a diagnostic decomposition into mixture/composition effects, barycentric cross-silo heterogeneity, and their interaction. The audit therefore provides both a global estimate of disparity and a diagnostic explanation of why local and global assessments may differ.

Auditing, not mitigation. Our contribution is an audit primitive. We do not propose a training algorithm, a post-processing rule, or a constraint for enforcing demographic parity. This distinction is important. A global audit may reveal a population-level disparity, but it does not determine by itself whether the appropriate response is to retrain the model, adjust thresholds, change data collection, modify deployment rules, or reconsider the target fairness criterion. Those choices depend on the application, the legal and institutional setting, and the costs of intervention. The role of the proposed method is to make the population-level disparity measurable under federated data constraints.

Score-level versus decision-level fairness. We audit score distributions rather than thresholded decisions. This is useful when downstream thresholds are client-specific, unknown to the central auditor, or adjusted over time. It is also relevant when the score itself is experienced by individuals, as in insurance pricing, credit scoring, clinical risk indices, ranking, or triage. However, score-level demographic parity is not a substitute for all fairness analyses. Decision-level criteria such as demographic parity after thresholding, equalized odds, predictive parity, and calibration may be more appropriate in some applications. The present framework can be extended to label-conditional audits by applying the same protocol to conditional laws such as

$$\mathcal{L}(Z \mid A = j, S = s, Y = y),$$

but this increases communication and creates additional small-cell disclosure risks.

Privacy and disclosure. The protocol is privacy-motivated, but it is not a formal privacy guarantee. It avoids centralizing individual records and individual scores, yet subgroup counts and quantile sketches can still leak information. If $n_{j,s} = 1$, all released quantiles coincide with the individual score; if $n_{j,s}$ is small, repeated or extreme quantiles may reveal partial information about local records. Practical deployment should therefore include safeguards such as minimum cell-size thresholds, suppression or coarsening of small cells, secure aggregation, and, where appropriate, calibrated noise. The stability bounds in this paper indicate how perturbations of counts and quantiles propagate to the audit estimates, but a full privacy–accuracy analysis, including formal differential privacy, is left for future work.

Choice of sketch size and uncertainty. The sketch size k controls the communication–accuracy trade-off. The theory gives a deterministic finite-grid term of order $O(1/k)$, and the experiments show that moderate sketch sizes can recover the centralized benchmark closely. Nevertheless, k should be treated as part of the audit design rather than as a purely technical parameter. We recommend reporting sensitivity curves in k, especially when underestimating disparity is costly. The present experiments focus on point estimation and finite-grid sensitivity. In deployment, uncertainty intervals could be obtained through a within-silo or silo-level bootstrap that respects the federated data structure; developing a full inference procedure is an important extension.

Theoretical scope. Several mathematical limitations remain. First, the method exploits the one-dimensional quantile representation of Wasserstein distances. Extending the same simplicity to multivariate

scores, vector-valued predictions, or structured outputs would require different summaries and different optimal-transport tools. Second, the clean rates rely on regularity assumptions such as compact support and densities bounded away from zero on the relevant trimmed region. The estimators remain well defined for discrete, tied, or clipped scores, but the sharp $O(1/k)$ approximation argument does not apply without qualification. Third, the ANOVA-style identity is specific to squared Wasserstein distance in one dimension. For $p = 1$, one can still derive triangle-inequality diagnostics, but not the exact expansion used here.

Empirical scope. The experiments cover controlled synthetic data, a standard fairness benchmark with simulated silos, and a large-scale natural federation based on state-level `ACSIncome` clients. These regimes illustrate the main local–global mismatch and show that quantile sketches can recover centralized benchmarks with modest communication. They do not exhaust all deployment settings. Real federated audits may involve more sensitive groups, smaller cells, nonstationary client populations, client-specific scoring rules, or institutional constraints on what summaries can be released. These settings would require additional reporting rules and, in some cases, stronger privacy safeguards.

Conclusion. Federated fairness auditing should target the population-level quantity of interest, not an average of local metrics. By combining distributional fairness functionals, quantile sketches, and a W_2^2 diagnostic decomposition, we provide a one-shot audit method for estimating global demographic disparity under data fragmentation. The results suggest that quantile sketches offer a practical interface between federated data constraints and population-level fairness measurement, while also making clear when local audits can be misleading.

Broader Impact Statement

This paper proposes a method for auditing demographic disparities when score data are distributed across institutions, jurisdictions, or clients and cannot be pooled in raw form. The intended positive impact is to make population-level fairness measurement more reliable in settings where silo-level audits may be misleading. By estimating score-distribution disparities from subgroup counts and quantile summaries, the protocol can help auditors detect aggregation effects that would be missed by local assessments alone, while avoiding the direct centralization of individual records or individual scores.

The method is relevant in domains where data fragmentation and selection effects are common, such as healthcare, insurance, credit, public administration, and criminal justice. In such settings, a scoring system may appear balanced within each local unit while producing a substantial population-level disparity after aggregation. The diagnostic decomposition can also help distinguish disparities driven by group-specific silo composition from disparities that persist across silos, which may support more informed monitoring and governance.

The work also has limitations and possible risks. First, demographic parity is only one fairness criterion. A small value of the proposed metric does not imply that a system is fair in all relevant respects, and a large value does not by itself identify the appropriate remedy. The protocol is an audit tool, not a mitigation procedure or a normative decision rule. Its outputs should therefore be interpreted alongside application-specific legal, ethical, and statistical considerations, as well as complementary decision-level and label-conditional fairness analyses when those are relevant.

Second, the protocol is privacy-motivated but does not provide a formal privacy guarantee. Subgroup counts and quantile summaries may leak information, especially for small cells, rare groups, or highly concentrated score distributions. Deployments should therefore include safeguards such as minimum cell-size thresholds, suppression or coarsening of small subgroups, secure aggregation, and, where appropriate, calibrated noise or formal differential-privacy mechanisms. The stability results in the paper help quantify how perturbations of communicated summaries affect audit accuracy, but a full privacy–accuracy analysis is left for future work.

Finally, fairness auditing can be misused if treated as a purely technical certification exercise. The proposed metric measures one population-level distributional property of scores. It should not be used to certify a system as fair without considering the broader institutional context, the choice of target variable, the legitimacy of the scoring task, the downstream decision rules, and the groups affected by deployment.

References

Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011. doi: 10.1137/100805741.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for federated learning on user-held data. In *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016)*. Curran Associates Inc., 2017.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression via plug-in estimator and recalibration with statistical guarantees. *Advances in Neural Information Processing Systems*, 33:19137–19148, 2020.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6478–6490, 2021.

Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. Fairness-aware agnostic federated learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining*, pp. 181–189, 2021. doi: 10.1137/1.9781611976700.21.

Ted Dunning and Otmar Ertl. Computing extremely accurate quantiles using t-digests. *arXiv*, 1902.04023, 2019.

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 1956. doi: 10.1214/aoms/1177728174.

Michael Greenwald and Sanjeev Khanna. Space-efficient online computation of quantile summaries. In *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, pp. 58–66, 2001. doi: 10.1145/375663.375670.

Steven A Julious and Mark A Mullee. Confounding and Simpson's paradox. *British Medical Journal*, 309 (6967):1480–1481, 1994.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

Zohar Karnin, Kevin Lang, and Edo Liberty. Optimal quantile approximation in streams. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, pp. 71–78, 2016. doi: 10.1109/FOCS.2016.17.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the COMPAS recidivism algorithm. *ProPublica*, 2016.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the Annual Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International conference on machine learning*, pp. 4615–4625. PMLR, 2019.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pp. 151–159, 2020.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Flavien Prost, Ben Packer, and Jilin Chen. Simpson's paradox in recommender fairness: Reconciling differences between per-user and aggregated evaluations. *arXiv*, 2210.07755, 2022.

Filippo Santambrogio. *Optimal transport for applied mathematicians*. Springer, 2015.

Christopher J Tosh and Daniel Hsu. Simple and near-optimal algorithms for hidden stratification and multi-group learning. In *International Conference on Machine Learning*, pp. 21633–21657. PMLR, 2022.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009. doi: 10.1007/978-3-540-71050-9.

A Additional details on quantile reconstruction

This appendix describes the finite-grid reconstructions induced by the communicated quantiles. The main protocol uses a stepwise reconstruction because it is simple, stable, and corresponds directly to an atomic distribution supported on the transmitted quantiles. We also describe an optional piecewise-linear reconstruction that uses the same communicated values and may reduce deterministic error when the underlying quantile functions are smooth.

A.1 Stepwise reconstruction

Fix $k \geq 1$ and the midpoint grid

$$u_\ell = \frac{\ell - \frac{1}{2}}{k}, \qquad \ell \in [k].$$

Let Q be a quantile function and write

$$q_\ell := Q(u_\ell).$$

The stepwise reconstruction used in the main protocol is the atomic law

$$\nu_k^{\mathrm{st}} = \frac{1}{k} \sum_{\ell=1}^{k} \delta_{q_\ell}.$$

Equivalently, its cdf is

$$F_k^{\mathrm{st}}(x) = \frac{1}{k} \sum_{\ell=1}^{k} \mathbf{1}\{q_\ell \leq x\}.$$

Its quantile function is constant on the bins

$$I_\ell = \left[\frac{\ell - 1}{k}, \frac{\ell}{k}\right), \qquad \ell \in [k],$$

namely

$$Q_k^{\mathrm{st}}(u) = q_\ell, \qquad u \in I_\ell.$$

This reconstruction has two practical advantages. First, mixtures of reconstructed local laws are weighted empirical distributions over the transmitted support points, so the server can form each group-mixture cdf by sorting dk values. Second, Cramér distances between reconstructed laws can be computed exactly by integrating between consecutive jump locations.

A.2 Piecewise-linear reconstruction

A smoother server-side alternative is to linearly interpolate the quantile function between consecutive communicated knots. For $u \in [u_\ell, u_{\ell+1}]$, define

$$Q_k^{\mathrm{lin}}(u) = q_\ell + \frac{u - u_\ell}{u_{\ell+1} - u_\ell}(q_{\ell+1} - q_\ell) = q_\ell + k(u - u_\ell)(q_{\ell+1} - q_\ell), \qquad \ell = 1, \ldots, k-1.$$

On the two boundary half-bins, use the constant extension

$$Q_k^{\mathrm{lin}}(u) = q_1 \quad \text{for } u \in [0, u_1], \qquad Q_k^{\mathrm{lin}}(u) = q_k \quad \text{for } u \in [u_k, 1].$$

This reconstruction uses exactly the same communicated quantiles as the stepwise version. It only changes the server-side interpolation and integration rule.

A.3 Approximation bounds

Let

$$\omega_Q(\delta) := \sup\{|Q(u) - Q(v)| : |u - v| \le \delta\}$$

be the modulus of continuity of Q. The following deterministic bounds concern the reconstruction error of a fixed quantile function.

Proposition A.1 (Stepwise and linear reconstruction errors)**.** *Let Q be nondecreasing on $[0, 1]$. Then, on the interior interval $[u_1, u_k]$,*

$$\|Q_k^{\mathrm{st}} - Q\|_{\infty,[u_1,u_k]} \le \omega_Q(1/k),$$

and

$$\|Q_k^{\mathrm{lin}} - Q\|_{\infty,[u_1,u_k]} \le \omega_Q(1/k).$$

If Q is L-Lipschitz on $[0, 1]$, both errors are at most L/k. If, in addition, $Q \in C^2([u_1, u_k])$, then

$$\|Q_k^{\mathrm{lin}} - Q\|_{\infty,[u_1,u_k]} \le \frac{\|Q''\|_{\infty,[u_1,u_k]}}{8k^2}.$$

Proof. For the stepwise reconstruction, each $u \in I_\ell$ is at distance at most $1/(2k)$ from u_ℓ, and the stated bound follows from the modulus of continuity. For the linear reconstruction, on each interval $[u_\ell, u_{\ell+1}]$, the interpolant is bounded between the endpoint values because Q is nondecreasing. Both endpoints are within distance $1/k$ of any point in the interval, so the same modulus bound applies. If Q is L-Lipschitz, this gives the L/k bound. If $Q \in C^2$ on the interior interval, the standard linear-interpolation error formula on a grid with mesh $1/k$ gives the second-order bound. □

The $O(k^{-2})$ term is an interior smoothness statement. With the constant extension on $[0, u_1] \cup [u_k, 1]$, boundary contributions may remain first order unless additional tail or boundary corrections are imposed. This is why the main protocol reports sensitivity to k rather than relying on a formal second-order guarantee in all settings.

A.4 Exact integration for U_2 under linear reconstruction

For $p = 2$, the Wasserstein–Fréchet functional is an integral of squared quantile differences. The midpoint estimator used in the main protocol is

$$U_{k,2}^{\mathrm{mid}} = \sum_{s \in \mathcal{S}} \alpha_s \frac{1}{k} \sum_{\ell=1}^{k} (q_{s,\ell} - q_\ell^*)^2, \qquad q_\ell^* = \sum_{s \in \mathcal{S}} \alpha_s q_{s,\ell}.$$

This estimator is simple and matches the stepwise reconstruction. If one instead uses the piecewise-linear reconstruction of the group quantile functions, then the squared differences are quadratic between knots and can be integrated exactly.

Let

$$d_{s,\ell} := q_{s,\ell} - q^*_\ell.$$

Under the constant boundary extension described above,

$$\int_0^1 d^{\text{lin}}_{s,k}(u)^2\, du = \frac{1}{2k} d^2_{s,1} + \sum_{\ell=1}^{k-1} \frac{1}{3k} \left(d^2_{s,\ell} + d_{s,\ell} d_{s,\ell+1} + d^2_{s,\ell+1} \right) + \frac{1}{2k} d^2_{s,k}. \tag{39}$$

Therefore the corresponding linear-exact estimator is

$$U^{\text{lin}}_{k,2} = \sum_{s\in\mathcal{S}} \alpha_s \left[\frac{1}{2k} d^2_{s,1} + \sum_{\ell=1}^{k-1} \frac{1}{3k} \left(d^2_{s,\ell} + d_{s,\ell} d_{s,\ell+1} + d^2_{s,\ell+1} \right) + \frac{1}{2k} d^2_{s,k} \right]. \tag{40}$$

Proposition A.2 (Interior second-order deterministic accuracy)**.** *Assume that each group quantile function Q_s belongs to $C^2([u_1, u_k])$, with uniformly bounded second derivative on this interval. Then the contribution of the interior interval $[u_1, u_k]$ to*

$$\left| U^{\text{lin}}_{k,2} - U_2 \right|$$

is $O(k^{-2})$. With the constant boundary extension on $[0, u_1] \cup [u_k, 1]$, the full error may also contain first-order boundary contributions unless additional boundary assumptions or corrections are imposed.

Proof. Since

$$Q^*(u) = \sum_{s\in\mathcal{S}} \alpha_s Q_s(u),$$

the barycenter quantile is also C^2 on $[u_1, u_k]$. Proposition A.1 gives $O(k^{-2})$ uniform error on the interior interval for each reconstructed Q_s and for the reconstructed barycenter quantile. A mean-value bound for squared differences then gives

$$\int_{u_1}^{u_k} \left| \left(Q^{\text{lin}}_{s,k}(u) - Q^{\text{lin}\,*}_k(u) \right)^2 - \left(Q_s(u) - Q^*(u) \right)^2 \right| du = O(k^{-2}),$$

because the relevant quantile functions are bounded on the compact interior interval. Formula (39) removes quadrature error by integrating the squared piecewise-linear reconstruction exactly. Summing over s with weights α_s gives the stated interior contribution. The boundary intervals are controlled separately and may contribute at first order under the constant extension. □

Practical implication. The main protocol uses the midpoint stepwise estimator because it is simple, stable, and well defined for arbitrary score distributions, including discrete or tied scores. The linear-exact estimator (40) is an optional server-side refinement: it uses the same communicated quantiles and can reduce deterministic error when score quantile functions are smooth on the interior of the unit interval. In empirical audits, reporting the sensitivity of $\widehat{G}_{k,2}$ to k remains the most direct way to assess whether the sketch resolution is adequate.

B Technical lemmas

This appendix collects the analytic facts used in the proofs. They are standard consequences of one-dimensional optimal transport, midpoint approximation, empirical-process concentration, and elementary Lipschitz stability. We state them in the notation of the paper for completeness.

B.1 Wasserstein distance and barycenters in one dimension

For a probability measure μ on $\mathbb{R}$, let

$$Q_\mu(u) := \inf\{x \in \mathbb{R} : F_\mu(x) \geq u\}, \qquad u \in (0, 1),$$

be its generalized inverse cdf.

Lemma B.1 (Wasserstein distance via quantiles)**.** *Let $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$, with $p \geq 1$. Then*

$$W_p(\mu, \nu)^p = \int_0^1 |Q_\mu(u) - Q_\nu(u)|^p \, du.$$

In particular, for $p = 2$,

$$W_2(\mu, \nu)^2 = \|Q_\mu - Q_\nu\|^2_{L^2([0,1])}.$$

This representation is classical in one-dimensional optimal transport; see, for example, Villani (2009) and Santambrogio (2015).

Lemma B.2 (One-dimensional W_2 barycenter)**.** *Let $\mu_1, \ldots, \mu_m \in \mathcal{P}_2(\mathbb{R})$, and let $\lambda_1, \ldots, \lambda_m \geq 0$ with $\sum_{r=1}^m \lambda_r = 1$. The unique W_2 barycenter μ^* satisfies*

$$Q_{\mu^*}(u) = \sum_{r=1}^m \lambda_r Q_{\mu_r}(u), \qquad u \in (0, 1).$$

Moreover,

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R})} \sum_{r=1}^m \lambda_r W_2(\mu_r, \mu)^2 = \sum_{r=1}^m \lambda_r \int_0^1 \left(Q_{\mu_r}(u) - \sum_{r'=1}^m \lambda_{r'} Q_{\mu_{r'}}(u) \right)^2 du.$$

Proof. By Lemma B.1, the barycenter problem is equivalent to minimizing

$$\int_0^1 \sum_{r=1}^m \lambda_r \left(Q_{\mu_r}(u) - q(u) \right)^2 du$$

over quantile functions q. Pointwise in u, the minimizer is

$$q^*(u) = \sum_{r=1}^m \lambda_r Q_{\mu_r}(u).$$

This function is nondecreasing because it is a nonnegative linear combination of nondecreasing functions. Hence it is itself a quantile function and defines a probability law. Substituting q^* into the objective gives the displayed value. □

The formula is the one-dimensional specialization of the Wasserstein barycenter construction; see Agueh & Carlier (2011).

B.2 Riemann approximation on the midpoint grid

The midpoint grid used throughout the paper is

$$u_\ell = \frac{\ell - \frac{1}{2}}{k}, \qquad \ell \in [k].$$

Lemma B.3 (Midpoint rule for Lipschitz functions)**.** *Let $g : [0, 1] \to \mathbb{R}$ be L_g-Lipschitz. Then*

$$\left| \int_0^1 g(u) \, du - \frac{1}{k} \sum_{\ell=1}^k g(u_\ell) \right| \leq \frac{L_g}{2k}.$$

Proof. Let $I_\ell = [(\ell - 1)/k, \ell/k]$. For $u \in I_\ell$,

$$|u - u_\ell| \leq \frac{1}{2k},$$

so

$$|g(u) - g(u_\ell)| \le \frac{L_g}{2k}.$$

Therefore,

$$\left| \int_{I_\ell} g(u)\, du - \frac{1}{k} g(u_\ell) \right| \le \frac{L_g}{2k^2}.$$

Summing over $\ell = 1, \dots, k$ gives the result. □

B.3 Step-cdf approximation from quantile knots

Let F be a cdf on $\mathbb{R}$ with generalized inverse Q. For the midpoint grid $u_\ell = (\ell - \frac{1}{2})/k$, set

$$q_\ell = Q(u_\ell), \qquad F_k(x) := \frac{1}{k} \sum_{\ell=1}^{k} \mathbf{1}\{q_\ell \le x\}.$$

Lemma B.4 (Uniform cdf approximation from midpoint quantiles)**.** *For every cdf F,*

$$\|F_k - F\|_\infty \le \frac{1}{2k}.$$

Proof. For any $x \in \mathbb{R}$, the generalized inverse satisfies

$$Q(u) \le x \quad \Longleftrightarrow \quad u \le F(x).$$

Hence

$$F_k(x) = \frac{1}{k} \#\{\ell : u_\ell \le F(x)\}.$$

The empirical cdf of the deterministic midpoint grid approximates the uniform cdf on $[0, 1]$ with supremum error at most $1/(2k)$. Therefore

$$|F_k(x) - F(x)| \le \frac{1}{2k}.$$

Taking the supremum over x proves the claim. □

Lemma B.5 (Step-cdf perturbation by knot error)**.** *Let $q_1 \le \cdots \le q_k$ and $q'_1 \le \cdots \le q'_k$ satisfy*

$$\max_{\ell \in [k]} |q_\ell - q'_\ell| \le \eta.$$

Define

$$F_q(x) = \frac{1}{k} \sum_{\ell=1}^{k} \mathbf{1}\{q_\ell \le x\}, \qquad F_{q'}(x) = \frac{1}{k} \sum_{\ell=1}^{k} \mathbf{1}\{q'_\ell \le x\}.$$

Then, for every $x \in \mathbb{R}$,

$$|F_q(x) - F_{q'}(x)| \le \frac{1}{k} \#\{\ell \in [k] : q_\ell \in (x - \eta, x + \eta]\}.$$

Proof. The indicators $\mathbf{1}\{q_\ell \le x\}$ and $\mathbf{1}\{q'_\ell \le x\}$ can differ only if q_ℓ and q'_ℓ lie on opposite sides of x. Since $|q_\ell - q'_\ell| \le \eta$, this can occur only when $q_\ell \in (x - \eta, x + \eta]$. Summing over ℓ and dividing by k gives the bound. □

Lemma B.6 (Counting quantile knots in short intervals)**.** *Assume F is continuous, and let $q_\ell = Q(u_\ell)$, where $u_\ell = (\ell - \frac{1}{2})/k$. For any $a < b$,*

$$\frac{1}{k} \#\{\ell : q_\ell \in (a, b]\} \le F(b) - F(a) + \frac{1}{k}.$$

Consequently, if F is M-Lipschitz on an interval containing $[x - \eta, x + \eta]$, then

$$\frac{1}{k} \#\{\ell : q_\ell \in (x - \eta, x + \eta]\} \le 2M\eta + \frac{1}{k}.$$

Proof. If $q_\ell \in (a, b]$, then $u_\ell \le F(b)$, and, by continuity of F, levels strictly below $F(a)$ cannot generate quantiles in $(a, b]$. Thus the relevant grid levels lie in an interval of length at most $F(b) - F(a)$, up to one boundary grid point. Since the grid spacing is $1/k$, their number is at most $k\{F(b) - F(a)\} + 1$. Dividing by k gives the first claim. The second follows from

$$F(x + \eta) - F(x - \eta) \le 2M\eta.$$

□

B.4 Empirical cdf and quantile concentration

Lemma B.7 (DKW–Massart inequality)**.** *Let $\widehat{F}_n$ be the empirical cdf of n i.i.d. samples from a distribution with cdf F. Then, for every $t > 0$,*

$$\mathbb{P}\left(\|\widehat{F}_n - F\|_\infty > t\right) \le 2e^{-2nt^2}.$$

Equivalently, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\|\widehat{F}_n - F\|_\infty \le \sqrt{\frac{1}{2n} \log \frac{2}{\delta}}.$$

The sharp constant form is due to Massart (1990), building on Dvoretzky et al. (1956).

Lemma B.8 (Quantile bracketing under cdf perturbations)**.** *Let F and G be cdfs, and set*

$$\Delta := \|F - G\|_\infty.$$

Then, for every $u \in (\Delta, 1 - \Delta)$,

$$F^{-1}(u - \Delta) \le G^{-1}(u) \le F^{-1}(u + \Delta),$$

where F^{-1} and G^{-1} denote generalized inverse cdfs.

Proof. Let $x_+ = F^{-1}(u + \Delta)$. Then $F(x_+) \ge u + \Delta$, hence

$$G(x_+) \ge F(x_+) - \Delta \ge u.$$

Thus $G^{-1}(u) \le x_+$. Similarly, if $x < F^{-1}(u - \Delta)$, then $F(x) < u - \Delta$, so $G(x) < u$. Hence $G^{-1}(u) \ge F^{-1}(u - \Delta)$. □

Lemma B.9 (Lipschitz inversion on a trimmed region)**.** *Fix $\varepsilon \in (0, 1/2)$. Suppose that F is absolutely continuous on*

$$[Q_F(\varepsilon/2), Q_F(1 - \varepsilon/2)]$$

and that its density satisfies $f(x) \ge m_\varepsilon > 0$ on this interval. Let G be another cdf, and assume

$$\|F - G\|_\infty \le \varepsilon/2.$$

Then, for all $u \in [\varepsilon, 1 - \varepsilon]$,

$$|Q_G(u) - Q_F(u)| \le \frac{1}{m_\varepsilon} \|F - G\|_\infty.$$

Proof. Let $\Delta = \|F - G\|_\infty$. Since $\Delta \le \varepsilon/2$, the points $u - \Delta$ and $u + \Delta$ belong to $[\varepsilon/2, 1 - \varepsilon/2]$. By Lemma B.8,

$$Q_F(u - \Delta) \le Q_G(u) \le Q_F(u + \Delta).$$

On the enlarged trimmed region, $f \ge m_\varepsilon$, so Q_F is $1/m_\varepsilon$-Lipschitz. Hence

$$|Q_F(u + \Delta) - Q_F(u)| \le \frac{\Delta}{m_\varepsilon}, \qquad |Q_F(u) - Q_F(u - \Delta)| \le \frac{\Delta}{m_\varepsilon}.$$

The result follows. □

B.5 Lipschitz stability of the discrete W_2^2 functional

The next lemma propagates perturbations in reconstructed quantiles and group weights to perturbations of the finite-grid W_2^2 disparity.

Let $q = (q_{s,\ell})_{s\in\mathcal{S},\ell\in[k]}$ be an array of quantiles, and let $\alpha = (\alpha_s)_{s\in\mathcal{S}}$ be a probability vector. Define

$$\Phi_\alpha(q) := \sum_{s\in\mathcal{S}} \alpha_s \frac{1}{k} \sum_{\ell=1}^{k} \left(q_{s,\ell} - \sum_{r\in\mathcal{S}} \alpha_r q_{r,\ell} \right)^2 .$$

Equivalently,

$$\Phi_\alpha(q) = \frac{1}{k} \sum_{\ell=1}^{k} \left[\sum_{s\in\mathcal{S}} \alpha_s q_{s,\ell}^2 - \left(\sum_{s\in\mathcal{S}} \alpha_s q_{s,\ell} \right)^2 \right] .$$

Lemma B.10 (Stability of the discrete W_2^2 functional)**.** *Assume that all entries of q and q' lie in $[-B, B]$, and that α, α' are probability vectors on $\mathcal{S}$. Then*

$$|\Phi_\alpha(q) - \Phi_{\alpha'}(q')| \le 4B\|q - q'\|_\infty + 3B^2\|\alpha - \alpha'\|_1.$$

Proof. We split the difference as

$$|\Phi_\alpha(q) - \Phi_{\alpha'}(q')| \le |\Phi_\alpha(q) - \Phi_\alpha(q')| + |\Phi_\alpha(q') - \Phi_{\alpha'}(q')|.$$

First fix α. For each grid point ℓ, write

$$m_\ell(q, \alpha) = \sum_{s\in\mathcal{S}} \alpha_s q_{s,\ell}.$$

The contribution of grid point ℓ is

$$\phi_\ell(q, \alpha) = \sum_{s} \alpha_s q_{s,\ell}^2 - m_\ell(q, \alpha)^2.$$

Its derivative with respect to $q_{t,\ell}$ is

$$\frac{\partial \phi_\ell}{\partial q_{t,\ell}} = 2\alpha_t \{q_{t,\ell} - m_\ell(q, \alpha)\}.$$

Since all entries lie in $[-B, B]$, we have $|q_{t,\ell} - m_\ell(q, \alpha)| \le 2B$, hence

$$\sum_{t\in\mathcal{S}} \left| \frac{\partial \phi_\ell}{\partial q_{t,\ell}} \right| \le 4B.$$

The mean-value theorem gives

$$|\Phi_\alpha(q) - \Phi_\alpha(q')| \le 4B\|q - q'\|_\infty.$$

Now fix q'. For each ℓ, the derivative of $\phi_\ell(q', \alpha)$ with respect to α_t is

$$\frac{\partial \phi_\ell}{\partial \alpha_t} = {q'_{t,\ell}}^2 - 2m_\ell(q', \alpha) q'_{t,\ell}.$$

Since $|q'_{t,\ell}| \le B$ and $|m_\ell(q', \alpha)| \le B$, this derivative is bounded in absolute value by $3B^2$. Therefore

$$|\Phi_\alpha(q') - \Phi_{\alpha'}(q')| \le 3B^2\|\alpha - \alpha'\|_1.$$

Combining the two bounds proves the claim. □

C Proofs for the distributional fairness target

This appendix proves the basic properties of the population-level functionals introduced in Section 2, as well as the centralized quantile-sketch approximation result used in Section 3.

C.1 Zero disparity and score-level demographic parity

Proposition C.1 (Zero Wasserstein–Fréchet disparity)**.** *Assume that $\alpha_s > 0$ for all $s \in \mathcal{S}$. Then*

$$U_p = 0 \quad \Longleftrightarrow \quad \nu_s = \nu_{s'} \quad \text{for all } s, s' \in \mathcal{S}.$$

Consequently, $U_p = 0$ if and only if score-level demographic parity holds.

Proof. If all group laws coincide, say $\nu_s = \nu_0$ for all s, then choosing $\nu = \nu_0$ in the definition of U_p gives

$$U_p \leq \sum_s \alpha_s W_p(\nu_s, \nu_0)^p = 0.$$

Since $U_p \geq 0$, this proves $U_p = 0$.

Conversely, suppose $U_p = 0$. Let ν^* be a Wasserstein barycenter attaining the infimum in (5). Then

$$0 = \sum_s \alpha_s W_p(\nu_s, \nu^*)^p.$$

Each term is nonnegative and each $\alpha_s > 0$. Hence

$$W_p(\nu_s, \nu^*) = 0 \qquad \text{for every } s \in \mathcal{S}.$$

Since W_p is a metric on $\mathcal{P}_p(\mathbb{R})$, this implies

$$\nu_s = \nu^* \qquad \text{for every } s \in \mathcal{S}.$$

Thus all group score laws coincide. □

Proposition C.2 (Zero Cramér heterogeneity)**.** *Assume that $\alpha_s > 0$ for all $s \in \mathcal{S}$. Then*

$$H_p = 0 \quad \Longleftrightarrow \quad \nu_s = \nu_{s'} \quad \text{for all } s, s' \in \mathcal{S}.$$

Proof. If all group laws coincide, then the pooled law

$$\nu^{\text{pool}} = \sum_s \alpha_s \nu_s$$

is equal to this common law. Hence $H_p = 0$.

Conversely, suppose $H_p = 0$. Then

$$0 = \sum_s \alpha_s C_p(\nu_s, \nu^{\text{pool}})^p.$$

Since all terms are nonnegative and all weights are positive,

$$C_p(\nu_s, \nu^{\text{pool}}) = 0 \qquad \text{for every } s.$$

Thus

$$F_s(x) = F_{\text{pool}}(x)$$

for Lebesgue-almost every x. Because cdfs are nondecreasing and right-continuous, equality almost everywhere implies equality everywhere. Indeed, if two right-continuous monotone functions differed at some point, right-continuity would imply that they differ on an interval of positive Lebesgue measure. Therefore $F_s = F_{\text{pool}}$ for every s, and hence

$$\nu_s = \nu^{\text{pool}} \qquad \text{for every } s.$$

The group laws all coincide. □

C.2 Two-group simplifications

Proposition C.3 (Two-group form of U_2)**.** *Let $\mathcal{S} = \{0, 1\}$, with $\alpha_0, \alpha_1 > 0$ and $\alpha_0 + \alpha_1 = 1$. Then*

$$U_2 = \alpha_0 \alpha_1 W_2(\nu_0, \nu_1)^2.$$

Proof. By Lemma B.2, the W_2 barycenter has quantile function

$$Q^*(u) = \alpha_0 Q_0(u) + \alpha_1 Q_1(u).$$

Therefore

$$Q_0(u) - Q^*(u) = \alpha_1 \{Q_0(u) - Q_1(u)\},$$

and

$$Q_1(u) - Q^*(u) = -\alpha_0 \{Q_0(u) - Q_1(u)\}.$$

Hence

$$\begin{aligned} U_2 &= \alpha_0 \int_0^1 (Q_0(u) - Q^*(u))^2 \, du + \alpha_1 \int_0^1 (Q_1(u) - Q^*(u))^2 \, du \\ &= \alpha_0 \alpha_1^2 \int_0^1 (Q_0(u) - Q_1(u))^2 \, du + \alpha_1 \alpha_0^2 \int_0^1 (Q_0(u) - Q_1(u))^2 \, du \\ &= \alpha_0 \alpha_1 (\alpha_0 + \alpha_1) \int_0^1 (Q_0(u) - Q_1(u))^2 \, du \\ &= \alpha_0 \alpha_1 W_2(\nu_0, \nu_1)^2. \end{aligned}$$

□

Proposition C.4 (Two-group form of H_2)**.** *Let $\mathcal{S} = \{0, 1\}$, with $\alpha_0, \alpha_1 > 0$ and $\alpha_0 + \alpha_1 = 1$. Then*

$$H_2 = \alpha_0 \alpha_1 C_2(\nu_0, \nu_1)^2.$$

Proof. Let

$$F_{\text{pool}} = \alpha_0 F_0 + \alpha_1 F_1.$$

Then

$$F_0 - F_{\text{pool}} = \alpha_1 (F_0 - F_1), \qquad F_1 - F_{\text{pool}} = -\alpha_0 (F_0 - F_1).$$

Therefore

$$\begin{aligned} H_2 &= \alpha_0 \int_{\mathbb{R}} (F_0(x) - F_{\text{pool}}(x))^2 dx + \alpha_1 \int_{\mathbb{R}} (F_1(x) - F_{\text{pool}}(x))^2 dx \\ &= \alpha_0 \alpha_1^2 \int_{\mathbb{R}} (F_0(x) - F_1(x))^2 dx + \alpha_1 \alpha_0^2 \int_{\mathbb{R}} (F_0(x) - F_1(x))^2 dx \\ &= \alpha_0 \alpha_1 C_2(\nu_0, \nu_1)^2. \end{aligned}$$

□

C.3 Exact computation of the Cramér sketch metric

Proposition C.5 (Step-function computation of C_p)**.** *Let F and G be right-continuous step cdfs with jump locations contained in $\{g_1, \ldots, g_M\}$, where $g_1 < \cdots < g_M$. Then*

$$C_p(F, G)^p = \sum_{m=1}^{M-1} (g_{m+1} - g_m) |F(g_m) - G(g_m)|^p.$$

Proof. Between two consecutive jump locations g_m and g_{m+1}, both step functions are constant. For $x \in [g_m, g_{m+1})$,

$$F(x) = F(g_m), \qquad G(x) = G(g_m).$$

Thus

$$\int_{g_m}^{g_{m+1}} |F(x) - G(x)|^p dx = (g_{m+1} - g_m)|F(g_m) - G(g_m)|^p.$$

Summing over $m = 1, \ldots, M-1$ gives the formula. Outside the smallest and largest jump locations, both cdfs are equal to either 0 or 1, so the contribution is zero. □

C.4 Proof of the centralized sketch approximation

We now prove the deterministic $O(1/k)$ approximation bound stated in Proposition 3.1.

Proof of Proposition 3.1. We prove the Wasserstein and Cramér bounds separately, and then record the fixed-k plug-in consistency.

Wasserstein approximation. First consider $p = 2$. By Lemma B.2, the population barycenter has quantile function

$$q^*(u) = \sum_{s \in \mathcal{S}} \alpha_s Q_s(u).$$

Define

$$g(u) := \sum_{s \in \mathcal{S}} \alpha_s \left|Q_s(u) - q^*(u)\right|^2.$$

Then

$$U_2 = \int_0^1 g(u)\, du, \qquad U_{k,2} = \frac{1}{k} \sum_{\ell=1}^{k} g(u_\ell).$$

Under the assumptions of the proposition, each group law has compact support and a density bounded away from zero on its support. Hence each Q_s is Lipschitz on $[0, 1]$, with a finite Lipschitz constant depending on the lower density bound. The quantile functions are also uniformly bounded by compact support. Therefore q^* is Lipschitz and bounded, and so is g. Lemma B.3 yields

$$|U_2 - U_{k,2}| \le \frac{C_U}{k}$$

for a finite constant C_U.

For $p = 1$, the same argument applies with

$$g(u) := \sum_{s \in \mathcal{S}} \alpha_s |Q_s(u) - q^*(u)|,$$

where $q^*(u)$ is a weighted median of the values

$$\{Q_s(u) : s \in \mathcal{S}\}.$$

The weighted median set is stable under sup-norm perturbations of its inputs, and one may choose a Lipschitz version q^* when all Q_s's are Lipschitz. Hence g is Lipschitz, and the midpoint rule again gives

$$|U_1 - U_{k,1}| \le \frac{C_U}{k}.$$

Cramér approximation. Let

$$F_{\text{pool}}(x) = \sum_{s \in \mathcal{S}} \alpha_s F_s(x)$$

be the pooled cdf, and let

$$F_k^{\text{pool}}(x) = \sum_{s \in \mathcal{S}} \alpha_s F_{s,k}(x)$$

be the pooled cdf built from the midpoint reconstructions. By Lemma B.4,

$$\|F_{s,k} - F_s\|_\infty \le \frac{1}{2k} \qquad \text{for all } s.$$

Therefore

$$\|F_k^{\text{pool}} - F_{\text{pool}}\|_\infty \le \sum_s \alpha_s \|F_{s,k} - F_s\|_\infty \le \frac{1}{2k}.$$

Because all group laws have compact support, there exists an interval $[a, b]$ containing the supports of all ν_s and all reconstructed laws $\nu_{s,k}$. Outside $[a, b]$, the relevant cdf differences vanish.

For $p = 1$, the reverse triangle inequality gives

$$\Big| |F_s - F_{\text{pool}}| - |F_{s,k} - F_k^{\text{pool}}| \Big| \le |F_s - F_{s,k}| + |F_{\text{pool}} - F_k^{\text{pool}}| \le \frac{1}{k}.$$

Integrating over $[a, b]$, multiplying by α_s, and summing over s yields

$$|H_1 - H_{k,1}| \le \frac{b-a}{k}.$$

For $p = 2$, use

$$|x^2 - y^2| = |x - y||x + y| \le 2|x - y|$$

for $x, y \in [-1, 1]$. Since all cdf differences lie in $[-1, 1]$,

$$\begin{aligned} &\Big| (F_s - F_{\text{pool}})^2 - (F_{s,k} - F_k^{\text{pool}})^2 \Big| \\ &\quad \le 2 \Big| (F_s - F_{\text{pool}}) - (F_{s,k} - F_k^{\text{pool}}) \Big| \\ &\quad \le 2 \Big(|F_s - F_{s,k}| + |F_{\text{pool}} - F_k^{\text{pool}}| \Big) \le \frac{2}{k}. \end{aligned}$$

Integrating over $[a, b]$, multiplying by α_s, and summing over s gives

$$|H_2 - H_{k,2}| \le \frac{2(b-a)}{k}.$$

This proves the claimed $O(1/k)$ Cramér approximation bound for $p \in \{1, 2\}$.

Fixed-k empirical plug-in consistency. For fixed k, the plug-in estimators depend on finitely many empirical quantiles and empirical group weights. Under the continuity and positive-density assumptions, each empirical quantile

$$\widehat{Q}_s(u_\ell)$$

is consistent for $Q_s(u_\ell)$, and the empirical weights $\widehat{\alpha}_s$ are consistent for α_s. The finite-grid functionals $U_{k,p}$ and $H_{k,p}$ are continuous functions of these finitely many quantiles and weights, as long as the relevant groups have positive probability. The continuous mapping theorem therefore gives convergence in probability of the plug-in estimators to their finite-grid population counterparts. □

D Proofs for the federated protocol

This appendix proves the main properties of the federated estimator introduced in Section 3. We first prove the population identity $G_p = U_p$, then define the population finite-grid target $G_{k,p}$, and finally prove the consistency, rate, and high-probability stability results.

D.1 Agreement between federated and centralized targets

Proof of Proposition 3.2. By the law of total probability, for every sensitive group $s \in \mathcal{S}$,

$$\nu_s = \mathcal{L}(Z \mid S = s) = \sum_{j=1}^{d} \mathbb{P}(A = j \mid S = s)\, \mathcal{L}(Z \mid A = j, S = s).$$

Using the notation of the paper, this is

$$\nu_s = \sum_{j=1}^{d} \pi_{j|s} \nu_{j,s} = \nu_s^{\times}.$$

Therefore the collection of group laws used in the centralized functional U_p is identical to the collection of group-mixture laws used in the federated functional G_p. The two Wasserstein–Fréchet optimization problems are the same, and hence

$$G_p = U_p.$$

The same argument applies to the Cramér companion metric. Since $\nu_s^{\times} = \nu_s$ for every s, the pooled laws also agree:

$$\nu^{\text{pool},\times} = \sum_{s \in \mathcal{S}} \alpha_s \nu_s^{\times} = \sum_{s \in \mathcal{S}} \alpha_s \nu_s = \nu^{\text{pool}}.$$

Thus

$$H_p^{\times} = H_p.$$

□

D.2 Population finite-grid federated target

For the convergence proof, it is useful to define the population finite-grid version of the federated estimator. This is the quantity that would be obtained if the server received exact population local quantiles $Q_{j,s}(u_\ell)$ and exact weights $\alpha_s, \pi_{j|s}$, but still used only the k-point grid.

For each (j, s), define the population step cdf

$$F_{j,s,k}(x) := \frac{1}{k} \sum_{\ell=1}^{k} \mathbf{1}\{Q_{j,s}(u_\ell) \le x\}.$$

Let $\nu_{j,s,k}$ be the corresponding atomic law. For each sensitive group s, define the finite-grid mixture law

$$\nu_{s,k}^{\times} := \sum_{j=1}^{d} \pi_{j|s} \nu_{j,s,k},$$

with cdf

$$F_{s,k}^{\times}(x) = \sum_{j=1}^{d} \pi_{j|s} F_{j,s,k}(x).$$

Let

$$q_{s,\ell}^{\times} := \inf\{x : F_{s,k}^{\times}(x) \ge u_\ell\}, \qquad \ell \in [k],$$

be the quantiles of this reconstructed group-mixture law. Finally, define

$$q_\ell^{\times *} \in \arg\min_{z\in\mathbb{R}} \sum_{s\in\mathcal{S}} \alpha_s |q_{s,\ell}^{\times} - z|^p.$$

The population finite-grid federated target is

$$G_{k,p} := \sum_{s\in\mathcal{S}} \alpha_s \frac{1}{k} \sum_{\ell=1}^{k} |q_{s,\ell}^{\times} - q_\ell^{\times *}|^p. \tag{41}$$

Similarly, define

$$\nu_k^{\mathrm{pool},\times} := \sum_{s\in\mathcal{S}} \alpha_s \nu_{s,k}^{\times}, \qquad F_k^{\mathrm{pool},\times} = \sum_{s\in\mathcal{S}} \alpha_s F_{s,k}^{\times},$$

and

$$H_{k,p}^{\times} := \sum_{s\in\mathcal{S}} \alpha_s C_p(\nu_{s,k}^{\times}, \nu_k^{\mathrm{pool},\times})^p. \tag{42}$$

The empirical estimator $\widehat{G}_{k,p}$ differs from $G_{k,p}$ only through empirical local quantiles and empirical weights. The difference between $G_{k,p}$ and G_p is the deterministic finite-grid approximation error.

D.3 Federated consistency and rate

Proof of Proposition 3.3. We prove the result for $\widehat{G}_{k,p}$. The proof for $\widehat{H}_{k,p}^{\times}$ is analogous, replacing quantile-space terms by cdf-space terms and using the step-function formula for the Cramér integral.

Write

$$|\widehat{G}_{k,p} - G_p| \le |\widehat{G}_{k,p} - G_{k,p}| + |G_{k,p} - G_p|. \tag{43}$$

Deterministic finite-grid term. The second term in (43) is deterministic. By Proposition 3.1, applied to the group-mixture laws

$$\{\nu_s^{\times} : s \in \mathcal{S}\},$$

we have

$$|G_{k,p} - G_p| \le \frac{C}{k}$$

under the stated regularity assumptions. The same argument gives

$$|H_{k,p}^{\times} - H_p^{\times}| \le \frac{C}{k}.$$

Fixed-k empirical consistency. For fixed k, the estimator depends on finitely many empirical quantities:

$$\widehat{\alpha}_s, \qquad \widehat{\pi}_{j|s}, \qquad \widehat{Q}_{j,s}(u_\ell), \qquad j \in [d],\ s \in \mathcal{S},\ \ell \in [k].$$

As $n_{\min} \to \infty$, the empirical weights satisfy

$$\widehat{\alpha}_s \xrightarrow{\mathbb{P}} \alpha_s, \qquad \widehat{\pi}_{j|s} \xrightarrow{\mathbb{P}} \pi_{j|s}$$

for every positive-probability group and cell. Under the local continuity assumptions,

$$\widehat{Q}_{j,s}(u_\ell) \xrightarrow{\mathbb{P}} Q_{j,s}(u_\ell)$$

for every fixed (j, s, ℓ) in a positive-probability cell.

The finite-grid reconstruction map is continuous at the population inputs under the stated quantile regularity conditions: small perturbations of the transmitted knots and weights induce small perturbations of the reconstructed mixture quantiles on the fixed grid. The finite-dimensional Fréchet objective is also continuous in these reconstructed quantiles and weights. Therefore, by the continuous mapping theorem,

$$\widehat{G}_{k,p} - G_{k,p} \xrightarrow{\mathbb{P}} 0$$

for fixed k. Combining this fixed-k consistency with the deterministic approximation term proves consistency whenever $k = k(n) \to \infty$, $n_{\min} \to \infty$, and both error terms vanish.

Quantitative rate. We now make the preceding argument quantitative. By Proposition 3.4, a union bound combined with the DKW–Massart inequality and trimmed quantile inversion gives, on the relevant trimmed grid,

$$\max_{j,s,\ell} \left| \widehat{Q}_{j,s}(u_\ell) - Q_{j,s}(u_\ell) \right| = O_{\mathbb{P}} \left(\sqrt{\frac{\log k}{n_{\min}}} \right), \tag{44}$$

where the constant depends on the lower density bound on the trimmed local regions. Empirical weight errors satisfy

$$\max_s |\widehat{\alpha}_s - \alpha_s| + \max_{j,s} |\widehat{\pi}_{j|s} - \pi_{j|s}| = O_{\mathbb{P}} \left(\sqrt{\frac{\log(d|\mathcal{S}|)}{n_{\min}}} \right),$$

under the same positive-cell regime. Since d and $|\mathcal{S}|$ are fixed in the proposition, this term is of the same or smaller order than (44).

The reconstructed group-mixture cdfs are weighted empirical distributions supported on the transmitted quantiles. Proposition D.1 implies that if the local knots are perturbed by η and the weights by γ, then

$$|\widehat{G}_{k,2} - G_{k,2}| \le C(\eta + \gamma + k^{-1})$$

for $p = 2$. The same Lipschitz argument applies for $p = 1$, using the stability of the pointwise weighted median in place of the linear barycenter formula. Taking

$$\eta = O_{\mathbb{P}} \left(\sqrt{\frac{\log k}{n_{\min}}} \right), \qquad \gamma = O_{\mathbb{P}} \left(\sqrt{\frac{\log(d|\mathcal{S}|)}{n_{\min}}} \right),$$

gives

$$|\widehat{G}_{k,p} - G_{k,p}| = O_{\mathbb{P}} \left(\sqrt{\frac{\log k}{n_{\min}}} + \frac{1}{k} \right).$$

Combining with the deterministic term yields

$$|\widehat{G}_{k,p} - G_p| = O_{\mathbb{P}} \left(\frac{1}{k} + \sqrt{\frac{\log k}{n_{\min}}} \right).$$

The proof for $\widehat{H}^{\times}_{k,p}$ follows the same structure. Both the empirical and population Cramér metrics are finite integrals between sorted jump locations on a common compact support. Uniform perturbations of the reconstructed cdfs and of the jump locations yield the same order of error, so

$$|\widehat{H}^{\times}_{k,p} - H^{\times}_p| = O_{\mathbb{P}} \left(\frac{1}{k} + \sqrt{\frac{\log k}{n_{\min}}} \right).$$

□

D.4 High-probability control of local empirical quantiles

Proof of Proposition 3.4. Fix a positive-probability cell (j, s). By the DKW–Massart inequality, for every $t > 0$,

$$\mathbb{P}\left(\|\widehat{F}_{j,s} - F_{j,s}\|_\infty > t \right) \le 2\exp(-2n_{j,s}t^2).$$

Since $n_{j,s} \ge n_{\min}$ for all positive-count cells, a union bound over all $j \in [d]$, $s \in \mathcal{S}$, and the k grid levels gives, with probability at least $1 - \delta$,

$$\max_{j,s} \|\widehat{F}_{j,s} - F_{j,s}\|_\infty \le \sqrt{\frac{1}{2n_{\min}} \log \left(\frac{2(k+1)d|\mathcal{S}|}{\delta} \right)} = \Delta_{n,k,\delta}.$$

The factor $k+1$ is harmless and upper-bounds the number of grid levels in the trimmed range.

On this event,

$$\|\widehat{F}_{j,s} - F_{j,s}\|_\infty \le \Delta_{n,k,\delta}$$

for every positive-probability cell (j,s). If $\Delta_{n,k,\delta} \le \varepsilon/2$, then for every

$$u_\ell \in [\varepsilon, 1-\varepsilon],$$

the points $u_\ell - \Delta_{n,k,\delta}$ and $u_\ell + \Delta_{n,k,\delta}$ lie in

$$[\varepsilon/2, 1-\varepsilon/2].$$

By Lemma B.8,

$$Q_{j,s}(u_\ell - \Delta_{n,k,\delta}) \le \widehat{Q}_{j,s}(u_\ell) \le Q_{j,s}(u_\ell + \Delta_{n,k,\delta}).$$

The lower density bound on

$$[Q_{j,s}(\varepsilon/2), Q_{j,s}(1-\varepsilon/2)]$$

implies that $Q_{j,s}$ is $1/m_\varepsilon$-Lipschitz on $[\varepsilon/2, 1-\varepsilon/2]$. Therefore

$$\left|\widehat{Q}_{j,s}(u_\ell) - Q_{j,s}(u_\ell)\right| \le \frac{\Delta_{n,k,\delta}}{m_\varepsilon}.$$

Taking the maximum over all positive-probability cells and all grid levels in the trimmed range gives the stated bound. □

D.5 High-probability stability of the reconstructed audit

The previous result controls the communicated local quantiles. We now record how such perturbations propagate to the reconstructed audit. This is the stability step used in Appendix D.3.

Proposition D.1 (Stability under perturbed summaries)**.** *Assume all score distributions are supported on a common compact interval and that the relevant group-mixture cdfs have densities bounded below on the trimmed region used for inversion. Suppose that*

$$\max_{j,s,\ell} |\widehat{q}_{j,s,\ell} - q_{j,s,\ell}| \le \eta, \qquad \max_s |\widehat{\alpha}_s - \alpha_s| + \max_{j,s} |\widehat{\pi}_{j|s} - \pi_{j|s}| \le \gamma.$$

Then, for $p=2$,

$$|\widehat{G}_{k,2} - G_{k,2}| \le C(\eta + \gamma + k^{-1}),$$

and

$$|\widehat{H}^{\times}_{k,2} - H^{\times}_{k,2}| \le C(\eta + \gamma + k^{-1}),$$

for a finite constant C depending only on the common support, density bounds, d, and $|\mathcal{S}|$. The same type of bound holds for

$$\widehat{V}_{\mathrm{mix}}, \qquad \widehat{V}_{\mathrm{bar}}, \qquad \widehat{R}.$$

Proof. A perturbation of the local quantile knots by at most η changes each local reconstructed step cdf only near its jump points. By Lemma B.5, the pointwise discrepancy between the original and perturbed local step cdfs is bounded by the fraction of knots lying in an interval of width 2η. By Lemma B.6, this fraction is bounded by $O(\eta + k^{-1})$ under an upper density bound. Perturbing the mixture weights adds $O(\gamma)$. Hence, uniformly in x,

$$\max_s \|\widehat{F}^{\times}_{s,k} - F^{\times}_{s,k}\|_\infty \le C(\eta + \gamma + k^{-1}).$$

On the trimmed region, the group-mixture quantile functions are Lipschitz. Therefore the reconstructed mixture quantiles satisfy

$$\max_{s,\ell} |\widehat{q}^{\times}_{s,\ell} - q^{\times}_{s,\ell}| \le C(\eta + \gamma + k^{-1}).$$

For $p = 2$, the finite-grid Wasserstein–Fréchet functional is Lipschitz in the quantile array and in the group weights on compact score ranges, by Lemma B.10. This gives the bound for $\widehat{G}_{k,2}$.

For the Cramér metric, both the population and empirical reconstructed cdfs are step functions supported on a common compact interval. The Cramér integral is a finite sum over intervals between sorted jump locations. Uniform cdf perturbations and knot-location perturbations together imply the same $O(\eta + \gamma + k^{-1})$ bound.

Finally, the diagnostic terms $\widehat{V}_{\text{mix}}$, $\widehat{V}_{\text{bar}}$, and $\widehat{R}$ are finite sums of products of bounded quantile differences and weights. The same Lipschitz argument gives the stated stability bound for each term. □

Combining Proposition D.1 with the high-probability quantile bound from Proposition 3.4 yields the rate stated in Proposition 3.3, as well as high-probability analogues for the diagnostic terms.

E Proofs and additional diagnostics for the local–global decomposition

This appendix proves the vanishing regimes stated in Section 4 and records a $p = 1$ diagnostic based on the triangle inequality. The exact ANOVA-style identity in Theorem 4.1 is specific to W_2^2, but the same objects remain informative for W_1.

E.1 Proof of the vanishing regimes

Proof of Proposition 4.2. We prove each item separately.

1. No cross-silo heterogeneity within sensitive groups. Assume that, for each sensitive group s, the local law $\nu_{j,s}$ does not depend on j whenever $\pi_{j|s} > 0$. Then there exists a law λ_s such that

$$\nu_{j,s} = \lambda_s \qquad \text{for all } j \text{ with } \pi_{j|s} > 0.$$

Therefore the actual group-mixture law is

$$\nu_s^{\times} = \sum_{j=1}^{d} \pi_{j|s} \nu_{j,s} = \sum_{j=1}^{d} \pi_{j|s} \lambda_s = \lambda_s.$$

The within-group Wasserstein barycenter of identical laws is also λ_s. Hence

$$\nu_s^* = \nu_s^{\times} \qquad \text{for every } s.$$

It follows that

$$W_2(\nu_s^{\times}, \nu_s^*)^2 = 0 \qquad \text{for every } s,$$

and therefore

$$V_{\text{mix}} = 0.$$

Since $Q_s^{\times} = Q_s^*$ for every s, the interaction term R also vanishes. The decomposition then gives

$$G_2 = V_{\text{bar}}.$$

The conditional independence $Z \perp\!\!\!\perp A \mid S$ implies this regime, up to the positive-cell qualification.

2. Within-silo score-level demographic parity. Assume that, for each silo j, the local law $\nu_{j,s}$ does not depend on s whenever $\mathbb{P}(A = j, S = s) > 0$. Then there exists a law λ_j such that

$$\nu_{j,s} = \lambda_j \qquad \text{for all } s \text{ with } \mathbb{P}(A = j, S = s) > 0.$$

The group-mixture law is therefore

$$\nu_s^{\times} = \sum_{j=1}^{d} \pi_{j|s} \lambda_j.$$

Thus any dependence of $\nu_s^\times$ on s comes only through the selection weights $\pi_{j|s}$.

If, in addition,

$$\pi_{j|s} = \beta_j \qquad \text{for all } j, s$$

on the relevant support, then

$$\nu_s^\times = \sum_{j=1}^{d} \beta_j \lambda_j$$

does not depend on s. Hence all group-mixture laws coincide, and the population disparity is zero:

$$G_2 = 0.$$

3. Both conditional independence conditions. Suppose that both $Z \perp\!\!\!\perp A \mid S$ and $Z \perp\!\!\!\perp S \mid A$ hold on the positive-probability cells. Let $\mathcal{G}$ be the bipartite graph with vertex set $\mathcal{S} \cup [d]$, and with an edge between $s \in \mathcal{S}$ and $j \in [d]$ whenever

$$\mathbb{P}(S = s, A = j) > 0.$$

The condition $Z \perp\!\!\!\perp A \mid S$ implies that, for each fixed s, all laws $\nu_{j,s}$ on edges incident to s are equal. The condition $Z \perp\!\!\!\perp S \mid A$ implies that, for each fixed j, all laws $\nu_{j,s}$ on edges incident to j are equal. Therefore the local score law is constant on each connected component of $\mathcal{G}$.

If all positive-probability sensitive groups belong to the same connected component, there is a common law λ such that

$$\nu_{j,s} = \lambda \qquad \text{for all positive-probability cells } (j, s).$$

Consequently,

$$\nu_s^\times = \lambda \qquad \text{for every } s \text{ with } \alpha_s > 0,$$

and hence

$$G_2 = V_{\text{mix}} = V_{\text{bar}} = R = 0.$$

This proves the stated positivity-qualified implication. □

E.2 A $p = 1$ triangle-inequality diagnostic

The exact identity in Theorem 4.1 relies on the Hilbertian geometry of W_2^2. For $p = 1$, there is no analogous squared-norm expansion. Nevertheless, the same objects yield useful upper and lower bounds.

Let ν_s^* be any within-group W_1 barycenter of the local laws $\{\nu_{j,s} : j \in [d]\}$ with weights $\pi_{j|s}$, and let $\nu^{\times *}$ be any W_1 barycenter of the group-mixture laws $\{\nu_s^\times : s \in \mathcal{S}\}$ with weights α_s. Define

$$V_{\text{mix}}^{(1)} := \sum_{s \in \mathcal{S}} \alpha_s W_1(\nu_s^\times, \nu_s^*),$$

and

$$V_{\text{bar}}^{(1)} := \sum_{s \in \mathcal{S}} \alpha_s W_1(\nu_s^*, \nu^{\times *}).$$

Proposition E.1 ($p = 1$ triangle-inequality diagnostic)**.** *For $p = 1$,*

$$\left| V_{\text{bar}}^{(1)} - V_{\text{mix}}^{(1)} \right| \le G_1 \le V_{\text{bar}}^{(1)} + V_{\text{mix}}^{(1)}.$$

Proof. For each sensitive group s, the triangle inequality gives

$$W_1(\nu_s^\times, \nu^{\times *}) \le W_1(\nu_s^\times, \nu_s^*) + W_1(\nu_s^*, \nu^{\times *}).$$

Multiplying by α_s and summing over s yields

$$G_1 \le V_{\text{mix}}^{(1)} + V_{\text{bar}}^{(1)}.$$

For the lower bound, the reverse triangle inequality gives

$$W_1(\nu_s^\times, \nu^{\times *}) \geq \left| W_1(\nu_s^*, \nu^{\times *}) - W_1(\nu_s^\times, \nu_s^*) \right|.$$

Thus

$$G_1 \geq \sum_{s \in \mathcal{S}} \alpha_s \left| W_1(\nu_s^*, \nu^{\times *}) - W_1(\nu_s^\times, \nu_s^*) \right|.$$

By Jensen's inequality,

$$\sum_{s \in \mathcal{S}} \alpha_s |a_s - b_s| \geq \left| \sum_{s \in \mathcal{S}} \alpha_s a_s - \sum_{s \in \mathcal{S}} \alpha_s b_s \right|.$$

Taking

$$a_s = W_1(\nu_s^*, \nu^{\times *}), \qquad b_s = W_1(\nu_s^\times, \nu_s^*),$$

gives

$$G_1 \geq \left| V_{\text{bar}}^{(1)} - V_{\text{mix}}^{(1)} \right|.$$

□

Computation from sketches. In one dimension,

$$W_1(\mu, \nu) = \int_0^1 |Q_\mu(u) - Q_\nu(u)| \, du.$$

A W_1 barycenter can be computed pointwise in u as a weighted median of the input quantiles. Consequently, the same communicated quantiles used by Algorithm 1 yield plug-in estimates of $V_{\text{mix}}^{(1)}$ and $V_{\text{bar}}^{(1)}$: the server replaces the integrals by midpoint sums over the grid $\{u_\ell : \ell \in [k]\}$, and replaces each W_1 barycenter quantile by a weighted median of the corresponding local or group-mixture quantiles.

Interpretation. Unlike the $p = 2$ case, the $p = 1$ bounds are not an exact decomposition. They nevertheless give a useful diagnostic range. If $V_{\text{mix}}^{(1)}$ is large, the global W_1 disparity may be strongly affected by mixture and composition effects. If $V_{\text{bar}}^{(1)}$ is large, disparity persists after replacing each within-group mixture by its W_1 transport center. The two quantities bound the possible size of G_1 through Proposition E.1.

F One-shot federated audit: full algorithm

This appendix gives an implementation-level description of the one-shot audit protocol introduced in Section 3. The protocol uses the common midpoint quantile grid

$$u_\ell = \frac{\ell - \frac{1}{2}}{k}, \qquad \ell \in [k],$$

and requires one communication round. Each silo sends subgroup counts and local empirical score quantiles. The server then reconstructs the group-mixture score distributions, estimates the population-level disparity, and, for $p = 2$, computes the diagnostic decomposition.

Implementation detail. The inversion step in Algorithm 1 does not require numerical root finding. For each sensitive group s, the reconstructed mixture cdf $\widehat{F}_{s,k}^\times$ is a weighted empirical distribution supported on the multiset

$$\{\widehat{q}_{j,s,\ell} : j \in [d], \ell \in [k]\}.$$

Each support point $\widehat{q}_{j,s,\ell}$ has mass $\widehat{\pi}_{j|s}/k$. The mixture quantiles $\widehat{q}_{s,\ell}^\times$ are obtained by sorting these dk support points and scanning cumulative masses. The computational cost per sensitive group is therefore $O(dk \log(dk))$, while the communication cost per silo is $|\mathcal{S}|(k+1)$ scalars.

Empty or small cells. The algorithm is written for positive subgroup counts. In practice, cells with $n_{j,s} = 0$ contribute zero weight to the mixture for group s, and cells with very small $n_{j,s}$ should not release raw quantile summaries without additional safeguards. Such cells can be suppressed, coarsened, combined with larger cells, or handled through secure aggregation or noise mechanisms. These disclosure controls are not part of the estimator itself, but are necessary in privacy-sensitive deployments. The stability results in Appendix D.5 describe how perturbations of counts and quantiles propagate to the final estimates.

Algorithm 1 One-shot federated audit from k quantiles

Require: Number of quantiles k, midpoint grid $u_\ell = (\ell - \frac{1}{2})/k$, $\ell \in [k]$.
Require: For each silo $j \in [d]$ and sensitive group $s \in \mathcal{S}$: subgroup count $n_{j,s}$ and local empirical quantiles

$$\widehat{q}_{j,s,\ell} := \widehat{Q}_{j,s}(u_\ell), \qquad \ell \in [k],$$

computed from the local scores $\{Z_i : A_i = j, S_i = s\}$.
Ensure: Global disparity estimate $\widehat{G}_{k,p}$, Cramér companion estimate $\widehat{H}^{\times}_{k,p}$, and, for $p = 2$, diagnostic terms $(\widehat{V}_{\mathrm{mix}}, \widehat{V}_{\mathrm{bar}}, \widehat{R})$.
Compute subgroup totals

$$n_s = \sum_{j=1}^{d} n_{j,s}, \qquad n = \sum_{s \in \mathcal{S}} n_s.$$

Compute empirical weights

$$\widehat{\alpha}_s = \frac{n_s}{n}, \qquad \widehat{\pi}_{j|s} = \frac{n_{j,s}}{n_s},$$

for all groups s such that $n_s > 0$.
for each $j \in [d]$ and $s \in \mathcal{S}$ with $n_{j,s} > 0$ **do**
Construct the local step cdf induced by the transmitted quantiles:

$$\widehat{F}_{j,s,k}(x) = \frac{1}{k} \sum_{\ell=1}^{k} \mathbf{1}\{\widehat{q}_{j,s,\ell} \le x\}.$$

Let $\widehat{\nu}_{j,s,k}$ denote the corresponding atomic law.
end for
for each $s \in \mathcal{S}$ **do**
Construct the within-group mixture cdf across silos:

$$\widehat{F}^{\times}_{s,k}(x) = \sum_{j=1}^{d} \widehat{\pi}_{j|s} \widehat{F}_{j,s,k}(x).$$

Let $\widehat{\nu}^{\times}_{s,k}$ denote the law with cdf $\widehat{F}^{\times}_{s,k}$.
Invert the mixture cdf on the common grid:

$$\widehat{q}^{\times}_{s,\ell} = \inf\{x : \widehat{F}^{\times}_{s,k}(x) \ge u_\ell\}, \qquad \ell \in [k].$$

end for
for each $\ell \in [k]$ **do**
Compute the barycenter quantile across sensitive groups:

$$\widehat{q}^{\times *}_{\ell} \in \arg\min_{z \in \mathbb{R}} \sum_{s \in \mathcal{S}} \widehat{\alpha}_s |\widehat{q}^{\times}_{s,\ell} - z|^p.$$

if $p = 2$ **then**
Use the closed form

$$\widehat{q}^{\times *}_{\ell} = \sum_{s \in \mathcal{S}} \widehat{\alpha}_s \widehat{q}^{\times}_{s,\ell}.$$

end if
end for
Output

$$\widehat{G}_{k,p} = \sum_{s \in \mathcal{S}} \widehat{\alpha}_s \frac{1}{k} \sum_{\ell=1}^{k} \left| \widehat{q}^{\times}_{s,\ell} - \widehat{q}^{\times *}_{\ell} \right|^p.$$

Algorithm 2 Cramér companion metric and W_2^2 diagnostics

Require: Quantities computed by Algorithm 1.

Construct the pooled reconstructed law across sensitive groups:

$$\widehat{\nu}_k^{\mathrm{pool},\times} = \sum_{s\in\mathcal{S}} \widehat{\alpha}_s \widehat{\nu}_{s,k}^{\times}.$$

Equivalently, its cdf is

$$\widehat{F}_k^{\mathrm{pool},\times}(x) = \sum_{s\in\mathcal{S}} \widehat{\alpha}_s \widehat{F}_{s,k}^{\times}(x).$$

for each $s \in \mathcal{S}$ **do**

Let $g_1 < \cdots < g_M$ be the distinct jump locations of $\widehat{F}_{s,k}^{\times}$ and $\widehat{F}_k^{\mathrm{pool},\times}$.

Compute the Cramér distance exactly from the step functions:

$$C_p(\widehat{\nu}_{s,k}^{\times}, \widehat{\nu}_k^{\mathrm{pool},\times})^p = \sum_{m=1}^{M-1} (g_{m+1} - g_m) \left| \widehat{F}_{s,k}^{\times}(g_m) - \widehat{F}_k^{\mathrm{pool},\times}(g_m) \right|^p.$$

end for

Output

$$\widehat{H}_{k,p}^{\times} = \sum_{s\in\mathcal{S}} \widehat{\alpha}_s C_p(\widehat{\nu}_{s,k}^{\times}, \widehat{\nu}_k^{\mathrm{pool},\times})^p.$$

if $p = 2$ **then**

for each $s \in \mathcal{S}$ and $\ell \in [k]$ **do**

Compute the within-group transport-barycenter quantile across silos:

$$\widehat{q}_{s,\ell}^{*} = \sum_{j=1}^{d} \widehat{\pi}_{j|s} \widehat{q}_{j,s,\ell}.$$

end for

Compute the mixture/composition term:

$$\widehat{V}_{\mathrm{mix}} = \sum_{s\in\mathcal{S}} \widehat{\alpha}_s \frac{1}{k} \sum_{\ell=1}^{k} \left(\widehat{q}_{s,\ell}^{\times} - \widehat{q}_{s,\ell}^{*} \right)^2.$$

Compute the barycentric cross-silo term:

$$\widehat{V}_{\mathrm{bar}} = \sum_{s\in\mathcal{S}} \widehat{\alpha}_s \frac{1}{k} \sum_{\ell=1}^{k} \left(\widehat{q}_{s,\ell}^{*} - \widehat{q}_{\ell}^{\times *} \right)^2.$$

Compute the interaction term:

$$\widehat{R} = 2 \sum_{s\in\mathcal{S}} \widehat{\alpha}_s \frac{1}{k} \sum_{\ell=1}^{k} \left(\widehat{q}_{s,\ell}^{\times} - \widehat{q}_{s,\ell}^{*} \right) \left(\widehat{q}_{s,\ell}^{*} - \widehat{q}_{\ell}^{\times *} \right).$$

Return the finite-grid identity

$$\widehat{G}_{k,2} = \widehat{V}_{\mathrm{mix}} + \widehat{V}_{\mathrm{bar}} + \widehat{R},$$

up to numerical rounding.

end if

G Additional experiments on ACSIncome

This appendix provides additional details for the ACSIncome experiment reported in Section 5.4. The purpose of this experiment is to evaluate the audit protocol in a setting where the silos are natural rather than simulated.

G.1 Dataset and federated structure

We use the 2018 ACSIncome benchmark derived from the American Community Survey Public Use Microdata Sample (Ding et al., 2021). The prediction target is

$$Y = \mathbf{1}\{\texttt{PINCP} > 50{,}000\},$$

where PINCP denotes annual personal income. The available covariates include

$$\texttt{AGEP},\ \texttt{COW},\ \texttt{SCHL},\ \texttt{MAR},\ \texttt{OCCP},\ \texttt{POBP},\ \texttt{RELP},\ \texttt{WKHP},\ \texttt{SEX},\ \texttt{RAC1P}.$$

We treat the 50 U.S. states and Puerto Rico as clients, yielding

$$d = 51$$

natural silos. This differs from the COMPAS, Adult, and Law experiments, where silo assignments are simulated.

We consider two binary sensitive-group comparisons:

$$\texttt{SEX} : \text{Male vs. Female},$$

and

$$\texttt{RAC1P} : \text{White alone vs. Black or African American alone.}$$

For the RAC1P audit, we restrict the sample to individuals in these two race categories.

G.2 Score construction

For each audit, we train a ridge logistic regression model to predict Y. The score is the predicted probability

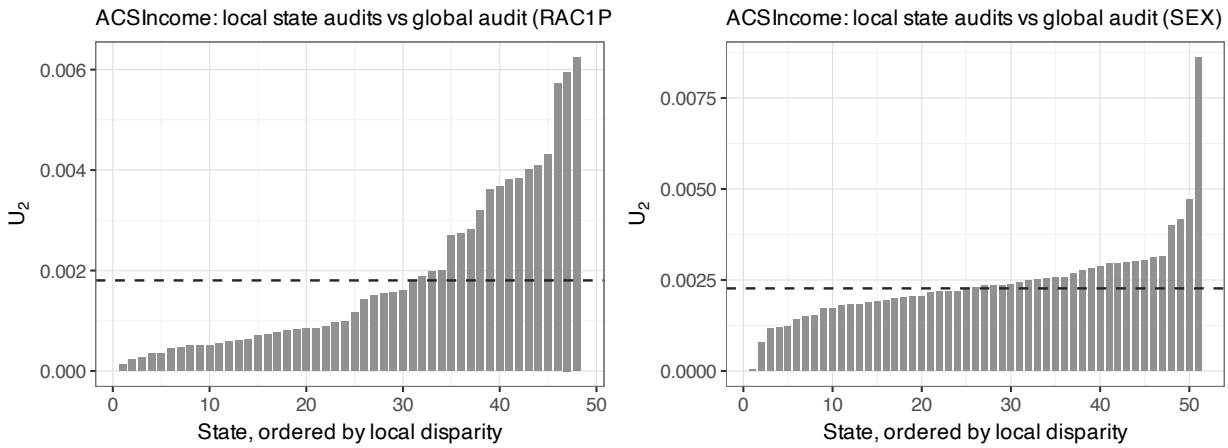
$$Z = \widehat{\mathbb{P}}(Y = 1 \mid X).$$

The audited sensitive attribute is excluded from the score model. Thus, for the SEX audit, SEX is excluded from the covariates; for the RAC1P audit, RAC1P is excluded. The variables AGEP and WKHP are treated as numeric, while the remaining ACS variables are encoded as categorical variables. For computational efficiency, the model is trained on a random subsample and then used to predict scores for the full audit sample.

G.3 State-level local audits

The main text reports a representative subset of state-level local audits in Table 4. Here we visualize all state-level local audits and compare them with the population-level benchmark. Each bar corresponds to the value of U_2 computed locally within one state, while the dashed horizontal line denotes the centralized population benchmark computed from pooled scores.

The SEX audit shows relatively stable disparities across large states, but local values still vary around the pooled benchmark. The RAC1P audit shows stronger local variation: some states have local U_2 values below the global benchmark, while others have larger local values. This is precisely the regime targeted by the federated audit: the population-level score-distribution gap must be reconstructed from group-specific mixtures across clients, rather than inferred from any single state-level audit.



(a) `RAC1P`: White vs. Black.

(b) `SEX`: Male vs. Female.

Figure 5: **`ACSIncome`: local state audits versus the population audit.** Each bar reports the local state-level U_2 value. The dashed line is the centralized population-level benchmark. Local state audits may lie below or above the pooled benchmark, confirming that state-level audits are not interchangeable with the population audit.

G.4 Communication cost

For `ACSIncome`, each audit has $d = 51$ clients and two sensitive groups. Hence the total number of transmitted scalars is

$$2d(k+1) = 102(k+1).$$

For example, with $k = 50$, the server receives

$$102 \times 51 = 5{,}202$$

scalars in total: 50 quantiles and one count per group and state. The server receives no individual-level records and no individual-level scores.

G.5 Interpretation

The `ACSIncome` experiment complements the simulated-silo experiments in two ways. First, the silos are natural geographic units rather than artificial partitions. Second, group proportions and score distributions vary across states, so the group-specific mixing weights $\pi_{j|s}$ are empirically meaningful. The experiment therefore illustrates the main methodological point of the paper in a large-scale setting: population-level demographic disparity is a property of the pooled group score laws, while the available evidence is distributed across clients.

H Additional experiments on `COMPAS`

This appendix provides additional details for the `COMPAS` experiment reported in Section 5.3. Unlike the `ACSIncome` experiment, where silos correspond to natural geographic units, the `COMPAS` silos are simulated. The goal is therefore not to model an institutional federation, but to stress-test the local–global mismatch under controlled allocation mechanisms.

We verified that the reported conclusions are insensitive to the jitter scale: using smaller independent jitters within each decile bin changes the numerical values only marginally and does not alter the ordering between the centralized benchmark and the local silo audits.

H.1 Dataset and simulated silos

We restrict the COMPAS data to African-American and Caucasian individuals and use the reported decile score as the audit score. Since this score is discrete, we add a small independent jitter within each integer bin before computing empirical quantiles. This preserves the decile ordering while avoiding ties in the quantile computations. The filtered sample contains

$$n_{\mathrm{AA}} = 3696, \qquad n_{\mathrm{C}} = 2454.$$

We create simulated federations by assigning individuals to d silos. The allocation mechanism is controlled by a selection parameter ρ. When ρ is close to zero, silo assignment is approximately independent of the score; when $|\rho|$ is large, silo membership becomes more strongly associated with the score. This allows us to vary the strength of score–silo dependence while keeping the pooled group-wise score distributions fixed. Hence the centralized population target remains unchanged, but local audits and federated reconstruction difficulty vary with the allocation regime.

H.2 Selection effects and sketch accuracy

Figure 6 reports the mean relative error of the federated estimator for a fixed sketch size $k = 10$, as a function of the selection parameter ρ and the number of simulated silos d. The error is smallest when score–silo selection is weak and increases when the allocation mechanism induces stronger dependence between scores and silos. This confirms that the difficulty of the federated audit depends not only on the number of clients, but also on how informative silo membership is about the score distribution.

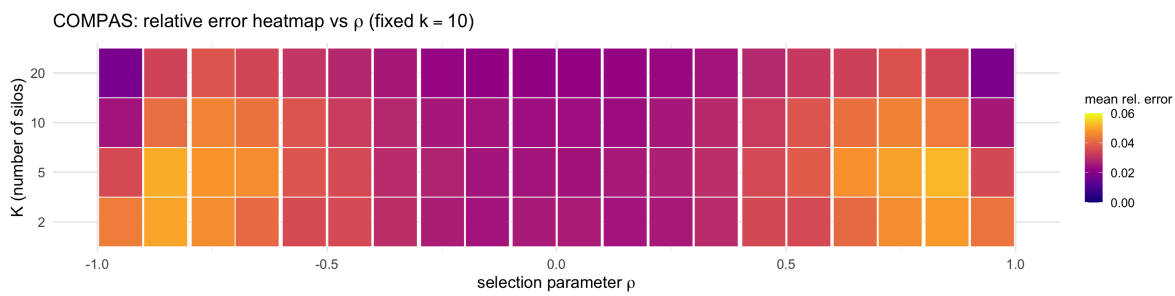


Figure 6: **COMPAS: relative error under score–silo selection.** Mean relative error $\mathbb{E}\left[|\widehat{G}_{k,2} - G_2|/G_2\right]$ for a fixed sketch size $k = 10$, as a function of the selection parameter ρ and the number of simulated silos d.

H.3 Minimal sketch size for reliable recovery

To summarize the communication–accuracy trade-off, we also report the smallest sketch size required to recover the centralized benchmark with high probability. For each k, define

$$p_{\mathrm{ok}}(k) = \mathbb{P}\left(\frac{|\widehat{G}_{k,2} - G_2|}{G_2} \leq 1\%\right),$$

where the probability is estimated over repeated simulated silo allocations. Then define

$$k_{95} = \min\{k : p_{\mathrm{ok}}(k) \geq 0.95\}.$$

Figure 7 reports k_{95} by allocation regime and number of silos. Selection-biased allocations require larger sketches than random allocation. In the random-allocation regime, relatively small sketches are sufficient. Under selection bias, the required k increases, reflecting the additional difficulty created by informative silo membership.

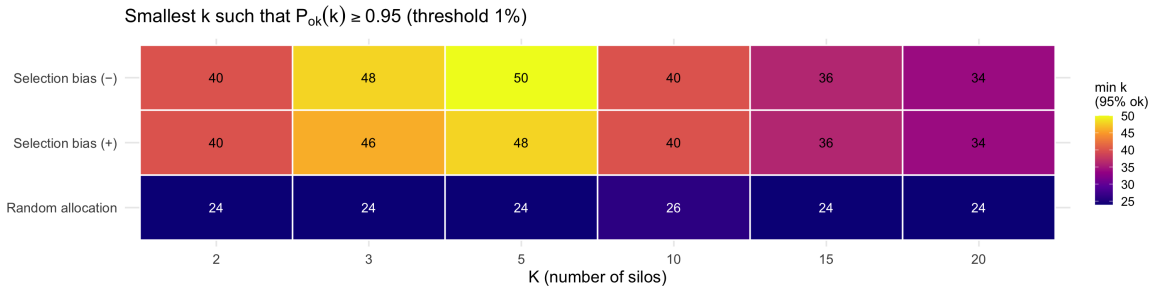


Figure 7: **COMPAS: minimal sketch size for reliable recovery.** Smallest k such that $p_{\mathrm{ok}}(k) = \mathbb{P}(|\widehat{G}_{k,2} - G_2|/G_2 \leq 1\%) \geq 0.95$, by allocation regime and number of simulated silos d.

H.4 Interpretation

The `COMPAS` experiment complements the synthetic experiment. The score distribution is empirical rather than generated from a parametric model, but the silo structure remains controlled. The results support the central claim of the paper: local audits can fail as estimators of the population audit when silo assignment is informative. They also show that the sketch size required for reliable recovery is not a fixed property of the dataset alone. It depends on the number of silos, the local subgroup sizes, and the strength of score–silo selection.

I Additional experiments on `Adult` and `Law`

This appendix reports additional robustness experiments on two standard fairness benchmarks, `Adult` and `Law`. Unlike `ACSIncome`, these datasets do not provide natural client partitions comparable to state-level silos. We therefore use simulated federations. The purpose is not to model real institutions, but to evaluate how the quantile-sketch audit behaves under different federation sizes and different strengths of score–silo dependence.

I.1 Experimental setup

For `Adult`, the prediction task is income classification,

$$Y = \mathbf{1}\{\texttt{income} > 50\mathrm{K}\},$$

and the score is the predicted probability from a logistic regression model trained without the audited sensitive attribute. We focus on a binary race comparison, restricting to individuals labeled `Black` or `White`. For `Law`, the outcome is bar passage, and the score is again the predicted probability from a logistic regression model trained on non-sensitive covariates. We restrict to individuals labeled `black` or `white`.

For both datasets, individuals are assigned to d simulated silos. We consider random allocation and two selection-biased allocation regimes. In the random regime, silo assignment is approximately independent of the score. In the selection-biased regimes, silo assignment is made informative about the score while preserving the pooled group-wise score distributions. Thus the centralized benchmark $G_2 = U_2$ remains fixed, while local audits and the difficulty of federated reconstruction vary across simulated federations.

For each simulated federation, each silo releases only subgroup counts and k score quantiles per sensitive group. The server reconstructs the group-mixture score distributions and computes $\widehat{G}_{k,2}$. We repeat the experiment over multiple simulated silo allocations and compare $\widehat{G}_{k,2}$ with the centralized benchmark.

I.2 Reliability metric

We summarize the accuracy of the federated sketch through the reliability profile

$$p_{\mathrm{ok}}(k) = \mathbb{P}\left(\frac{|\widehat{G}_{k,2} - G_2|}{G_2} \leq 1\%\right),$$

where the probability is estimated over repeated simulated silo allocations. We then define

$$k_{95} = \min\{k : p_{\mathrm{ok}}(k) \geq 0.95\}.$$

The value k_{95} is a communication-oriented summary: it is the smallest number of quantiles per group and silo needed to recover the centralized benchmark within a 1% relative-error tolerance with probability at least 95%.

I.3 `Adult`

Figure 8 reports the mean relative error for `Adult` at a fixed sketch size, as a function of the selection parameter ρ and the number of simulated silos d. The pattern shows that informative silo assignment can substantially increase the error of a fixed-resolution sketch. This is expected: when silo membership becomes more informative about the score, the group-mixture reconstruction becomes more sensitive to local subgroup quantiles and group-specific client weights.

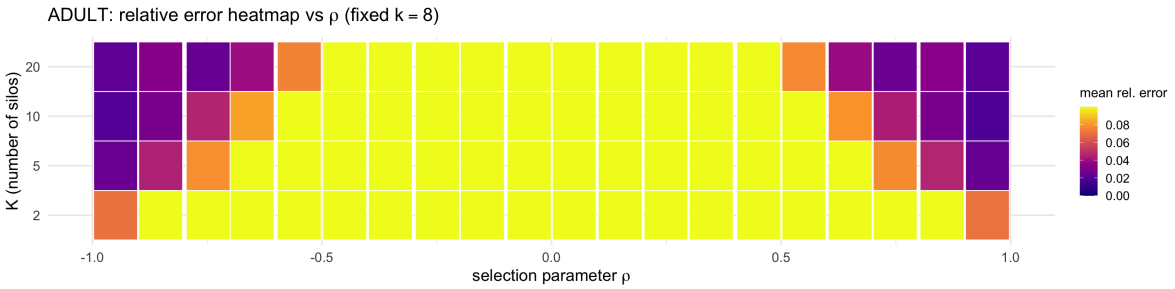


Figure 8: **`Adult`: relative error under score–silo selection.** Mean relative error $\mathbb{E}\left[|\widehat{G}_{k,2} - G_2|/G_2\right]$ for a fixed sketch size, as a function of the selection parameter ρ and the number of simulated silos d.

Figure 9 reports the corresponding k_{95} values. Random allocation requires much smaller sketches than selection-biased allocation. Under selection bias, larger values of k are needed to reach the same reliability level. The figure also shows that the required sketch size depends on the number of silos: increasing d changes the local subgroup sample sizes and the stability of empirical quantiles.

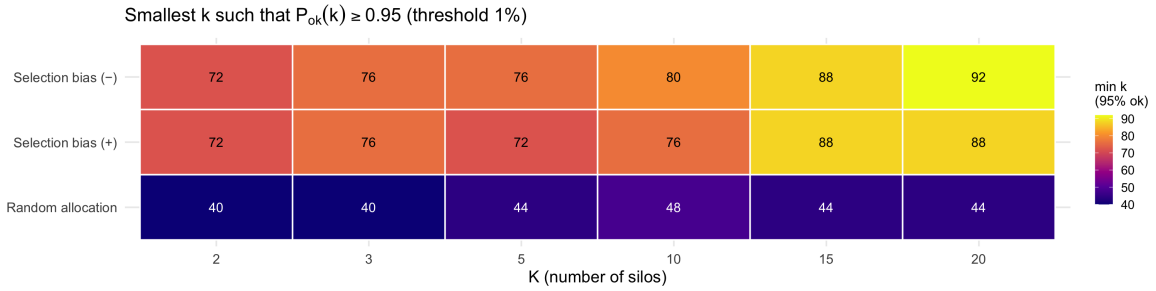


Figure 9: **`Adult`: minimal sketch size for reliable recovery.** Smallest k such that $p_{\mathrm{ok}}(k) = \mathbb{P}(|\widehat{G}_{k,2} - G_2|/G_2 \leq 1\%) \geq 0.95$, by allocation regime and number of simulated silos d.

I.4 Law

Figure 10 reports the same k_{95} summary for Law. The required sketch sizes are generally larger than in the Adult and COMPAS experiments. This indicates that the Law simulated federations are more demanding for the quantile-sketch audit, either because of smaller effective subgroup sizes, stronger sensitivity of the score distribution to silo assignment, or greater finite-sample variability of local quantiles.

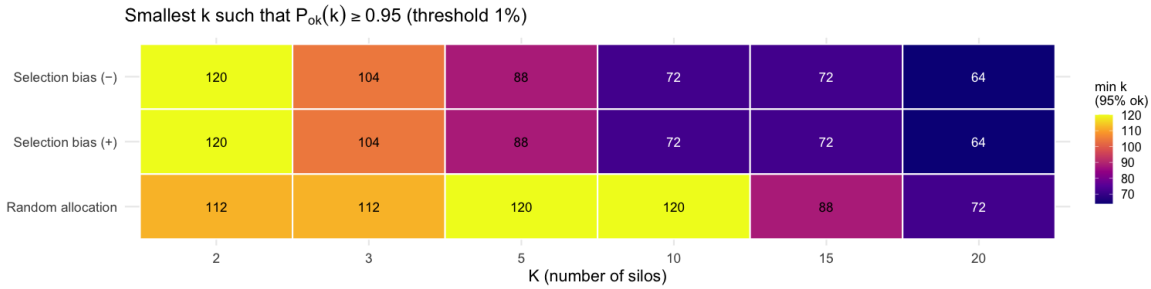


Figure 10: **Law: minimal sketch size for reliable recovery.** Smallest k such that $p_{\mathrm{ok}}(k) = \mathbb{P}(|\widehat{G}_{k,2} - G_2|/G_2 \leq 1\%) \geq 0.95$, by allocation regime and number of simulated silos d.

I.5 Summary

The Adult and Law experiments reinforce the conclusions of the main text. First, local audits are not reliable substitutes for the population-level audit when silo assignment is informative. Second, the quantile-sketch estimator can recover the centralized target with a moderate number of communicated quantiles, but the required sketch size is dataset- and federation-dependent. Third, reliability depends jointly on the number of silos, local subgroup sample sizes, and the strength of score–silo dependence. These robustness experiments therefore support the main methodological recommendation: federated fairness audits should reconstruct the population-level group score laws rather than average local fairness metrics.

J Reproducibility details

This appendix summarizes implementation details common to all experiments. Dataset-specific choices are reported in the corresponding appendices.

J.1 Quantile grid and reference values

All federated audits use the midpoint quantile grid

$$u_\ell = \frac{\ell - \frac{1}{2}}{k}, \qquad \ell \in [k].$$

Unless otherwise stated, we evaluate

$$k \in \{5, 10, 20, 30, 50, 75, 100, 150, 200\}.$$

For each experiment, the centralized benchmark is computed from pooled individual-level scores using a fine grid with

$$k_{\mathrm{fine}} = 2000.$$

This centralized benchmark is used only for evaluation. The federated estimator uses only client-level counts and quantile summaries.

For two sensitive groups, the main reported metrics are

$$|\Delta \bar{z}| = \left| \frac{1}{n_0} \sum_{i:S_i=0} Z_i - \frac{1}{n_1} \sum_{i:S_i=1} Z_i \right|,$$

$$U_2 = \alpha_0\alpha_1 W_2(\nu_0, \nu_1)^2, \qquad H_2 = \alpha_0\alpha_1 C_2(\nu_0, \nu_1)^2,$$

and

$$W_2 = W_2(\nu_0, \nu_1).$$

The quantity U_2 is the main Wasserstein–Fréchet audit target. The quantity H_2 is the Cramér companion metric. The unweighted W_2 distance is reported as an interpretable scale measure between the two group score distributions.

J.2 Federated summaries

For each silo j and sensitive group s, the communicated summary is

$$(n_{j,s}, \widehat{q}_{j,s,1}, \ldots, \widehat{q}_{j,s,k}),$$

where

$$\widehat{q}_{j,s,\ell} = \widehat{Q}_{j,s}(u_\ell)$$

is the empirical quantile of the local scores

$$\{Z_i : A_i = j, S_i = s\}.$$

Thus, for $|\mathcal{S}|$ sensitive groups, the communication cost per silo is

$$|\mathcal{S}|(k+1)$$

scalars, and the total communication cost over d silos is

$$d|\mathcal{S}|(k+1).$$

In all experiments reported in the main text, $|\mathcal{S}| = 2$, so the total cost is

$$2d(k+1).$$

The server reconstructs the group-mixture cdfs as described in Algorithm 1. For each group s, the reconstructed mixture cdf is a weighted empirical distribution supported on the transmitted quantiles

$$\{\widehat{q}_{j,s,\ell} : j \in [d], \ell \in [k]\},$$

with mass $\widehat{\pi}_{j|s}/k$ assigned to each support point from silo j. The mixture quantiles $\widehat{q}^{\times}_{s,\ell}$ are obtained by sorting these support points and scanning the cumulative weights.

J.3 Score models

All audits are performed on one-dimensional scores Z.

COMPAS. The score is the COMPAS decile score, restricted to African-American and Caucasian individuals. Since the raw decile score is discrete, we add a small independent jitter within each integer bin before computing empirical quantiles. This avoids ties while preserving the decile ordering.

ACSIncome. The target is

$$Y = \mathbf{1}\{\texttt{PINCP} > 50{,}000\}.$$

We train a ridge logistic regression model and use the predicted probability as the score. For the SEX audit, the variable SEX is excluded from the score model. For the RAC1P audit, the variable RAC1P is excluded. The variables AGEP and WKHP are treated as numeric; the remaining ACS variables are treated as categorical.

Adult. The target is income above 50K. The score is the predicted probability from a logistic regression model trained without the audited sensitive attribute. In the default specification, sex is also excluded from the score model.

Law. The target is bar passage. The score is the predicted probability from a logistic regression model trained on non-sensitive covariates, excluding the audited sensitive attribute.

J.4 Silo construction

The experiments use two kinds of federated structures.

Natural silos. For `ACSIncome`, silos are not simulated. Each U.S. state and Puerto Rico is treated as one client, yielding

$$d = 51.$$

Simulated silos. For `COMPAS`, `Adult`, and `Law`, silo assignments are simulated. We use random assignment as a baseline and selection-biased assignments to stress-test local–global mismatch. The selection-biased mechanisms induce dependence between score values and silo membership while preserving the pooled group-wise score distributions. Hence the centralized target remains fixed, while local audits and federated reconstruction difficulty vary with the allocation regime.

J.5 Monte Carlo repetitions and reliability profiles

For simulated-silo experiments, reported averages, boxplots, and reliability profiles are computed over repeated random silo allocations. The centralized benchmark $G_2 = U_2$ is computed once from the pooled scores, while $\widehat{G}_{k,2}$ is recomputed for each simulated federation and each value of k.

For a tolerance level τ, the reliability profile is

$$p_{\mathrm{ok}}(k) = \mathbb{P}\left(\frac{|\widehat{G}_{k,2} - G_2|}{G_2} \leq \tau\right),$$

where the probability is estimated over Monte Carlo silo allocations. In the reported reliability figures, we use

$$\tau = 1\%.$$

We also report

$$k_{95} = \min\{k : p_{\mathrm{ok}}(k) \geq 0.95\},$$

which summarizes the number of quantiles per group and silo needed to reach the target reliability level.

J.6 Software

The experiments are implemented in `R`. The `ACSIncome` data are accessed through the `fairlearn` Python package using `reticulate`. Logistic regression models use ridge regularization. Quantile summaries, Wasserstein metrics, Cramér metrics, and decomposition terms are computed from the same shared quantile grid.