# FedCF: Fair Federated Conformal Prediction

**Anutam Srinivasan [1,*†], Aditya Vadlamani [2,*], Amin Meghrazi [2], Srinivasan Parthasarathy [2]**
[1]Georgia Institute of Technology [2]The Ohio State University
asrinivasan350@gatech.edu, vadlamani.12@osu.edu,
meghrazi.1@osu.edu, srini@cse.ohio-state.edu

## Abstract

Conformal Prediction (CP) is a widely used technique for quantifying uncertainty in machine learning models. In its standard form, CP offers probabilistic guarantees on the coverage of the true label, but it is agnostic to sensitive attributes in the dataset. Several recent works have sought to incorporate fairness into CP by ensuring conditional coverage guarantees across different subgroups. One such method is Conformal Fairness (CF). In this work, we extend the CF framework to the Federated Learning setting and discuss how we can audit a federated model for fairness by analyzing the fairness-related gaps for different demographic groups. We empirically validate our framework by conducting experiments on several datasets spanning multiple domains, fully leveraging the exchangeability assumption.

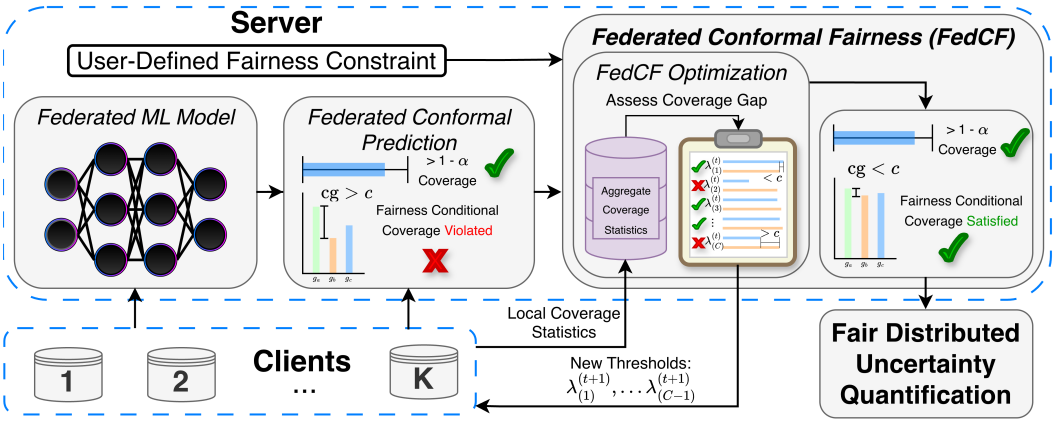



Figure 1: Overview of the FedCF framework. (1) Local Calibration: Decentralized clients compute non-conformity scores using private data. (2) Server Aggregation: Clients send only lightweight summary statistics (group-specific error rates) to the server. (3) Global Adjustment: The server updates and broadcasts the global quantiles to ensure equalized coverage across all demographic groups without ever pooling raw data.

## 1 Introduction

Ensuring model fairness is an important construct for trustworthy machine learning (ML). ML models, when not calibrated for fairness, are prone to developing biases at each stage of an ML pipeline, as reflected by their predictions [Mehrabi et al., 2021]. We define bias as disparate performance (e.g., classification accuracy) across different sub-populations. During data collection, measurement bias may arise from disproportionate sampling across subpopulations, whereas representation bias may stem from insufficient training data for specific strata. During training, these biases are inductively

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

learned by the model–leading to incorrect predictions in safety-critical tasks. These models are also susceptible to algorithmic bias from regularization and optimization techniques during model training, leading to incorrect generalization for marginalized groups. To mitigate these risks, many ML models must comply with regulations issued by local governing bodies [Hirsch et al., 2023]. Towards model compliance Komala et al. [2024], Agrawal et al. [2024], Jones et al. [2025] proposed approaches to enhance model fairness in tasks, including federated graph learning and representation learning.

Developing trustworthy ML frameworks with mathematically rigorous guarantees is an important requirement for regulated safety-critical tasks. In this context, researchers have explored Conformal Prediction (CP), an uncertainty quantification (UQ) technique that only assumes statistical exchangeability, to develop trustworthy models [Vovk et al., 2005]. Practitioners have adopted CP due to its model-free assumption and post-hoc application [Cherian and Bronner, 2020]. Additionally, users can apply CP to structured data (e.g., graphs), which cannot be used with traditional IID-based methods [Maneriker et al., 2025]. However, vanilla CP is not calibrated for fairness [Cresswell et al., 2025].

Several approaches have been proposed at the intersection of fairness and CP. Romano et al. [2020b] developed a CP approach for the regression setting to ensure equalized coverage across protected groups. Lu et al. [2022] considers equalized coverage in a classification task for medical imaging. Zhou and Sesia [2024] extend Romano et al. [2020b] and provide an algorithm adaptive to sensitive groups to increase the predictive power of the CP sets when several sensitive attributes are present, and focus on the classification task. Lastly, Vadlamani et al. [2025] provides a framework that ensures fair coverage of positive outcomes without requiring protected attributes at inference time, unlike prior work. Orthogonally, CP has been used to enhance the fairness of other tasks. To mitigate bias in LLM-based recommender systems, Fayyazi et al. [2025] explores iteratively using fairness-aware CP.

The above fair CP methods rely on centralized data, a challenge in domains such as healthcare and finance, where data is often decentralized and data privacy regulations apply. Federated learning offers a solution by enabling collaborative model training without transferring raw data off local clients. However, while recent work on federated CP exists [Lu et al., 2023, Humbert et al., 2023], the shift toward privacy-preserving decentralized training often exacerbates unfairness [Kanaparthy et al., 2022, Lo et al., 2025] in ML models. Simultaneously balancing fairness and federated requirements in CP is non-trivially challenging due to variability in client requirements and capabilities, cross-client distribution of sensitive attributes, and efficiency.

**Key Contributions**: We present Federated Conformal Fairness (FedCF), see Figure 1, which adapts Conformal Fairness (CF) [Vadlamani et al., 2025] to a federated setting, addressing the above non-trivial challenges while maintaining CF's theoretical guarantees in a distributed setting.

**Extending CF Theory to FL.** We discuss how to bound conditional coverage according to a user-specified fairness notion when data is decentralized. To facilitate this, we develop a novel *mixture-of-clients formalization*, in which we decompose the conditional coverage into a sufficient set of terms that each client can compute using local data, determine how the server should aggregate these terms to bound the conditional coverage, and theoretically prove the validity of our approach.

**Descent-Based CF Formulation.** We revise the original CF algorithm to reduce the number of communication rounds required. We do this by reformulating the original iterative approach presented in Vadlamani et al. [2025] into a descent-based optimization framework. This allows FedCF to be embedded directly into an FL algorithm and to integrate naturally within existing FL systems.

**Flexible Aggregation Protocols.** We consider the client-server communication overhead and its tradeoff with preserving data privacy. Specifically, we propose two approaches, with one having less communication overhead and the other being more privacy-preserving of client data.

**Real-World Empirical Validation.** We evaluate FedCF on several datasets with naturally induced ***data heterogeneity***, across different modalities and popular fairness metrics. We observe that FedCF can control for a particular coverage gap level while maintaining the original CP guarantee.

## 2 Background

### 2.1 Conformal Prediction

Conformal Prediction (CP) [Vovk et al., 2005] is a widely used framework for quantifying predictive uncertainty in ML. CP provides rigorous statistical guarantees without imposing assumptions on the

model, requiring only that the data is *exchangeable*–a broader condition than IID and compatible with non-IID or structured settings (e.g., graphs).

We focus on split (inductive) CP in the classification setting. Let $\boldsymbol{x}_i \in \mathcal{X} = \mathbb{R}^d$ and $y_i \in \mathcal{Y} = \{0, \ldots, C-1\}$ denote features and labels. Given a calibration dataset, $\mathcal{D}_{\text{calib}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, our goal is to construct a set-valued predictor $\mathcal{C}$ such that, for an exchangeable test point $(\boldsymbol{x}_{\text{test}}, y_{\text{test}})$,

$$1 - \alpha \leq \Pr[y_{\text{test}} \in \mathcal{C}(\boldsymbol{x}_{\text{test}})] \leq 1 - \alpha + \tfrac{1}{|\mathcal{D}_{\text{calib}}|+1}, \tag{1}$$

where $1 - \alpha \in (0, 1)$ is the target *coverage level*. We refer to Equation 1 as the *coverage guarantee*. Concretely, given a non-conformity score $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, define the *conformal quantile* as

$$\hat{q}(\alpha) = \text{Quantile}\left(\frac{\lceil (n+1)(1-\alpha) \rceil}{n}; \{s(\boldsymbol{x}_i, y_i)\}_{i=1}^n\right). \tag{2}$$

The resulting prediction set $\mathcal{C}_{\hat{q}(\alpha)}(\boldsymbol{x}_{\text{test}}) = \{y \in \mathcal{Y} : s(\boldsymbol{x}_{\text{test}}, y) \leq \hat{q}(\alpha)\}$ satisfies the guarantee in 1.

**Evaluating CP**: Two standard metrics are considered (averaged over the test set): (1) *Coverage*, the estimated test-time probability, $\Pr\left[y_{\text{test}} \in \mathcal{C}_{\hat{q}(\alpha)}(\boldsymbol{x}_{\text{test}})\right]$; and (2) *Efficiency*, the prediction set size, $\left|\mathcal{C}_{\hat{q}(\alpha)}(\boldsymbol{x}_{\text{test}})\right|$. These are typically in tension as a higher desired coverage leads to larger sets.

### 2.2 Federated Learning

A key contributor to developing strong deep learning models is providing a large amount of quality training data [Kaplan et al., 2020]. However, in domains including healthcare and finance, collecting large amounts of data may be prohibitive due to privacy concerns. Federated learning (FL) McMahan et al. [2017] is a framework for collaborative learning that keeps training data decentralized and private. Given $K$ clients that participate in training, each client has its own training data that it wants to keep private. The goal is to optimize a global loss function $L$ that is the weighted average of local risk functions $\ell_k$. Formally, FL finds weights $\theta^*$ s.t. $\theta^* = \underset{\theta}{\text{argmin}}\left\{L(\theta) = \sum_{k=1}^K w_k \mathbb{E}_{(x^{(k)}, y^{(k)}) \sim P_k}\left[\ell_k\left(\theta; x^{(k)}, y^{(k)}\right)\right]\right\}$, where $P_k$ is client $k$'s local distribution and $\boldsymbol{w} \in \Delta^K$ are weights [Lu et al., 2023].

### 2.3 Federated Conformal Prediction (FCP)

**Setting:** In FL, the development and calibration datasets are partitioned over $K$ clients. Meaning, each client $k \in \{1, \ldots, K\} = \mathcal{K}$ retains a private calibration set $\mathcal{D}_{\text{calib}}^{(k)} = \left\{\left(\boldsymbol{x}_i^{(k)}, y_i^{(k)}\right)\right\}_{i=1}^{n_k}$ drawn from an unknown local distribution $P_k$. The goal is to still construct a prediction set function $\mathcal{C}$ such that for any test point $(\boldsymbol{x}_{\text{test}}, y_{\text{test}}) \sim Q_{\text{test}}$, where $Q_{\text{test}} = \sum_{k=1}^K \gamma_k P_k$ is the mixture distribution with weights $\gamma_k \propto (n_k + 1)$ [Lu et al., 2023], Equation 1 is satisfied. This is done while respecting the communication and privacy constraints of FL.

**Partial exchangeability and the FCP algorithm.** FCP [Lu et al., 2023] relies on the notion of partial exchangeability [de Finetti, 1971], by requiring that within each client, with probability $\gamma_k$, $\{s(\boldsymbol{x}_1^{(k)}, y_1^{(k)}), \ldots, s(\boldsymbol{x}_{n_k}^{(k)}, y_{n_k}^{(k)}), s(\boldsymbol{x}_{\text{test}}, y_{\text{test}})\}$ is exchangeable. This assumption helps account for real-world label skew and data heterogeneity amongst clients. Under this assumption, FCP aggregates all non-conformity scores and selects the $((1-\alpha)(N+K))$-th statistic as follows

$$\hat{q}(\alpha) = \text{Quantile}\left(\frac{\lceil (N+K)(1-\alpha) \rceil}{N}; \left\{\left(\boldsymbol{x}_i^{(k)}, y_i^{(k)}\right)\right\}_{k,i}\right) \tag{3}$$

where $N = \sum_{k=1}^K n_k$. The prediction set $C_\alpha(\boldsymbol{x}) = \{y : s(\boldsymbol{x}, y) \leq \hat{q}_\alpha\}$ then satisfies,

$$1 - \alpha \;\leq\; \Pr[y_{\text{test}} \in C_\alpha(\boldsymbol{x}_{\text{test}})] \;\leq\; 1 - \alpha + \tfrac{K}{N+K}. \tag{4}$$

**Communication-efficient quantile sketches.** To preserve privacy, instead of transmitting all $N$ scores to the server, each client can send a mergeable sketch (e.g., T-Digest [Dunning, 2021], DDSketch [Masson et al., 2019]). Doing so loosens the guarantee given in Eq. (4) [Lu et al., 2023].

### 2.4 Conformal Fairness

The Conformal Fairness (CF) framework [Vadlamani et al., 2025] formalizes the notion of fair prediction sets by considering the disparity in conditional coverage between sensitive groups and

improves upon the marginal coverage guarantees of traditional CP, where different groups face disparate coverage. Formally, let $M$ denote a fairness metric (e.g. Equal Opportunity) and defines the filter function $F_M : \mathcal{X} \times \mathcal{Y} \times \mathcal{G} \times \mathcal{Y}^+ \to \{0, 1\}$, where $\mathcal{G}$ is a set of sensitive attributes, and $\mathcal{Y}^+ \subseteq \mathcal{Y}$ is the set of advantaged outcomes, to assess conditional coverages. For example, $F_M$ for Equal Opportunity is defined as $F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, y) := \mathbf{1}[\boldsymbol{x}_{\text{test}} \in g \wedge y_{\text{test}} = y]$. Then, CF finds class-wise thresholds, $\lambda_{\text{opt}} \in \mathbb{R}^{|\mathcal{Y}|}$ and defines $\mathcal{C}_{\lambda_{\text{opt}}}(\boldsymbol{x}_{\text{test}}) = \{y \in \mathcal{Y} : s(\boldsymbol{x}_{\text{test}}, y) \leq \lambda_{\text{opt};(y)}\}$ where $\lambda_{\text{opt};(y)}$ is the threshold for class $y$, such that, $\forall\, g_a, g_b \in \mathcal{G},\ \tilde{y} \in \mathcal{Y}^+$:

$$\left|\Pr\left[\tilde{y} \in \mathcal{C}_{\lambda_{\text{opt}}}(\boldsymbol{x}_{\text{test}}) \mid F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g_a, \tilde{y}) = 1\right] - \Pr\left[\tilde{y} \in \mathcal{C}_{\lambda_{\text{opt}}}(\boldsymbol{x}_{\text{test}}) \mid F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g_b, \tilde{y}) = 1\right]\right| \leq c, \quad (5)$$

where $c$ is a user-specified closeness criterion dictating how strictly fairness must be satisfied. CF searches over $\Lambda = [\lambda_0, \lambda_{\text{max}}]$ to minimize $\lambda_{\text{opt};(y)}$ (thereby the prediction set size) while satisfying Equation 5. CF sets $\lambda_0 = \hat{q}(\alpha)$ to retain the coverage guarantee $\Pr\left[y \in \mathcal{C}_{\lambda_{\text{opt}}}(\boldsymbol{x}_{\text{test}})\right] \geq 1 - \alpha$, while $\lambda_{\text{max}}$ is the largest value the non-conformity score may take. To determine if a threshold $\lambda$ satisfies (5) under exchangeability, CF finds tight upper and lower bounds $U_{\text{cov}}(\lambda, F_M, g, \tilde{y})$ and $L_{\text{cov}}(\lambda, F_M, g, \tilde{y})$ such that $L_{\text{cov}}(\lambda, F_M, g, \tilde{y}) \leq \Pr\left[\tilde{y} \in \mathcal{C}_{\lambda_{\text{opt}}}(\boldsymbol{x}_{\text{test}}) \mid F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, \tilde{y}) = 1\right] \leq U_{\text{cov}}(\lambda, F_M, g, \tilde{y})$ and formally computes the coverage gap,

$$\text{cg}(\lambda, F_M, \tilde{y}, \mathcal{G}) := \max_{g_a \in \mathcal{G}} \{U_{\text{cov}}(\lambda, F_M, g_a, \tilde{y})\} - \min_{g_b \in \mathcal{G}} \{L_{\text{cov}}(\lambda, F_M, g_b, \tilde{y})\}. \quad (6)$$

For a given positive label $\tilde{y}$, CF returns the minimum $\lambda$ that satisfies $\text{cg}(\lambda, F_M, \tilde{y}, \mathcal{G}) \leq c$ as $\lambda_{\text{opt};(\tilde{y})}$. For labels $y \notin \mathcal{Y}^+$, CF returns $\lambda_{\text{opt};(y)} = \lambda_0$. With this procedure, CF offers a theoretically grounded way to bound fairness disparities (Eq. (5)) without sacrificing CP's finite-sample coverage. Note that CF does not aim to equalize coverage across groups or classes (like other works [Romano et al., 2020b]), but instead controls the *coverage gap* across groups for different labels. More details on Conformal Fairness can be found in Appendix C.

### 2.5 Related Works

While standard conformal prediction (CP) provides marginal finite-sample coverage, it can yield unequal coverage for protected groups [Cresswell et al., 2025, Gao et al., 2026]. Equalized-coverage methods address this by applying conformal calibration to each group [Romano et al., 2020a]. Extensions of this work have been applied to medical imaging, post-deployment clinical bias auditing, equal opportunity, and overlapping subpopulations [Lu et al., 2022, Sun et al., 2025, Wang et al., 2024, Jung et al., 2023, Bastani et al., 2022]. Related conditional methods consider predefined, class-based, incomplete-data, or learned group structures [Vovk, 2012, Zhou and Sesia, 2024, Kong et al., 2026, Xu et al., 2026]. However, equal coverage need not imply equal downstream impact, and equalized coverage and set size may conflict [Cresswell et al., 2025, Liu et al., 2026]. Counterfactually fair CP instead targets invariance under interventions on protected attributes, but requires causal assumptions and counterfactual access [Guldogan et al., 2026]. We build on Conformal Fairness [Vadlamani et al., 2025] because its filter-based formulation supports multiple group-fairness criteria, controls disparities in decision-relevant label inclusion, preserves marginal coverage, and does not require group membership at inference time. These properties are well suited to federated settings, where group distributions differ across clients and existing federated CP methods primarily provide marginal guarantees [Lu et al., 2023, Humbert et al., 2023, 2024].

## 3 FedCF Theory and Methodology

In this section, we begin by establishing the theoretical and methodological foundations. We introduce a descent-based reformulation of the CF Framework (3.1) that is conducive to the FL setting. Next, we extend the notion of the *coverage gap* (3.2) and clarify our assumptions and the theoretical guarantees we can derive. Finally, we present the Federated Conformal Fairness (FedCF) Framework.

### 3.1 Revisiting the Conformal Fairness Algorithm

One drawback of the original CF algorithm is that it performs a grid search over a *discretized space* to find the minimal $\lambda$ satisfying the fairness specification. This requires computing the coverage

gap for each positive label every iteration, which is inefficient in the FL setting due to client-server communication. We introduce a *descent-based*[1] CF algorithm also seen in Algorithm C.1.

**Input:** The core optimization step requires a positive label, $\tilde{y} \in \mathcal{Y}^+$, the clients' calibration data, $\{\mathcal{D}^{(k)}_{\text{calib}}\}_{k\in\mathcal{K}}$ and a filter function, $F_M$ to compute the coverage gap across the set of groups $\mathcal{G}$. For fairness evaluation, we specify a closeness criterion, $c$. For the optimization hyperparameters, we specify the number of optimization rounds, initial learning rate $\eta$, and momentum $\mu$. We initialize $\lambda_0 = \hat{q}(\alpha)$ as defined in (3) to ensure the original coverage guarantee (4) is satisfied.

**Problem:** Given the inputs, let $\lambda \in \Lambda$ be a threshold and $\text{cg}(\lambda, F_M, \tilde{y}, \mathcal{G})$ be the coverage gap evaluated at $\lambda$. The optimization problem for CF can be formulated as

$$\forall \tilde{y} \in \mathcal{Y}^+ : \lambda_{\text{opt};(\tilde{y})} = \min_{\lambda\in\Lambda} \lambda \quad \text{subject to } \text{cg}(\lambda, F_M, \tilde{y}, \mathcal{G}) - c \leq 0. \tag{7}$$

**Procedure:** In round $t$, for each $\tilde{y} \in \mathcal{Y}^+$, given the current threshold $\lambda^{(t)}_{(\tilde{y})}$ and current optimal (and satisfactory) threshold $\bar{\lambda}^{(t)}_{\text{opt};(\tilde{y})}$, we can compute the coverage gap $\text{cg}_t \coloneqq \text{cg}(\lambda^{(t)}_{(\tilde{y})}, F_M, \tilde{y}, \mathcal{G})$ and evaluate whether it adheres to our fairness constraint. The update rule for $\lambda^{(t)}_{(\tilde{y})}$ is,

$$\lambda^{(t+1)}_{(\tilde{y})} = \lambda^{(t)}_{(\tilde{y})} + \eta\, 2^{-p_t}\, b_{t+1}, \quad b_{t+1} = \mu b_t + (\text{cg}_t - c), \quad p_t = \max\{\lceil \log_2(\eta/a_{\max})\rceil, 0\}, \tag{8}$$

$$a_{\max} = \min\left\{\eta,\ \frac{\Delta\lambda}{|b_{t+1}| + \epsilon}\right\}, \quad \Delta\lambda = \begin{cases} \bar{\lambda}^{(t)}_{\text{opt};(\tilde{y})} - \lambda^{(t)}_{(\tilde{y})}, & b_{t+1} \geq 0, \\ \lambda^{(t)}_{(\tilde{y})} - \lambda_0, & b_{t+1} < 0, \end{cases} \tag{9}$$

Our approach does not stop once a satisfactory $\lambda$ is found. We continue exploring smaller values (unlike in a discretized grid). In the case where our modified step size is $b_{t+1} > 0$, meaning the update rule takes us to a higher $\lambda_{t+1}$, and satisfies the fairness constraint, we perform a random restart to promote further exploration and escape local minima, since $\text{cg}(\lambda, F_M, \tilde{y}, \mathcal{G})$ is piecewise-monotonic Milnor and Thurston [2006]. For more details, see Appendix C and Algorithm C.1 (lines 13 -15). We note that this algorithm directly applies to the federated setting, with one important consideration: the coverage gap computation.

### 3.2 Extending Coverage Gap to the Federated Setting

Table 1: Important notation used for coverage gap calculation.

| Notation | Definition | Notation | Definition |
|---|---|---|---|
| $n_k$ | $\left\lvert\mathcal{D}^{(k)}_{\text{calib}}\right\rvert$ | $\mathcal{D}^{(k)}_{\text{calib}}$ | Client $k$'s calibration dataset. |
| $n^{(g,\tilde{y})}_k$ | $\left\lvert\mathcal{S}^{(g,\tilde{y})}_k\right\rvert$ | $\mathcal{S}^{(g,\tilde{y})}_k$ | $\left\{(\boldsymbol{x}_i, y_i) \in \mathcal{D}^{(k)}_{\text{calib}} \mid F_M(\boldsymbol{x}_i, y_i, g, \tilde{y}) = 1\right\}$ |
| $\gamma_k$ | $\Pr[E_k]$ | $E_k$ | The event $\boldsymbol{x}_{\text{test}}$ is exchangeable with $\mathcal{D}^{(k)}_{\text{calib}}$. |
| $\pi^{(g,\tilde{y})}$ | Point estimate for Term (IV) in Equation 10. | $L^{(g,\tilde{y})}, U^{(g,\tilde{y})}$ | Bounds for Term (IV) in Equation 10. |
| | | $\alpha^{(g,\tilde{y});\lambda}_k$ | $\sum_{(\boldsymbol{x}_i,_)\in\mathcal{S}^{(g,\tilde{y})}_k} \mathbf{1}[s(\boldsymbol{x}_i, \tilde{y}) \leq \lambda]$ |

**Assumptions & Notation:** To account for heterogeneity in client data in FL, the presented theory assumes partial exchangeability. It requires only that one instance of a group-label pair be observed across all clients (cf. Remark 3.3). No further distributional assumptions are made here. For clarity, Table 1 summarizes the notation used in our main theorem, with a complete table in Appendix A.

**Mixture-of-Clients Coverage Gap Formulation.** In the federated setting, since the data is decentralized, we cannot directly estimate the terms in (5). Thus, we rewrite the quantity as:

$$\underbrace{\Pr[s(\boldsymbol{x}_{\text{test}}, \tilde{y}) \leq \lambda \mid F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, \tilde{y}) = 1]}_{\Pi_{cov}} = \sum_{k=1}^{K} \Big( \underbrace{\Pr[s(\boldsymbol{x}_{\text{test}}, \tilde{y}) \leq \lambda \mid F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, \tilde{y}) = 1, E_k]}_{\text{(I)}}$$
$$\cdot \underbrace{\Pr[F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, \tilde{y}) = 1 \mid E_k]}_{\text{(II)}} \cdot \underbrace{\Pr[E_k]}_{\text{(III)}} \Big) \cdot \Big( \underbrace{\Pr[F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, \tilde{y}) = 1]}_{\text{(IV)}} \Big)^{-1} \tag{10}$$

[1] We use the term descent-based since we are adaptively searching for a minimal $\lambda$ that satisfies the fairness constraint. The term descent-based does not imply the use of gradients [Liu et al., 2020, Golovin et al., 2020].

With this reformulation, we can estimate individual terms–either locally on each client or globally on the server. The following theorem presents bounds for the fairness-specific coverage level. Bounds for the individual terms are given in Lemmas D.1, D.2, and D.3. The corresponding proofs for the lemmas and theorems are provided in Appendix D.

**Theorem 3.1** (Interval Bounds)**.** *The fairness-specific coverage level, $\Pi_{cov}$, given by Equation 10, can be bounded as $L_{cov}(\lambda, F_M, g, \tilde{y}) \leq \Pi_{cov} \leq U_{cov}(\lambda, F_M, g, \tilde{y})$, where*

$$L_{cov}(\lambda, F_M, g, \tilde{y}) = \sum_{k=1}^{K} \frac{\gamma_k \alpha_k^{(g,\tilde{y});\lambda}}{(n_k+1)U^{(g,\tilde{y})}} \text{ and } U_{cov}(\lambda, F_M, g, \tilde{y}) = \sum_{k=1}^{K} \frac{\gamma_k (\alpha_k^{(g,\tilde{y});\lambda}+1)}{(n_k+1)L^{(g,\tilde{y})}}. \quad (11)$$

*and $L^{(g,\tilde{y})} = \sum_{k=1}^{K} \gamma_k \frac{n_k^{(g,\tilde{y})}}{n_k+1}$ and $U^{(g,\tilde{y})} = \sum_{k=1}^{K} \gamma_k \frac{n_k^{(g,\tilde{y})}+1}{n_k+1}$*

Theorem 3.1 gives us bounds for the coverage level, which translate into bounds for the coverage gap between groups for fairness evaluation. We note that $\lambda_{\text{opt}} = \lambda_{\text{max}}$ is a degenerate solution satisfying the closeness criterion, and formally show that the condition to ensure entries of $\lambda_{\text{opt}} < \lambda_{\text{max}}$:

**Theorem 3.2.** *Given a closeness criterion, c and $\gamma_k = \frac{n_k+1}{N+K}$, in order to have non-vacuous prediction sets (i.e., $\lambda < \lambda_{max}$), we need $N^{(g,\tilde{y})} := \sum_{k=1}^{K} n_k^{(g,\tilde{y})} \geq \frac{2K}{c}$ for all $(g, \tilde{y}) \in \mathcal{G} \times \mathcal{Y}^+$.*

**Remark 3.3.** We only require a certain amount of data **across** the federation–not per-client.

### 3.2.1 Approaches to Improve Prediction Set Size (Efficiency)

As seen in Vadlamani et al. [2025], improving fairness inevitably comes at the cost of utility (efficiency). To address this tension, we explore two approaches.

**I. MLE:** FedCF relies on the broad assumption of partial exchangeability, similar to FCP, resulting in more conservative bounds. With stronger assumptions, we can get a less conservative bound. For example, suppose the data satisfies a notion of *partial IID*, that is, for each client $k$, with probability $\gamma_k$, $\left\{ s(\boldsymbol{x}_1^{(k)}, y_1^{(k)}), \ldots, s(\boldsymbol{x}_{n_k}^{(k)}, y_{n_k}^{(k)}), s(\boldsymbol{x}_{\text{test}}, y_{\text{test}}) \right\}$ is *IID*. Note this is **not** the same as assuming IID across the federation. Then we can construct a point estimate of our coverage bound using maximum likelihood estimation (MLE). This gives us a tighter estimate, but may violate the earlier finite-sample guarantees. We formalize this in the following theorem with the proof in Appendix D.

**Theorem 3.4** (Point (MLE) Estimate)**.** *If we assume partial-IID, using MLE estimates for each term, we get the following estimate for the fairness-specific coverage level, $\Pi_{cov} = \sum_{k=1}^{K} \frac{\gamma_k \alpha_k^{(g,\tilde{y});\lambda}}{n_k \pi^{(g,\tilde{y})}}$.*

**II. Using Group Information** FedCF does not require group-information at inference time. However, FedCF can be modified to use group information, such that $\lambda_{\text{opt}} \in \Lambda^{|\mathcal{Y}| \times |\mathcal{G}|}$. For a label $y$ the threshold vector is $\lambda_{\text{opt};(y)} \in \Lambda^{|\mathcal{G}|}$, the problem becomes,

$$\min_{\lambda \in \Lambda^{|\mathcal{G}|}} f(\lambda) \quad \text{subject to} \quad \text{cg}(\lambda, F_M, \tilde{y}, \mathcal{G}) - c \leq 0, \quad (12)$$

where $f$ is the scalar objective, and cg is redefined to accommodate different thresholds for each group. In this work, we take $f(\lambda) = \lambda^\top 1$ to be the elementwise sum and approximately solve (12). Theoretical details of problem (12) and how we solve it are in Appendix F.

### 3.3 FedCF: Federated Conformal Fairness framework

Having established the sufficient terms to compute the fairness-specific coverage gap, we now present the FedCF framework, which is depicted in Figure 1. We discuss FedCF in the context of the interval-bounds estimates from Theorem 3.1, noting the discussion also applies to the point-estimate case by setting $L_{\text{cov}} = U_{\text{cov}} = \Pi_{\text{cov}}$. Recall, the coverage gap is given by Equation 6 or equivalently,

$$\text{cg}(\lambda, F_M, \tilde{y}, \mathcal{G}) = \max_{g_a, g_b \in \mathcal{G}} \{U_{\text{cov}}(\lambda, F_M, g_a, \tilde{y}) - L_{\text{cov}}(\lambda, F_M, g_b, \tilde{y})\}. \quad (13)$$

While Equations 6 and 13 are mathematically equivalent, their formulations lead to two different communication and aggregation strategies, demonstrating the tradeoff between **communication overhead** and **privacy**. We present the *communication efficient* protocol in the main paper and the *enhanced privacy* protocol in Appendix H. In Appendix H, we also present a hybrid protocol, where clients select whether to use the *communication efficient* or *enhanced privacy* protocol. We include FedCF extensions concerning differential privacy in Appendix I.

Note that $U_{\text{cov}}$ and $L_{\text{cov}}$ depend on $L^{(g,\tilde{y})}$ and $U^{(g,\tilde{y})}$, respectively. Since $L^{(g,\tilde{y})}$ and $U^{(g,\tilde{y})}$ are also computed in a federated manner on the server, we compute them *prior* to computing the coverage gap for any particular $\lambda$ value[2]. Given that these priors are available on the server, we can compute the fairness-specific coverage gap. From Theorem 3.1, each client computes and sends two values for each $(g, \tilde{y}) \in \mathcal{G} \times \mathcal{Y}^+$ pair: $\frac{\alpha_k^{(g,\tilde{y});\lambda}}{n_k+1}$ and $\frac{\alpha_k^{(g,\tilde{y});\lambda}+1}{n_k+1}$. Once the server receives these pairs from each client, it aggregates these quantities to derive $L_{\text{cov}}$ and $U_{\text{cov}}$ for each $(g, \tilde{y})$ and computes the coverage gap (Equation 6). $U_{\text{cov}}$ is limited to 1 to reconcile $\Pr[\,\cdot\,] \leq 1$ for any event.

**Communication Complexity and Privacy Implications.** Each client is responsible for sending messages of size totaling $\mathcal{O}(2 \cdot |\mathcal{G}||\mathcal{Y}^+|)$ to the server per server round. While this is linear in terms of the number of $(g, \tilde{y})$ pairs, we note that with enough $\lambda$s, the server can learn the distribution of $\Pr[s(\boldsymbol{x}_{\text{test}}, \tilde{y}) \leq \lambda \mid F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, \tilde{y}) = 1, E_k]$. For the communication-efficient protocol, the total computational time complexity over $T$ server rounds is $\mathcal{O}\Big(T|\mathcal{Y}^+|\Big(K + \max_k \sum_{g\in\mathcal{G}} n_k^{(g,\tilde{y})}\Big)\Big)$, where $K$ denotes the number of clients.

# 4 Experiments

## 4.1 Setup.

**Datasets.** We evaluate the FedCF framework on four multi-class datasets: (1, 2) ACSIncome and ACSEducation [Ding et al., 2021], (3) Pokec-{n, z} [Takac and Zabovsky, 2012], (4) Fitzpatrick Groh et al. [2021]. Since these datasets were not originally designed for FL, we partition them into clients. For ACS, we use U.S. state/territory data with **six** regional schemes, yielding 4 (small), 8 (large), or 51 (all) clients, plus continental-only variants (4, 8, and 48 clients, respectively). For Pokec-{n, z}, each graph is treated as a client, as they are from distinct partitions of the larger *Pokec* social network. Finally, for Fitzpatrick, since there is no natural partitioning scheme, we apply a Dirichlet partitioner [Yurochkin et al., 2019] with $\alpha = 0.5$ to create $K \in \{2, 4, 8\}$ clients. We use a $30/20/25/25$ stratified split for the full $\mathcal{D}_{\text{train}}/\mathcal{D}_{\text{valid}}/\mathcal{D}_{\text{calib}}/\mathcal{D}_{\text{test}}$.

**Base Models.** For the ACS datasets, we use XGBoost [Chen and Guestrin, 2016]. For Pokec-n,z, we use GraphSAGE with GCN aggregation [Hamilton et al., 2017]. For Fitzpatrick, we use ResNet-18 [He et al., 2016]. GraphSAGE and ResNet are trained with FedAvg [McMahan et al., 2017], while XGBoost uses FedXgbBagging from Flower [Beutel et al., 2020].

**Baseline.** We construct a federated (fairness-agnostic) conformal predictor targeting a coverage level of $1 - \alpha = 0.9$ using FCP with T-Digest. We set $\gamma_k \propto (n_k + 1)$. For the non-conformity score, we use APS [Romano et al., 2020c] and RAPS [Angelopoulos et al., 2022] for all datasets, and DAPS [H. Zargarbashi et al., 2023], a graph-specific method, for Pokec-{n,z}. Descriptions of the scores are in Appendix E.5. We then assess fairness using $\lambda = \hat{q}(\alpha)$ for three popular group-fairness metrics, reformulated in Table B.1–Demographic Parity, Equal Opportunity, and Predictive Equality.

**Evaluation Metrics:** We report two key metrics: (1) *efficiency*, and (2) *worst-case fairness disparity*. The latter captures the largest difference in conditional coverage across groups under the chosen fairness metric. For example, under *Demographic Parity*, we report:

$$\max_{\tilde{y}\in\mathcal{Y}^+} \max_{g_a,g_b\in\mathcal{G}} \Big|\Pr\big[\tilde{y} \in \mathcal{C}_\lambda(\boldsymbol{x}_{test}) \mid \boldsymbol{x}_{test} \in g_a\big] - \Pr\big[\tilde{y} \in \mathcal{C}_\lambda(\boldsymbol{x}_{test}) \mid \boldsymbol{x}_{test} \in g_b\big]\Big|. \tag{14}$$

All methods achieve the desired $1 - \alpha$ coverage since FedCF only ever increases $\lambda$. Coverage results are included in Appendix K. Additional details on datasets and experimental setup are in Appendix E.

## 4.2 Results

In each figure, we use a **solid** line to represent the *average* efficiency of the **base federated conformal predictors** across different thresholds and a **dashed** line to represent the corresponding *average* worst-case fairness disparity. The bar plot shows the efficiency and worst-case fairness disparity using FedCF, while the **dots** indicate the *desired* fairness disparity. We report the average base performance

[2]The prior term is computed analogously to the coverage gap.

for clarity and readability. In all experiments, FedCF meets the fairness disparity bound under the closeness criterion $c$ (up to minor finite-sample variation), unlike the baseline FCP.

**Preserves Key Characteristics of CF.** Two important characteristics of the CF framework are that it is (1) agnostic to the specific non-conformity score function and (2) supports intersectional fairness. We demonstrate that our FedCF framework preserves these two characteristics via the Pokec-{n, z} dataset. Pokec-{n, z} each have two sensitive attributes: *region* and *gender*. In addition to considering each attribute individually, we can treat each *pair* of attributes as distinct and apply FedCF. Furthermore, Pokec-{n, z} is a graph dataset. Recently, several developments have been made in graph CP research on non-conformity scores that utilize the graph structure. In addition to two standard CP methods–APS and RAPS–we also provide results using DAPS. Figure 2 shows how the FedCF framework can achieve the desired fairness criterion with minimal cost to efficiency for different non-conformity scores and when considering multiple groups.

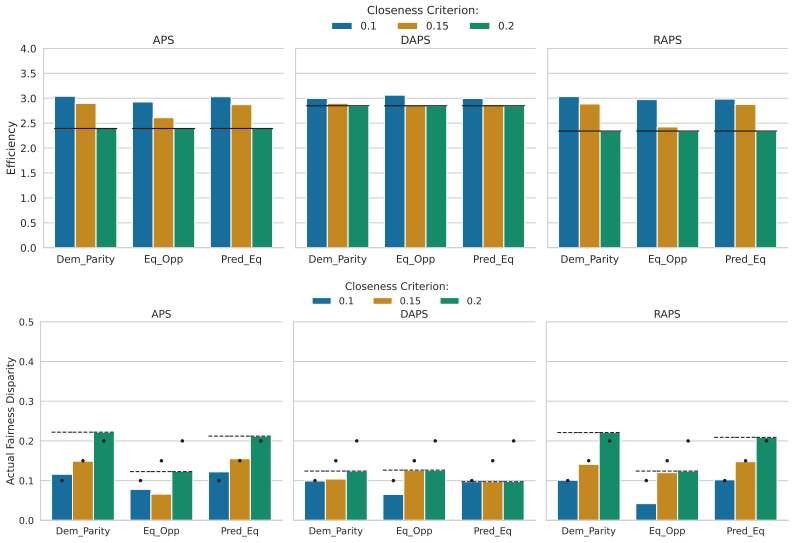



Figure 2: **Pokec-{n, z}** using **both** sensitive attributes. The top plots present the efficiency results, while the bottom plots are for the fairness disparities for (a) APS, (b) DAPS, and (c) RAPS. In all cases, FedCF achieves the desired closeness criteria better than the base federated conformal predictors.

**Robust Performance with Different Numbers of Clients.** An important trade-off in trustworthy FL is between predictive utility and maintaining fairness/privacy guarantees for each of its clients, which is increasingly challenging as the number of clients grows [Wen et al., 2023]. To demonstrate how our framework can adapt to the number of clients, we use a Dirichlet partitioner with the Fitzpatrick dataset to evaluate performance with $K \in \{2, 4, 8\}$ clients in addition to a centralized setup with a single client. In Table 2, as the number of clients increases, the baseline performance worsens, but the FedCF framework can still control for the necessary closeness criterion. We omit Equal Opportunity for Fitzpatrick as it is not meaningful in the context of this dataset, which aims to predict a skin condition with the sensitive attribute being skin type. People with certain skin types are more likely to develop certain skin conditions, so the true positive rates of a classifier will typically not equalize, resulting in a degenerate (meaningless) solution.

**Efficiency vs Fairness Trade-Off.** To make FedCF an actionable framework, it is essential to understand the utility trade-off when imposing fairness constraints. Using the interval-based approach to estimate the coverage gap gives a finite-sample guarantee for controlling fairness gaps. However, sometimes imposing fairness results in a severe cost to utility. For example, a degenerate conformal predictor (one with near-full efficiency) is "fair," but completely impractical for use. We provide an elbow curve illustrating the efficiency-fairness tradeoff in Figure K.8 (Appendix K.4). By relaxing theoretical guarantees, we can improve efficiency by using tighter estimates of the coverage gap, such as MLE point estimates. Figure 3 compares the efficiency and fairness disparities when using

Table 3: **ACSEducation (continental small,** 4 **clients), RAPS** for **Demographic Parity** *Average* Efficiency.

| $c$ | Classwise | (Group, Class)-wise |
|---|---|---|
| 0.1 | 5.010 | **4.779** |
| 0.15 | 4.826 | **4.439** |
| 0.2 | 3.982 | **3.957** |

Table 2: **Fitzpatrick using RAPS**. Each entry is of the form, **efficiency/fairness disparity**. We bold the lower fairness disparity value. This table contains the results for $K \in \{1, 2, 4, 8\}$ clients. We observe that FedCF consistently matches or outperforms the base federated conformal predictor.

(a) $c = 0.1$

| | 1 client | | 2 clients | | 4 clients | | 8 clients | |
|---|---|---|---|---|---|---|---|---|
| Metric | Base | Ours | Base | Ours | Base | Ours | Base | Ours |
| **Dem_Parity** | 2.356 / 0.151 | 3.647 / **0.103** | 2.542 / 0.161 | 4.122 / **0.153** | 2.846 / 0.294 | 4.684 / **0.099** | 3.504 / 0.164 | 5.241 / **0.089** |
| **Pred_Eq** | 2.356 / 0.111 | 3.647 / **0.109** | 2.543 / 0.177 | 4.139 / **0.177** | 2.841 / 0.288 | 4.564 / **0.102** | 3.502 / 0.169 | 5.293 / **0.083** |

(b) $c = 0.15$

| | 1 client | | 2 clients | | 4 clients | | 8 clients | |
|---|---|---|---|---|---|---|---|---|
| Metric | Base | Ours | Base | Ours | Base | Ours | Base | Ours |
| **Dem_Parity** | 2.356 / **0.151** | 2.356 / **0.151** | 2.542 / 0.161 | 2.775 / **0.161** | 2.840 / 0.294 | 3.349 / **0.117** | 3.505 / 0.164 | 3.863 / **0.087** |
| **Pred_Eq** | 2.356 / **0.111** | 2.356 / **0.111** | 2.543 / 0.177 | 2.778 / **0.177** | 2.839 / 0.288 | 3.331 / **0.114** | 3.500 / 0.168 | 3.870 / **0.087** |

(c) $c = 0.2$

| | 1 client | | 2 clients | | 4 clients | | 8 clients | |
|---|---|---|---|---|---|---|---|---|
| Metric | Base | Ours | Base | Ours | Base | Ours | Base | Ours |
| **Dem_Parity** | 2.356 / **0.151** | 2.356 / **0.151** | 2.550 / **0.162** | 2.550 / **0.162** | 2.841 / 0.292 | 3.259 / **0.164** | 3.504 / 0.162 | 3.586 / **0.146** |
| **Pred_Eq** | 2.356 / **0.111** | 2.356 / **0.111** | 2.543 / **0.177** | 2.543 / **0.177** | 2.837 / 0.287 | 3.256 / **0.167** | 3.503 / 0.169 | 3.525 / **0.150** |

interval bounds vs the MLE estimate on the ACSEducation dataset. We observe that with the interval bounds, we are always within the closeness criterion, but the efficiencies are quite high. Alternatively, using MLE estimates may exceed the desired closeness criterion, but be fairer than the baseline and not sacrifice as much efficiency. If available, using group information at inference time is another way to improve efficiency. Table 3 compares the efficiency of FedCF using the classwise $\lambda_{\text{opt};(y)}$ versus using a (group, class)-wise $\lambda_{\text{opt};(y,g)}$. We observe that FedCF can flexibly use available group information, when available, to improve efficiency, though it can still provide guarantees even when such information is unavailable.

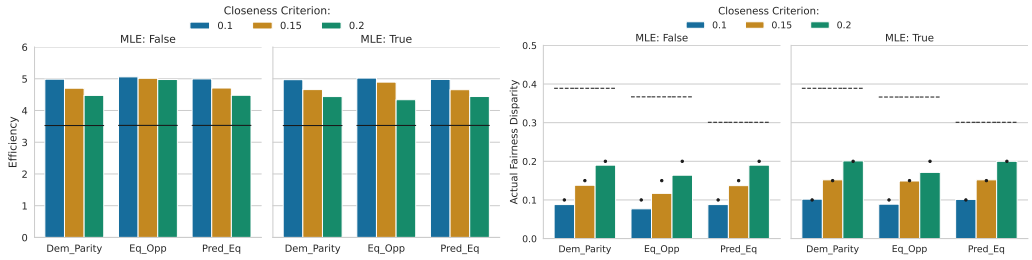



Figure 3: **ACSEducation (continental all,** 48 **clients) using RAPS.** The left two plots show the efficiency for (a) using the interval bounds and (b) using the MLE estimate. Similarly, the two right plots show the fairness disparity (c) using the interval bounds and (d) using the MLE estimate. We observe that with the MLE estimates, FedCF achieves better efficiency at the cost of a higher worst-case fairness disparity. Both outperform the baseline in terms of fairness disparity.

**Robustness to Violations of Partial Exchangeability.** To evaluate the sensitivity of FedCF to violations of the partial exchangeability assumption, we introduce four test-time perturbations: (1) *Gaussian noise*, where we perturb the predicted softmax outputs with $\mathcal{N}(0, 0.1)$ noise; (2) *shuffle group*, where we randomly permute sensitive group memberships for 25% of clients; (3) *subsample*, where we randomly remove instances from each (group, label) pair; and (4) *swap extreme groups*, where we exchange memberships between the most and least represented groups. As shown in Table 4, FedCF demonstrates strong robustness to group membership perturbations, with fairness violations remaining comparable to or below the no-shift baseline and negligible changes in efficiency. In contrast, subsampling leads to substantially larger fairness violations, increasing from $0.1051$ to $0.1700$ at $c = 0.1$ and from $0.2049$ to $0.3781$ at $c = 0.2$. Gaussian perturbations also increase fairness violations at tighter closeness criteria, although the effect is less pronounced. These results highlight that FedCF's robustness depends on the nature of the distributional perturbation. and that fairness guarantees may no longer hold when partial exchangeability is violated.

Table 4: **Robustness to violations of partial exchangeability.** Each entry is of the form **efficiency / worst-case fairness violation**. We evaluate FedCF under various distribution shifts and data perturbation (compared to a baseline with no perturbation). We observe that under some shifts, FedCF can satisfy the fairness specification.

| | **No Shift** | **Data Perturbation** | **Distribution Shift** | | |
|---|---|---|---|---|---|
| $c$ | FedCF | Gaussian Noise | Shuffle Group (25%) | Subsample | Swap Extreme Groups |
| 0.10 | 2.9791 / 0.1051 | 2.8696 / 0.1345 | 2.9785 / 0.0850 | 2.9599 / 0.1700 | 2.9788 / 0.1051 |
| 0.15 | 2.7163 / 0.1553 | 2.6754 / 0.1630 | 2.7160 / 0.1211 | 2.7661 / 0.2273 | 2.7162 / 0.1553 |
| 0.20 | 2.5846 / 0.2049 | 2.5664 / 0.1918 | 2.5847 / 0.1535 | 2.5982 / 0.3781 | 2.5841 / 0.2049 |

# 5 Discussion

**On Data Heterogeneity.** In a practical federated setting, the data distribution varies across clients–resulting in data heterogeneity, which can affect performance at inference time [Wen et al., 2023]. To address these concerns, we evaluate FedCF on varying partitioning schemes. For Fitzpatrick, we use a probabilistic partitioning scheme to ensure the data is distributed in a particular manner. For the ACS and Pokec-{n, z} datasets, the partitioning is naturally induced by state and region properties, resulting in real-world data heterogeneity. Across all partitioning schemes, we show FedCF controls for the desired closeness criterion, $c$, and adapts to data heterogeneity.

**On Data Requirements.** A limitation of CP is that, to achieve a desired coverage rate, practically, you require a large enough calibration dataset so that the interval width for the CP guarantee is tight enough. This is true in both the CF and FedCF framework as it requires sufficient calibration data for each group-positive label pair (across the federation for FedCF). For intersectional fairness, the multiplicative increase in the number of groups further increases the data requirements.

**On Framework Extensibility.** In this work, we've presented the base FedCF framework and introduced several key axes along which FedCF can be extended. First, enforcing fairness can have utility (efficiency) costs, and FedCF can be extended to mitigate these costs using tighter MLE estimates instead of interval bounds or by leveraging group information at test time (Appendix F). Second, the data privacy can be improved in FedCF using the *Enhanced Privacy* communication approach in Appendix H or by adding truncated Laplace noise, which gives formal $(\epsilon, \delta)$-differential privacy while preserving the closeness guarantee (Appendix I.2). Third, we discuss how FedCF can assist with ML fairness auditing in Appendix J, thereby closing the ML development loop.

# 6 Conclusion

In this work, we extended the Conformal Fairness framework to a federated setting, introducing the novel and comprehensive FedCF framework. We reformulated the CF framework to use a descent-based approach to make it more efficient for FL applications. Additionally, we developed theoretically grounded protocols to enable coverage gap calculations in a federated manner. The FedCF framework offers clients a choice of participation protocols, including *communication efficient* and *enhanced privacy* options. We conducted experiments on various non-conformity scores and datasets–including graph data, where we leverage the exchangeability assumption from CP.

**Limitations.** This work assumes the notion of partial exchangeability, similar to FCP, which we build on top of; however, in real-world settings, this may be violated, for example, in settings where there is dynamic drift [Jiang et al., 2021, Ramezani-Kebrya et al., 2023, Zec et al., 2025]. To redress, one option we hope to explore is the use of non-exchangeable CP [Tibshirani et al., 2019, Barber et al., 2023] to develop fair and federated solutions [Plassier et al., 2024].

**Future Work.** We plan to explore using Robust FCP [Kang et al., 2024] in lieu of FCP and also apply the FedCF method post-hoc to work under adversarial conditions. We will also explore how FedCF can be extended to split learning [Gupta and Raskar, 2018]. Unlike FL, which trains full models locally and aggregates updates, split learning divides the model across clients and server, sharing only partial computations (enhanced privacy, reduced compute).

# 7 Statements

## 7.1 Acknowledgements

The authors acknowledge support from National Science Foundation (NSF) grant #2112471 (AI-EDGE). Any opinions and findings are those of the author(s) and do not necessarily reflect the views of the granting agencies. The authors also thank the reviewers for constructive feedback on this work.

## 7.2 Reproducibility Statement

In the spirit of reproducibility, the proofs for the theoretical aspects of our work can be found in Appendices C, D, and F. Algorithmic details are available in Appendices C and H. Details of the experiments, including datasets, non-conformity scores, and hyperparameters, are provided in Appendix E. The source code is available at `https://github.com/AdityaVadlamani/FedCF`.

## 7.3 Impact Statement

This work develops a trustworthy machine learning framework for fair uncertainty quantification and decision-making in federated settings. Using a principled mixture of clients' formulations in the federated, we can provide theoretical bounds and guarantees that translate into both rigorous fairness evaluation under uncertainty as well as guidance for ML fairness auditing. By accounting for the privacy constraints of FL, our method is suitable for use and will have positive impacts across several domains (e.g., finance and healthcare).

# References

Nimesh Agrawal, Anuj Kumar Sirohi, Sandeep Kumar, et al. No prejudice! fair federated graph neural networks for personalized recommendation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 10775–10783, 2024.

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Uncertainty sets for image classifiers using conformal prediction, 2022. URL `https://arxiv.org/abs/2009.14193`.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. The Annals of Statistics, 51(2):816 – 845, 2023. doi: 10.1214/23-AOS2276. URL `https://doi.org/10.1214/23-AOS2276`.

Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivalid conformal prediction. Advances in Neural Information Processing Systems, 35:29362–29373, 2022.

Daniel J Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Hei Li Kwing, Titouan Parcollet, Pedro PB de Gusmão, and Nicholas D Lane. Flower: A friendly federated learning research framework. arXiv preprint arXiv:2007.14390, 2020.

T.-H. Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. ACM Transactions on Information and System Security, 14(3):26:1–26:24, 2011.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL `https://doi.org/10.1145/2939672.2939785`.

John Cherian and Lenny Bronner. How the washington post estimates outstanding votes for the 2020 presidential election. Retrieved September, 13:2023, 2020.

Jesse C Cresswell, Bhargava Kumar, Yi Sui, and Mouloud Belbahri. Conformal prediction sets can cause disparate impact. In The Thirteenth International Conference on Learning Representations, 2025.

Bruno de Finetti. On the condition of partial exchangeability. In Richard C. Jeffrey, editor, Studies in Inductive Logic and Probability, pages 193–205. University of California Press, 1971.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. Advances in neural information processing systems, 34:6478–6490, 2021.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. Journal of the Royal Statistical Society Series B: Statistical Methodology, 84(1):3–37, 2022.

Ted Dunning. The t-digest: Efficient estimates of distributions. Software Impacts, 7:100049, 2021. ISSN 2665-9638. doi: https://doi.org/10.1016/j.simpa.2020.100049. URL https://www.sciencedirect.com/science/article/pii/S2665963820300403.

Cynthia Dwork. Differential privacy. In International colloquium on automata, languages, and programming, pages 1–12. Springer, 2006.

Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N. Rothblum. Differential privacy under continual observation. In Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC), pages 715–724, 2010a.

Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. Boosting and differential privacy. In IEEE 51st Annual Symposium on Foundations of Computer Science (FOCS), pages 51–60, 2010b.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. Foundations and trends® in theoretical computer science, 9(3–4):211–407, 2014.

Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 2468–2479. SIAM, 2019.

European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2019/882. Official Journal of the European Union, 2024. URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

Yahya H. Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and Salman Avestimehr. Fairfed: enabling group fairness in federated learning. In Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i6.25911. URL https://doi.org/10.1609/aaai.v37i6.25911.

Arya Fayyazi, Mehdi Kamal, and Massoud Pedram. Facter: Fairness-aware conformal thresholding and prompt engineering for enabling fair llm-based recommender systems. In Forty-second International Conference on Machine Learning, 2025.

Ziang Gao, Pengqi Liu, Archer Yi Yang, Mouloud Belbahri, Jesse C Cresswell, and Masoud Asgharian. On the burden of achieving fairness in conformal prediction. arXiv preprint arXiv:2605.14260, 2026.

Quan Geng, Wei Ding, Ruiqi Guo, and Sanjiv Kumar. Tight analysis of privacy and utility tradeoff in approximate differential privacy. In Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS), pages 89–99, 2020.

Daniel Golovin, John Karro, Greg Kochanski, Chansoo Lee, Xingyou Song, and Qiuyi Zhang. Gradientless descent: High-dimensional zeroth-order optimization. In International Conference on Learning Representations, 2020. URL https://openreview.net/forum?id=Skep6TVYDB.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1820–1828, 2021.

John Guckenheimer. Sensitive dependence to initial conditions for one dimensional maps. Communications in Mathematical Physics, 70(2):133–160, 1979.

Ozgur Guldogan, Neeraj Sarna, Yuanyuan Li, and Michael Berger. Counterfactually fair conformal prediction. In The 29th International Conference on Artificial Intelligence and Statistics, 2026.

Otkrist Gupta and Ramesh Raskar. Distributed learning of deep neural network over multiple agents. Journal of Network and Computer Applications, 116:1–8, 2018. ISSN 1084-8045. doi: https://doi.org/10.1016/j.jnca.2018.05.003. URL https://www.sciencedirect.com/science/article/pii/S1084804518301590.

Soroush H. Zargarbashi, Simone Antonelli, and Aleksandar Bojchevski. Conformal prediction sets for graph neural networks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pages 12292–12318. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/h-zargarbashi23a.html.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, page 1025–1035, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Dennis Hirsch, Timothy Bartley, Aravind Chandrasekaran, Davon Norris, Srinivasan Parthasarathy, and Piers Norris Turner. Business Data Ethics: Emerging Models for Governing AI and Advanced Analytics. Springer Nature, 2023.

Pierre Humbert, Batiste Le Bars, Aurélien Bellet, and Sylvain Arlot. One-shot federated conformal prediction. In International Conference on Machine Learning, pages 14153–14177. PMLR, 2023.

Pierre Humbert, Batiste Le Bars, Aurélien Bellet, and Sylvain Arlot. Marginal and training-conditional guarantees in one-shot federated conformal prediction. arXiv preprint arXiv:2405.12567, 2024.

Thomas Humphries, Rasoul Akhavan Mahdavi, Shannon Veitch, and Florian Kerschbaum. Selective MPC: Distributed computation of differentially private key-value statistics. arXiv preprint arXiv:2107.12407, 2021.

Meirui Jiang, Zirui Wang, and Qi Dou. Harmofl: Harmonizing local and global drifts in federated learning on heterogeneous medical images. In AAAI Conference on Artificial Intelligence, 2021. URL https://api.semanticscholar.org/CorpusID:245353364.

Charles Jones, Fabio De Sousa Ribeiro, Mélanie Roschewitz, Daniel C Castro, and Ben Glocker. Rethinking fair representation learning for performance-sensitive tasks. In The Thirteenth International Conference on Learning Representations, 2025.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. 11th International Conference on Learning Representations (ICLR), 2023.

Samhita Kanaparthy, Manisha Padala, Sankarshan Damle, and Sujit Gujar. Fair federated learning for heterogeneous data. In Proceedings of the 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD), CODS-COMAD '22, page 298–299, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450385824. doi: 10.1145/3493700.3493750. URL https://doi.org/10.1145/3493700.3493750.

Mintong Kang, Zhen Lin, Jimeng Sun, Cao Xiao, and Bo Li. Certifiably byzantine-robust federated conformal prediction. In Proceedings of the 41st International Conference on Machine Learning, pages 23022–23057, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. CoRR, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

CR Komala, Ashok Kumar, N Hema, S Nagarani, Ajay Singh Yadav, M Rajendiran, R Srinivasan, and V Vijayan. Fair-automl: Enhancing fairness in machine learning predictions through automated machine learning and bias mitigation techniques. In AIP Conference Proceedings, volume 3193, page 020005. AIP Publishing LLC, 2024.

Jingsen Kong, Yiming Liu, Guangren Yang, and Ding Zhong. Fair conformal prediction for incomplete covariate data. Statistics and Computing, 36(1):23, 2026.

Pengqi Liu, Zijun Yu, Mouloud Belbahri, Arthur Charpentier, Masoud Asgharian, and Jesse C. Cresswell. Beyond procedure: Substantive fairness in conformal prediction. In International Conference on Machine Learning, 2026.

Sijia Liu, Pin-Yu Chen, Bhavya Kailkhura, Gaoyuan Zhang, Alfred O. Hero III, and Pramod K. Varshney. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. IEEE Signal Processing Magazine, 37(5):43–54, 2020. doi: 10.1109/MSP.2020.3003837.

Kuan-Chieh Lo, Yuntian He, Yuze Jiang, and Srinivasan Parthasarathy. Fairwag: Fairness-aware weighted aggregation for graph learning in a federated setting. In Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO '25, page 119–150, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400721403. doi: 10.1145/3757887.3763013. URL https://doi.org/10.1145/3757887.3763013.

Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. Proceedings of the AAAI Conference on Artificial Intelligence, 36(11):12008–12016, Jun. 2022. doi: 10.1609/aaai.v36i11.21459. URL https://ojs.aaai.org/index.php/AAAI/article/view/21459.

Charles Lu, Yaodong Yu, Sai Praneeth Karimireddy, Michael Jordan, and Ramesh Raskar. Federated conformal predictors for distributed uncertainty quantification. In International Conference on Machine Learning, pages 22942–22964. PMLR, 2023.

Pranav Maneriker, Codi Burley, and Srinivasan Parthasarathy. Online fairness auditing through iterative refinement. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '23, page 1665–1676, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599454. URL https://doi.org/10.1145/3580305.3599454.

Pranav Maneriker, Aditya T Vadlamani, Anutam Srinivasan, Yuntian He, Ali Payani, et al. Conformal prediction: A theoretical note and benchmarking transductive node classification in graphs. Transactions on Machine Learning Research, May 2025.

Charles Masson, Jee E. Rim, and Homin K. Lee. Ddsketch: a fast and fully-mergeable quantile sketch with relative-error guarantees. Proc. VLDB Endow., 12(12):2195–2205, August 2019. ISSN 2150-8097. doi: 10.14778/3352063.3352135. URL https://doi.org/10.14778/3352063.3352135.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, volume 54 of Proceedings of Machine Learning Research, pages 1273–1282. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6):1–35, 2021.

John Milnor and William Thurston. On iterated maps of the interval. In Dynamical Systems: Proceedings of the Special Year held at the University of Maryland, College Park, 1986–87, pages 465–563. Springer, 2006.

New York City Council. Local law 4 of 2021: Fair chance act, 7 2021. URL https://www.nyc.gov/site/cchr/law/fair-chance-act.page.

New York City Council. Local law 144 of 2021: Prohibiting automated employment decision tools. New York City Register, 7 2023. URL https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page.

Vincent Plassier, Nikita Kotelevskii, Aleksandr Rubashevskii, Fedor Noskov, Maksim Velikanov, Alexander Fishkov, Samuel Horvath, Martin Takac, Eric Moulines, and Maxim Panov. Efficient conformal prediction under data heterogeneity, 2024. URL https://arxiv.org/abs/2312.15799.

Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H. Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. How to DP-fy ML: A practical guide to machine learning with differential privacy. Journal of Artificial Intelligence Research, 77:1113–1201, 2023.

Ali Ramezani-Kebrya, Fanghui Liu, Thomas Pethick, Grigorios Chrysos, and Volkan Cevher. Federated learning under covariate shifts with generalization guarantees, 2023. URL https://arxiv.org/abs/2306.05325.

Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel Candès. With Malice Toward None: Assessing Uncertainty via Equalized Coverage. Harvard Data Science Review, 2(2), apr 30 2020a. https://hdsr.mitpress.mit.edu/pub/qedrwcz3.

Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel Candès. With malice toward none: Assessing uncertainty via equalized coverage. Harvard Data Science Review, 2020b.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. Advances in neural information processing systems, 33:3581–3591, 2020c.

Xiaotan Sun, Makiya Nakashima, Christopher Nguyen, Po-Hao Chen, WH Wilson Tang, Deborah Kwon, and David Chen. Fairicp: identifying biases and increasing transparency at the point of care in post-implementation clinical decision support using inductive conformal prediction. Journal of the American Medical Informatics Association, 32(8):1299–1309, 2025.

Lubos Takac and Michal Zabovsky. Data analysis in public social networks. In International scientific conference and international workshop present day trends of innovations, volume 1, 2012.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2019. Curran Associates Inc.

U.S. Equal Employment Opportunity Commission. Questions and answers to clarify and provide a common interpretation of the uniform guidelines on employee selection procedures. Federal Register, 44(43), 1979. URL https://www.eeoc.gov/laws/guidance/questions-and-answers-clarify-and-provide-common-interpretation-uniform-guidelines.

Aditya T. Vadlamani, Anutam Srinivasan, Pranav Maneriker, Ali Payani, and Srinivasan Parthasarathy. A generic framework for conformal fairness. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=xiQNfYl33p.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In Steven C. H. Hoi and Wray Buntine, editors, Proceedings of the Asian Conference on Machine Learning, volume 25 of Proceedings of Machine Learning Research, pages 475–490, Singapore Management University, Singapore, 04–06 Nov 2012. PMLR. URL https://proceedings.mlr.press/v25/vovk12.html.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. Algorithmic learning in a random world, volume 29. Springer, 2005.

Fangxin Wang, Lu Cheng, Ruocheng Guo, Kay Liu, and Philip S. Yu. Equal opportunity of coverage in fair regression. In Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, and Wensheng Zhang. A survey on federated learning: challenges and applications. International journal of machine learning and cybernetics, 14(2):513–535, 2023.

Senrong Xu, Yanke Zhou, Yuhao Tan, Zenan Li, Yuan Yao, Taolue Chen, Feng Xu, and Xiaoxing Ma. Fair conformal classification via learning representation-based groups. In The Fourteenth International Conference on Learning Representations, 2026.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 7252–7261. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/yurochkin19a.html.

Edvin Listo Zec, Adam Breitholtz, and Fredrik D. Johansson. Overcoming label shift with target-aware federated learning, 2025. URL https://arxiv.org/abs/2411.03799.

Yanfei Zhou and Matteo Sesia. Conformal classification with equalized coverage for adaptively selected groups. Advances in Neural Information Processing Systems, 37:108760–108823, 2024.

Hong Zhu, Lin Li, Yingying Zeng, and Zhiheng Yu. On piecewise monotone functions with height being infinity. Journal of Applied Analysis & Computation, 11(2):1062–1073, 2021.

## Appendix Table of Contents


**A Notation Table** **18**

**B Conformal Fairness Background** **18**

**C CF Optimization Approach** **19**

C.1 Existence of a Solution . . . . . 19

C.2 Convergence Analysis . . . . . 20

C.3 Empirical Justification of Design Choice . . . . . 20

C.4 Example: Synthetic Adversarial Coverage Gap Landscape . . . . . 21

**D Proofs** **23**

D.1 Proof of Theorem 3.1 . . . . . 23

D.2 Sufficient Data Counts . . . . . 25

D.3 Tightness of Theorem 3.1 . . . . . 26

**E Additional Experiment Details** **27**

E.1 Datasets . . . . . 27

E.2 Folktables Datasets . . . . . 27

E.3 Non-Tabular Datasets . . . . . 28

E.4 Hyperparameters and Implementation . . . . . 28

E.5 Non-Conformity Scores . . . . . 28

**F On Using Group Information** **29**

F.1 Quantifying the Efficiency Improvement . . . . . 30

**G FedCF vs FairFed + FCP** **31**

**H FedCF with Enhanced Privacy** **31**

H.1 Enhanced Privacy . . . . . 31

H.2 Hybrid . . . . . 35

H.3 Empirical Comparison . . . . . 35

**I Differential Privacy in FedCF** **38**

I.1 **Example:** Differential Privacy Bounds for Enhanced Privacy Protocol . . . . . 38

I.2 Strict Differential Privacy . . . . . 39

**J FedCF for Auditing** **40**

**K More Results** **41**

K.1 Coverage Results . . . . . 41

K.2 Impact of Data Heterogeneity on ACSEducation: US vs Continental US . . . . . 42

K.3 Impact of different sensitive attributes for Pokec-{n,z} . . . . . 44

K.4 Efficiency vs. Closeness Criterion . . . . . 44

# A Notation Table

Table A.1: Common notation used in FedCF.

| Notation | Definition |
|---|---|
| $\mathcal{G}$ | The set of all demographic groups. |
| $\mathcal{Y}, \mathcal{Y}^+$ | The set of labels and positive/advantaged labels, respectively. |
| $g$ and $\tilde{y}$ | The group $g \in \mathcal{G}$ and positive label $\tilde{y} \in \mathcal{Y}^+$. |
| $F_M$ | Filter function for fairness metric $M$. |
| $c$ | Closeness criterion for a fairness specification. |
| $\lambda$ | Threshold used for constructing test prediction sets. |
| $\mathcal{K}$ | The set of clients, $\{1, \ldots, K\}$. |
| $\mathcal{D}^{(k)}_{\text{train}}/\mathcal{D}^{(k)}_{\text{valid}}/\mathcal{D}^{(k)}_{\text{calib}}$ | Client $k$'s train/validation/calibration dataset. |
| $n_k$ and $N$ | $\lvert\mathcal{D}^{(k)}_{\text{calib}}\rvert$ and $\sum_{k=1}^{K} n_k$, respectively. |
| $\mathcal{S}_k^{(g,\tilde{y})}$ and $n_k^{(g,\tilde{y})}$ | $\left\{(\boldsymbol{x}_i, y_i) \in \mathcal{D}^{(k)}_{\text{calib}} \mid F_M(\boldsymbol{x}_i, y_i, g, \tilde{y}) = 1\right\}$ and $\left\lvert\mathcal{S}_k^{(g,\tilde{y})}\right\rvert$ |
| $\alpha_k^{(g,\tilde{y});\lambda}$ | $\sum_{(\boldsymbol{x}_i,_)\in\mathcal{S}_k^{(g,\tilde{y})}} \mathbf{1}[s(\boldsymbol{x}_i, \tilde{y}) \leq \lambda]$ |
| $E_k$ and $\gamma_k$ | The event $\boldsymbol{x}_{\text{test}}$ is exchangeable with data from client $k$ and $\Pr[E_k]$. |
| $L^{(g,\tilde{y})}, U^{(g,\tilde{y})}$ | Bounds for prior (term (IV)). |
| $\pi^{(g,\tilde{y})}$ | Point estimate for prior (term (IV)). |
| $L_{\text{cov}}, U_{\text{cov}}$ | Bounds for fairness-specific coverage level. |
| $\Pi_{\text{cov}}$ | Point estimate for fairness-specific coverage level. |

# B Conformal Fairness Background

**From point predictions to set-based fairness.** Let $\mathcal{C}_\lambda(\boldsymbol{x}) = \{y \in \mathcal{Y} : s(\boldsymbol{x}, y) \leq \lambda\}$ denote the CP prediction set at score threshold $\lambda$. CF adapts classical group-fairness metrics by replacing point-prediction events (i.e., the predicted label equals the advantaged label: $\tilde{y} = \hat{y}$) with set-membership events (i.e., the advantaged label is in the prediction set: $\tilde{y} \in \mathcal{C}_\lambda(\boldsymbol{x})$) and evaluating disparities across groups $\mathcal{G}$ and, when appropriate, advantaged labels $\mathcal{Y}^+$. For example, a set-based Demographic Parity-style constraint can be written as, $\left\lvert \Pr\left[\tilde{y} \in \mathcal{C}_\lambda(\boldsymbol{x}) \mid \boldsymbol{x} \in g_a\right] - \Pr\left[\tilde{y} \in \mathcal{C}_\lambda(\boldsymbol{x}) \mid \boldsymbol{x} \in g_b\right]\right\rvert \leq c$, $\forall\, g_a, g_b \in \mathcal{G},\ \tilde{y} \in \mathcal{Y}^+$ with analogous set-based forms for other common group-fairness metrics as seen in Table B.1.

**Conditional coverage as the fairness control knob.** To evaluate a chosen fairness notion, CF filters the calibration data to the relevant subpopulation (e.g., a group or a group-and-label slice) via a filter function, $F_M$, and uses conditional coverage estimates under that filter. (For example, $F_M$ for Equal Opportunity is defined as $F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, y) := \mathbf{1}[\boldsymbol{x}_{\text{test}} \in g \cap y_{\text{test}} = y]$.) It then searches a threshold space $\Lambda$ to identify $\lambda_{\text{opt}}$ that satisfies the closeness criterion across groups (and labels, if required) while maintaining CP validity. A key ingredient is that CP coverage holds when labels are fixed to a particular $\tilde{y}$, which underpins group- and class-conditional control in CF.

**Guarantees and trade-offs.** CF provides a theoretically grounded and empirically validated procedure to bound fairness disparities (Equation (5)) without sacrificing CP's finite-sample coverage guarantees. In practice, satisfying stricter fairness requirements (smaller closeness criterion $c$) increases the average prediction set size, reflecting the fairness–efficiency trade-off.

**Practical advantages.** Unlike many conditional-CP methods that require group membership at inference time or are model-specific, CF's set-based metrics and thresholding procedure do not require access to protected attributes at test time and apply to different non-conformity scores and data modalities, making it compatible with downstream deployment constraints.

Table B.1: Formulations for Conformal Fairness Metrics. Definitions hold $\forall g_a, g_b \in \mathcal{G}$ and $\forall \tilde{y} \in \mathcal{Y}^+$

| Metric | Definition |
|---|---|
| Demographic (or Statistical) Parity | $\left\lvert \Pr\left[\tilde{y} \in \mathcal{C}_\lambda(\boldsymbol{x}) \,\middle\vert\, \boldsymbol{x} \in g_a\right] - \Pr\left[\tilde{y} \in \mathcal{C}_\lambda(\boldsymbol{x}) \,\middle\vert\, \boldsymbol{x} \in g_b\right] \right\rvert < c$ |
| Equal Opportunity | $\left\lvert \Pr\left[\tilde{y} \in \mathcal{C}_\lambda(\boldsymbol{x}) \,\middle\vert\, y = \tilde{y}, \boldsymbol{x} \in g_a\right] - \Pr\left[\tilde{y} \in \mathcal{C}_\lambda(\boldsymbol{x}) \,\middle\vert\, y = \tilde{y}, \boldsymbol{x} \in g_b\right] \right\rvert < c$ |
| Predictive Equality | $\left\lvert \Pr\left[\tilde{y} \in \mathcal{C}_\lambda(\boldsymbol{x}) \,\middle\vert\, y \neq \tilde{y}, \boldsymbol{x} \in g_a\right] - \Pr\left[\tilde{y} \in \mathcal{C}_\lambda(\boldsymbol{x}) \,\middle\vert\, y \neq \tilde{y}, \boldsymbol{x} \in g_b\right] \right\rvert < c$ |

# C CF Optimization Approach

**Algorithm C.1** Descent-Based[3]CF Optimization

**function** FAIR_OPT_DESCENT($\tilde{y}$, $\lambda_0$, $\mathcal{G}$, $c$, $F_M$, num_rounds, $\eta$, $\mu$)
  $\bar{\lambda}_{\text{opt};(\tilde{y})} \leftarrow 1$
  $b_0 \leftarrow 0$
  **for** $t = 0$ to num_rounds $- 1$ **do**
    $\text{cg}_t \leftarrow \text{cg}(\lambda^{(t)}_{(\tilde{y})}, F_M, \tilde{y}, \mathcal{G})$
    $u_t \leftarrow \mathbf{1}\left[\text{cg}_t \leq c \ \wedge \ \lambda^{(t)}_{(\tilde{y})} < \bar{\lambda}_{\text{opt};(\tilde{y})}\right]$
    **if** $\text{cg}_t \leq c$ **and** $t = 0$ **then**
      **return** $\lambda_0$
    **else if** $u_t$ **then**
      $\bar{\lambda}_{\text{opt};(\tilde{y})} \leftarrow \lambda^{(t)}_{(\tilde{y})}$
    **end if**
    $b_{t+1} \leftarrow \mu\, b_t + (\text{cg}_t - c)$
    **if** $u_t$ **and** $b_{t+1} > 0$ **then**
      $b_{t+1} \leftarrow 0$
      $\lambda^{(t+1)}_{(\tilde{y})} \sim \text{Uniform}(\lambda_0, \ \bar{\lambda}_{\text{opt};(\tilde{y})})$
    **else**
      $\Delta\lambda \leftarrow (\bar{\lambda}_{\text{opt};(\tilde{y})} - \lambda^{(t)}_{(\tilde{y})})\mathbf{1}[b_{t+1} \geq 0] + (\lambda^{(t)}_{(\tilde{y})} - \lambda_0)\mathbf{1}[b_{t+1} < 0]$
      $a_{\max} \leftarrow \min\left(\eta, \ \frac{\Delta\lambda}{|b_{t+1}|+\epsilon}\right)$
      $p_t \leftarrow \max\left\{\left\lceil \log_2\left(\frac{\eta}{a_{\max}}\right)\right\rceil, \ 0\right\}$
      $\eta_t \leftarrow \min(\eta \cdot 2^{-p_t}, \ a_{\max})$
      $\lambda^{(t+1)}_{(\tilde{y})} \leftarrow \lambda^{(t)}_{(\tilde{y})} + \eta_t\, b_{t+1}$
    **end if**
  **end for**
  **return** $\bar{\lambda}_{\text{opt};(\tilde{y})}$
**end function**

## C.1 Existence of a Solution

By construction, our approach guarantees an admissible solution. We begin with the trivial solution where $\lambda$ is at its maximum (ensuring zero disparity), then use our coverage gap computation to verify if a new $\lambda$ satisfies the closeness criterion. To find the minimal satisfying $\lambda$, we employ a descent-based approach. Since the objective is non-convex, we mitigate local minima by performing random restarts within the admissible set whenever we reach the boundary of the admissible set (i.e., the current $\lambda_{\text{opt}}$).

[3]We use the term descent-based since we are adaptively searching for a minimal $\lambda$ that satisfies the fairness constraint. The term descent-based does not imply the use of gradients [Liu et al., 2020, Golovin et al., 2020].

### C.2 Convergence Analysis

For simplicity, suppose we remove the random restart step in Algorithm C.1. We will show that our algorithm finds and halts at a local minimum if we remove the random restart (Lines 13-15). This would imply that with random restarts, the algorithm can escape and explore other local minima, (heuristically) improving performance.

**Theorem C.1.** *WLOG, fix some $\tilde{y} \in \mathcal{Y}^+$. Suppose at iteration $T$, the algorithm finds a valid upper-bound threshold $\bar{\lambda}^{(T)}_{opt;(\tilde{y})}$ that satisfies $cg(\bar{\lambda}^{(T)}_{opt;(\tilde{y})}) \leq c$. Suppose further that at a subsequent iteration $t > T$, the explored threshold $\lambda^{(t)}_{(\tilde{y})} < \bar{\lambda}^{(T)}_{opt;(\tilde{y})}$ falls into a local valley where $cg(\lambda^{(t)}_{(\tilde{y})}) > c$. Then, without the random restart mechanism, the algorithm actively bounds the subsequent explored thresholds by $\bar{\lambda}^{(T)}_{opt;(\tilde{y})}$, converging asymptotically, and permanently traps itself in the local valley.*

*Proof.* For brevity, we will drop the $(\tilde{y})$ subscript and use $\lambda_t \coloneqq \lambda^{(t)}_{(\tilde{y})}$ and $\lambda_{opt} \coloneqq \bar{\lambda}^{(T)}_{\text{opt};(\tilde{y})}$. Because we are in a region where the constraint is violated, the error term is $e_t = \text{cg}(\lambda_t) - c > 0$. Per the update rule, the momentum $b_{t+1} = \mu b_t + e_t$ accumulates this positive error term. Since $e_t > 0$ and $\mu \in (0, 1)$, $b_{t+1}$ will eventually become strictly positive ($b_{t+1} > 0$). This positive momentum term indicates that the algorithm intends to step upward to satisfy the coverage gap constraint.

However, in Lines 17-21, the algorithm performs spatial clipping to prevent overshooting the known valid bound, enforcing $\lambda_{t+1} \leq \lambda_{\text{opt}}$. Let $D_t = \lambda_{opt} - \lambda_t$ be the available distance to the ceiling of our search space. In Line 18, we bound the maximum allowable step scale as:

$$a_{\max} = \min\left\{\eta, \frac{D_t}{b_{t+1} + \epsilon}\right\},$$

where $\epsilon > 0$ is a small stability term. To find the precise step scale $\eta_t$, the algorithm applies a base-2 logarithmic scaling to ensure it is the largest power of $1/2$ scaling of the base learning rate that remains less than or equal to $a_{\max}$. This discretization mathematically guarantees that:

$$\frac{1}{2} a_{\max} < \eta_t \leq a_{\max}.$$

The parameter update rule is $\lambda_{t+1} = \lambda_t + \eta_t b_{t+1}$. Subtracting both sides from $\lambda_{opt}$, we express this in terms of the distance $D_t$:

$$D_{t+1} = D_t - \eta_t b_{t+1}.$$

Substituting the lower bound for $\eta_t$ and the definition of $a_{\max}$, we obtain two cases: (1) $a_{\max} = \eta$ and (2) $a_{\max} = \frac{D_t}{b_{t+1}+\epsilon}$. In the first case, we get that

$$D_{t+1} = D_t - \eta \cdot b_{t+1},$$

meaning $\lambda_t \to \lambda_{\text{opt}}$ *linearly*. In the second case, using the lower bound for $\eta_t$, we get

$$D_{t+1} \leq D_t - \frac{1}{2}\left(\frac{D_t}{b_{t+1} + \epsilon}\right) b_{t+1} = D_t \left(1 - \frac{b_{t+1}}{2(b_{t+1} + \epsilon)}\right).$$

Let $\kappa = 1 - \frac{b_{t+1}}{2(b_{t+1}+\epsilon)}$. Since $b_{t+1} > 0$ and $\epsilon > 0$, it strictly holds that $0 < \kappa < 1$. Therefore, $D_{t+1} \leq \kappa D_t$. Thus, $\lambda_t \to \lambda_{\text{opt}}$ at a *geometric* rate. Thus, without the random restart, in either case, the algorithm eventually gets stuck at some (local) valley. □

### C.3 Empirical Justification of Design Choice

To justify the novel descent-based approach for the federated setting in terms of reducing the number of communications rounds, we define the discrete search space for the iterative method as $\Lambda = \texttt{linspace}(\hat{q}(\alpha), 2, \text{num_rounds})$, where $\hat{q}(\alpha)$ is the minimal lambda to satisfy $1 - \alpha$ coverage of standard federated CP, and 2 is the upper bound of the RAPS non-conformity score. Table C.1 shows that the iterative (grid-search) approach requires more communication rounds to achieve performance comparable to our new descent-based approach. This is because discretizing the space limits the precision of the $\lambda$ values, so, by using fewer rounds, we find a less precise, and larger (more conservative) $\lambda$. This is not a limitation for our descent-based approach, which converges to a constraint-satisfying $\lambda$ using our adaptive update rule. The results in Table C.1 show that the Iterative approach requires 1000 rounds to achieve sufficient granularity and match the performance of the Descent-Based approach, which only used 100 rounds. Thus, we demonstrate approximately

$10\times$ speedup to achieve comparable performance, justifying the descent-based adaptation for the federated learning setting.

Table C.1: **ACSIncome (small,** 4 **clients) using RAPS.** Each entry is of the form, **efficiency/fairness disparity**. We compare the Descent-Based (with 100 communication rounds) and the Iterative (grid-search) method (with 100 and 1000 communication rounds) across different closeness criteria ($c$).

| Method (Number of rounds) | $c = 0.1$ Eff / Disp | $c = 0.15$ Eff / Disp | $c = 0.2$ Eff / Disp |
|---|---|---|---|
| **Descent-Based** (100) | 3.108 / 0.145 | 2.977 / 0.179 | 2.817 / 0.257 |
| **Iterative** (100) | 3.238 / 0.119 | 3.140 / 0.140 | 2.922 / 0.213 |
| **Iterative** (1000) | 3.121 / 0.137 | 2.989 / 0.171 | 2.825 / 0.249 |

### C.4 Example: Synthetic Adversarial Coverage Gap Landscape

The coverage gap values that we encountered in our experiments demonstrated a piecewise-monotonic behavior [Milnor and Thurston, 2006, Guckenheimer, 1979, Zhu et al., 2021], meaning there are a small, but finite number of turning points (see Figure C.1). To demonstrate that our algorithm works in more adversarial settings, we provide an experiment with a synthetic, high-oscillation, coverage gap function and show that Algorithm C.1 finds a satisfactory solution (see Figure C.2).

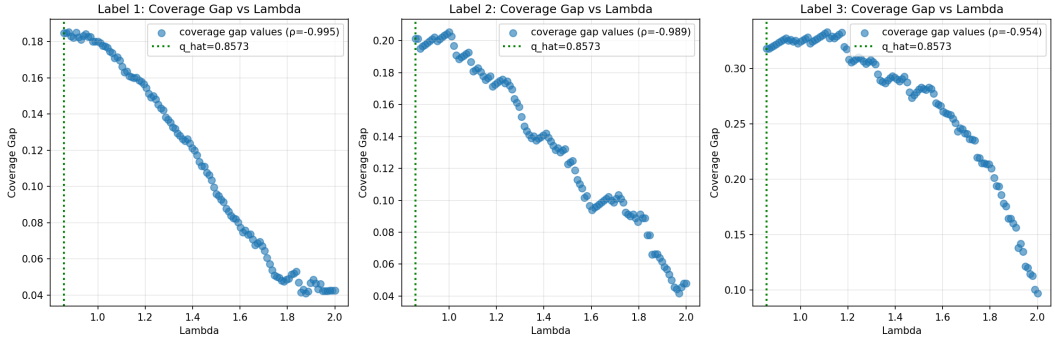



Figure C.1: **ACSIncome (small,** 4 **clients).** These plots show how the coverage gap (cg_curr) varies as we increase $\lambda$ starting from $\hat{q}$. As we can see, the relationship is nearly monotonic with $|\rho| > 0.95$ for all labels.

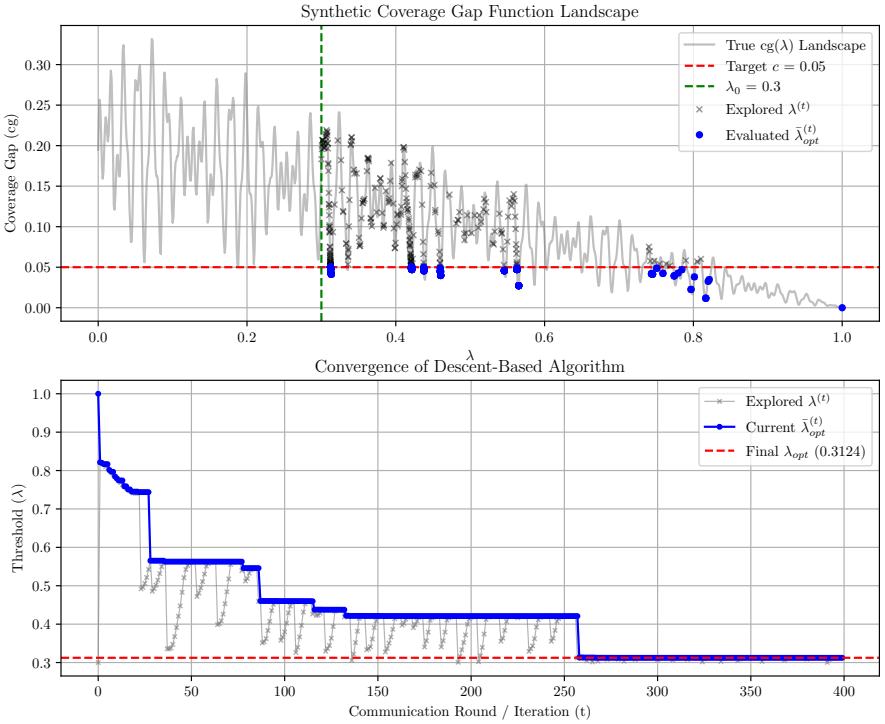



Figure C.2: We run Algorithm C.1, which includes *random restarts*–the mechanism we use to continue searching once a new $\bar{\lambda}_{\text{opt}}$ is found–on a synthetic, jagged function, representing an adversarial coverage gap function. Specifically, we use $\mu = 0.9$, $\eta = 0.1$. **Note:** we use $\text{cg}(\lambda)$ for notational brevity instead of $\text{cg}(\lambda, F_M, \tilde{y}, \mathcal{G})$ as $F_M, \tilde{y},$ and $\mathcal{G}$ are implicitly fixed in the synthetic function.

# D Proofs

## D.1 Proof of Theorem 3.1

Recall, since the data is distributed across clients in the federated setting, we reformulated the fairness-specific coverage level as Equation 10. In doing so, the computation of the coverage level is split between the clients and the server. We present bounds and point estimates for each of the terms in Equation 10 across Lemmas D.1, D.2, and D.3, leading to a proof of Theorem 3.1.

### D.1.1 Client-Side Estimates

Since each client operates independently with its own dataset, we can derive interval bounds for terms (I) and (II) assuming partial exchangeability. For the point estimates approach, we use maximum likelihood estimators (MLEs) for each term, providing the tightest estimates assuming partial IID.

**Lemma D.1.** *For each client $k$, group $g$, positive label $\tilde{y}$, and threshold $\lambda$, we get the following interval bounds:*

$$\frac{\alpha_k^{(g,\tilde{y});\lambda}}{n_k^{(g,\tilde{y})}+1} \leq \Pr[s(\boldsymbol{x}_{test}, \tilde{y}) \leq \lambda \mid F_M(\boldsymbol{x}_{test}, y_{test}, g, \tilde{y}) = 1, E_k] \leq \frac{\alpha_k^{(g,\tilde{y});\lambda}+1}{n_k^{(g,\tilde{y})}+1} \tag{15}$$

*If the data satisfy the partial-IID assumption, then we can use an MLE point estimate, given by the following:*

$$\Pr[s(\boldsymbol{x}_{test}, \tilde{y}) \leq \lambda \mid F_M(\boldsymbol{x}_{test}, y_{test}, g, \tilde{y}) = 1, E_k] = \frac{\alpha_k^{(g,\tilde{y});\lambda}}{n_k^{(g,\tilde{y})}} \tag{16}$$

The proof of Lemma D.1 is as follows:

*Proof.* We first observe that,

$$\begin{aligned}&\Pr[s(\boldsymbol{x}_{\text{test}}, \tilde{y}) \leq \lambda \mid F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, \tilde{y}) = 1,\ E_k] \\ &= \Pr_{\boldsymbol{x}_{\text{test}} \sim P_k}[s(\boldsymbol{x}_{\text{test}}, \tilde{y}) \leq \lambda \mid F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, \tilde{y}) = 1],\end{aligned}$$

since exchangeability with the elements in $k$ is true iff $\boldsymbol{x}_{\text{test}}$ is sampled from $k$'s local distribution, $P_k$. The interval bounds follow from the conditional coverage guarantees given in CF [Vadlamani et al., 2025].

For the point estimate, we can model the event that the predicted score $s(\boldsymbol{x}_{test}, \tilde{y})$ falls below $\lambda$ as a Bernoulli random variable with success probability $p$. We can treat the $n_k^{(g,\tilde{y})}$ calibration points as individual Bernoulli trials, to then construct a maximum likelihood estimate (MLE) for $p$, which will be $\hat{p} = \frac{\alpha_k^{(g,\tilde{y});\lambda}}{n_k^{(g,\tilde{y})}}$. □

Lemma D.1 bounds the fair-conditional coverage for a particular group-label pair for the test covariate $(\boldsymbol{x}_{\text{test}}, y_{\text{test}})$. We next bound the coverage of the test covariate satisfying the Fairness Metric ($F_M$), conditioned on the test point being exchangeable with data from client $k$ using Lemma D.2, and provide the proof below.

### D.1.2 Server-Side Estimates

Terms (III) and (IV) require a global view of the clients' data, so they are handled on the server.

For term (III), we follow the setup by Lu et al. [2023], where given $n_k = \left|\mathcal{D}_{\text{calib}}^{(k)}\right|$, $\gamma_k \coloneqq \Pr[E_k] \propto n_k + 1$ and $\sum_{k=1}^{K} \gamma_k = 1$. Finally, for term (IV), we have that

$$\Pr[F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, \tilde{y}) = 1] = \sum_{k=1}^{K} \Pr[F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, \tilde{y}) = 1 \mid E_k] \cdot \Pr[E_k].$$

Using Lemma D.2, we can get an interval-bound and point-estimate as shown in the following lemma.

**Lemma D.2.** *For each client $k$, group $g$, and positive label $\tilde{y}$, we get the following interval bounds:*

$$\frac{n_k^{(g,\tilde{y})}}{n_k+1} \leq \Pr[F_M(\boldsymbol{x}_{test}, y_{test}, g, \tilde{y}) = 1 \mid E_k] \leq \frac{n_k^{(g,\tilde{y})}+1}{n_k+1} \tag{17}$$

*If the data satisfy the partial-IID assumption, then we can use an MLE point estimate, given by the following:*

$$\Pr[F_M(\boldsymbol{x}_{test}, y_{test}, g, \tilde{y}) = 1 \mid E_k] = \frac{n_k^{(g,\tilde{y})}}{n_k} \tag{18}$$

*Proof.* To demonstrate the finite sample guarantee, we note that $F_M(\boldsymbol{x}, y, g, \tilde{y}) = 1$ for all $(\boldsymbol{x}, y) \in \mathcal{D}_{\text{calib}}^{(k)}$ are all exchangeable Bernoulli trials. Observe that conditioning on $E_k$ implies $\mathcal{D}_{\text{calib}_+}^{(k)} := \mathcal{D}_{\text{calib}}^{(k)} \cup \{\boldsymbol{x}_{\text{test}}\}$ is an exchangeable sequence of length $n_k + 1$. Treating this as a finite 'bag' of covariates, we have

$$\forall_{(\boldsymbol{x},y)\in\mathcal{D}_{\text{calib}_+}^{(k)}},\ \Pr[F_M(\boldsymbol{x}, y, g, \tilde{y}) = 1 \mid E_k] = \frac{\sum_{(\boldsymbol{x}_i,y_i)\in\mathcal{D}_{\text{calib}_+}^{(k)}} F_M(\boldsymbol{x_i}, y_i, g, \tilde{y})}{n_k + 1}.$$

In other words, we have defined the probability of randomly selecting a covariate with $F_M(\boldsymbol{x}, y, g, \tilde{y}) = 1$. Since this applies to all points we know,

$$\Pr[F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, \tilde{y}) = 1 \mid E_k] = \frac{\sum\limits_{(\boldsymbol{x}_i,y_i)\in\mathcal{D}_{\text{calib}_+}^{(k)}} F_M(\boldsymbol{x}_i, y_i, g, \tilde{y})}{n_k + 1}. \tag{19}$$

Since we implicitly condition on $F_M(\boldsymbol{x}, y, g, \tilde{y}),\ \forall(\boldsymbol{x}, y) \in k$ (effectively making them deterministic), we can calculate the following bounds,

$$\frac{\sum\limits_{(\boldsymbol{x}_i,y_i)\in\mathcal{D}_{\text{calib}}^{(k)}} F_M(\boldsymbol{x_i}, y_i, g, \tilde{y})}{n_k + 1} \le \Pr[F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, \tilde{y}) = 1 \mid E_k] \le \frac{\sum\limits_{(\boldsymbol{x}_i,y_i)\in\mathcal{D}_{\text{calib}}^{(k)}} F_M(\boldsymbol{x_i}, y_i, g, \tilde{y}) + 1}{n_k + 1}, \tag{20}$$

where the $+1$ term comes from the unknown value of $F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, \tilde{y})$. Substituting the sums with $n_k^{(g,\tilde{y})}$, proves the interval bounds.

For the point estimate, the event that $F_M = 1$ can be modeled as a Bernoulli random variable with success probability $p$. We can use the full $n_k$ calibration points as $n_k$ Bernoulli trials to construct an MLE for $p$, which will be $\hat{p} = \frac{n_k^{(g,\tilde{y})}}{n_k}$. □

Lastly, we use Lemma D.3, to bound $\Pr[F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, \tilde{y}) = 1]$. The proof of Lemma D.3 uses the result of Lemma D.2.

**Lemma D.3.** *For each client $k$, group $g$, and positive label $\tilde{y}$, we get the following interval bounds:*

$$L^{(g,\tilde{y})} = \sum_{k=1}^{K} \gamma_k \frac{n_k^{(g,\tilde{y})}}{n_k+1} \le \Pr[F_M(\boldsymbol{x}_{test}, y_{test}, g, \tilde{y}) = 1] \le \sum_{k=1}^{K} \gamma_k \frac{n_k^{(g,\tilde{y})}+1}{n_k+1} = U^{(g,\tilde{y})}. \tag{21}$$

*If the data satisfy the partial-IID assumption, then we can use an MLE point estimate, given by the following:*

$$\Pr[F_M(\boldsymbol{x}_{test}, y_{test}, g, \tilde{y}) = 1] = \sum_{k=1}^{K} \gamma_k \frac{n_k^{(g,\tilde{y})}}{n_k} = \pi^{(g,\tilde{y})}. \tag{22}$$

*Proof.* To achieve this result, we first use the law of total probability to decompose $\Pr[F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, \tilde{y}) = 1]$ into terms known by the server and the client:

$$\Pr[F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, \tilde{y}) = 1] = \sum_{k=1}^{K} \Pr[F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, \tilde{y}) = 1 \mid E_k] \cdot \Pr[E_k] \tag{23}$$

Then, substituting the bounds for term (II) (see Lemma D.2), and $\gamma_k = \Pr[E_k]$ from term (III) into Equation 23, we complete the proof. □

On closer inspection, we observe that Terms (I) and (II) can be combined and bound together to provide a tighter lower bound.

**Lemma D.4.** *Using the definitions of $\alpha_k^{(g,\tilde{y});\lambda}$ and $n_k + 1$ we have,*

$$\frac{\alpha_k^{(g,\tilde{y});\lambda}}{n_k+1} \leq \Pr\Big[s(\boldsymbol{x}_{test}, \tilde{y}) \leq \lambda \mid F_M(\boldsymbol{x}_{test}, y_{test}, g, \tilde{y}) = 1,\ \boldsymbol{x}_{test} \overset{exc.}{\sim} k\Big] \cdot \Pr\Big[F_M(\boldsymbol{x}_{test}, y_{test}, g, \tilde{y}) = 1 \mid \boldsymbol{x}_{test} \overset{exc.}{\sim} k\Big] \leq \frac{\alpha_k^{(g,\tilde{y});\lambda} + 1}{n_k + 1}. \tag{24}$$

*Proof.* First, observe that,

$$\Pr[s(\boldsymbol{x}_{\text{test}}, \tilde{y}) \leq \lambda \mid F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, \tilde{y}) = 1,\ E_k] \cdot \Pr[F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, \tilde{y}) = 1 \mid E_k]$$
$$= \Pr[s(\boldsymbol{x}_{\text{test}}, \tilde{y}) \leq \lambda,\ F_M(\boldsymbol{x}_{\text{test}}, y_{\text{test}}, g, \tilde{y}) = 1 \mid E_k]$$

Then consider the following Bernoulli random variables (R.V) $\mathbf{1}[s(\boldsymbol{x}, y) \leq \lambda] \cdot F_M(\boldsymbol{x}, y, g, \tilde{y})$ for all $(\boldsymbol{x}, y) \in k \cup \{(\boldsymbol{x}_{\text{test}}, y_{\text{test}})\}$ which form an exchangeable sequence (using the assumption $\boldsymbol{x}_{\text{test}} \overset{\text{exc.}}{\sim} k$). Additionally, observe $\alpha_k^{(g,\tilde{y});\lambda} = \sum_{(\boldsymbol{x}_i, y_i) \in k} \mathbf{1}[s(\boldsymbol{x}, y) \leq \lambda] \cdot F_M(\boldsymbol{x}, y, g, \tilde{y})$ is an equivalent definition of $\alpha_k^{(g,\tilde{y});\lambda}$. The rest of the proof follows from the proof of Lemma D.2 by using $\mathbf{1}[s(\boldsymbol{x}, y) \leq \lambda] \cdot F_M(\boldsymbol{x}, y, g, \tilde{y})$ as the Bernoulli R.V instead of $F_M(\boldsymbol{x}, y, g, \tilde{y})$ and $\alpha_k^{(g,\tilde{y});\lambda}$ in place of $n_k^{(g,\tilde{y})}$. □

Having proved the Lemmas, we can move on to proving Theorems 3.1 and 3.4:

**Theorem 3.1** (Interval Bounds)**.** *The fairness-specific coverage level, $\Pi_{cov}$, given by Equation 10, can be bounded as $L_{cov}(\lambda, F_M, g, \tilde{y}) \leq \Pi_{cov} \leq U_{cov}(\lambda, F_M, g, \tilde{y})$, where*

$$L_{cov}(\lambda, F_M, g, \tilde{y}) = \sum_{k=1}^{K} \frac{\gamma_k \alpha_k^{(g,\tilde{y});\lambda}}{(n_k+1)U^{(g,\tilde{y})}} \text{ and } U_{cov}(\lambda, F_M, g, \tilde{y}) = \sum_{k=1}^{K} \frac{\gamma_k (\alpha_k^{(g,\tilde{y});\lambda} + 1)}{(n_k+1)L^{(g,\tilde{y})}}. \tag{11}$$

*and $L^{(g,\tilde{y})} = \sum_{k=1}^{K} \gamma_k \frac{n_k^{(g,\tilde{y})}}{n_k+1}$ and $U^{(g,\tilde{y})} = \sum_{k=1}^{K} \gamma_k \frac{n_k^{(g,\tilde{y})}+1}{n_k+1}$*

*Proof.* Substituting the bounds for terms ① and ② via Lemma D.4 and ④ which were established via Lemmas D.2 and D.3, respectively, and the definition of ③ into Equation 10 completes the proof. □

**Theorem 3.4** (Point (MLE) Estimate)**.** *If we assume partial-IID, using MLE estimates for each term, we get the following estimate for the fairness-specific coverage level, $\Pi_{cov} = \sum_{k=1}^{K} \frac{\gamma_k \alpha_k^{(g,\tilde{y});\lambda}}{n_k \pi^{(g,\tilde{y})}}$.*

*Proof.* Substituting the estimates for terms ① and ② via Lemma D.4 and ④ which were established via Lemmas D.2 and D.3, respectively, and the definition of ③ into Equation 10 completes the proof. □

### D.2 Sufficient Data Counts

Below, we consider the sufficient number of data points to get a non-degenerate solution (i.e., $\lambda_{\text{opt};y} = \lambda_{\text{max}}$).

**Theorem 3.2.** *Given a closeness criterion, $c$ and $\gamma_k = \frac{n_k+1}{N+K}$, in order to have non-vacuous prediction sets (i.e., $\lambda < \lambda_{max}$), we need $N^{(g,\tilde{y})} := \sum_{k=1}^{K} n_k^{(g,\tilde{y})} \geq \frac{2K}{c}$ for all $(g, \tilde{y}) \in \mathcal{G} \times \mathcal{Y}^+$.*

*Proof.* Define the interval width for a particular group label pair as,

$$\text{IW}(\lambda, F_M, g, \tilde{y}) \coloneqq U_{\text{cov}}(\lambda, F_M, g, \tilde{y}) - L_{\text{cov}}(\lambda, F_M, g, \tilde{y}) \tag{25}$$
$$= \sum_{k=1}^{K} \gamma_k \left( \frac{(\alpha_k^{(g,\tilde{y});\lambda} + 1)}{(n_k+1)L^{(g,\tilde{y})}} - \frac{\alpha_k^{(g,\tilde{y});\lambda}}{(n_k+1)U^{(g,\tilde{y})}} \right), \tag{26}$$

where $U_{\text{cov}}$ and $L_{\text{cov}}$ are from Theorem 3.1. We note that we can achieve non-vacuous bounds for each $\tilde{y} \in \mathcal{Y}^+$ if for each $g \in \mathcal{G}$, $\text{IW}(\lambda, F_M, g, \tilde{y}) < c$ for some $\lambda < \lambda_{max}$. Observe that if the interval

width is greater than $c$ for all $\lambda$, then our algorithm cannot find non-vacuous bounds. Also, recall by definition, $L^{(g,\tilde{y})} = N^{(g,\tilde{y})}$ and $U^{(g,\tilde{y})} = N^{(g,\tilde{y})} + K$ Then consider,

$$\text{IW}(\lambda, F_M, g, \tilde{y}) = \sum_{k=1}^{K} \gamma_k \left( \frac{(\alpha_k^{(g,\tilde{y});\lambda} + 1)}{(n_k+1)L^{(g,\tilde{y})}} - \frac{\alpha_k^{(g,\tilde{y});\lambda}}{(n_k+1)U^{(g,\tilde{y})}} \right) \tag{27}$$

$$= \frac{1}{N+K} \sum_{k=1}^{K} \left( \frac{(\alpha_k^{(g,\tilde{y});\lambda} + 1)}{L^{(g,\tilde{y})}} - \frac{\alpha_k^{(g,\tilde{y});\lambda}}{U^{(g,\tilde{y})}} \right) \tag{28}$$

$$= \sum_{k=1}^{K} \left( \frac{(\alpha_k^{(g,\tilde{y});\lambda} + 1)}{N^{(g,\tilde{y})}} - \frac{\alpha_k^{(g,\tilde{y});\lambda}}{N^{(g,\tilde{y})} + K} \right) \tag{29}$$

Given $\sum_{k=1}^{K} \alpha_k^{(g,\tilde{y})} \leq \sum_{k=1}^{K} n_k^{(g,\tilde{y})} = N^{(g,\tilde{y})}$, we have

$$\left( \frac{N^{(g,\tilde{y})} - \sum_{k=1}^{K} \alpha_k^{(g,\tilde{y});\lambda}}{N^{(g,\tilde{y})}} - \frac{N^{(g,\tilde{y})} - \sum_{k=1}^{K} \alpha_k^{(g,\tilde{y});\lambda}}{N^{(g,\tilde{y})} + K} \right) \geq 0. \tag{30}$$

Thus,

$$(29) \leq \left[ \sum_{k=1}^{K} \left( \frac{(\alpha_k^{(g,\tilde{y});\lambda}+1)}{N^{(g,\tilde{y})}} - \frac{\alpha_k^{(g,\tilde{y});\lambda}}{N^{(g,\tilde{y})}+K} \right) + \left( \frac{N^{(g,\tilde{y})} - \sum_{k=1}^{K} \alpha_k^{(g,\tilde{y});\lambda}}{N^{(g,\tilde{y})}} - \frac{N^{(g,\tilde{y})} - \sum_{k=1}^{K} \alpha_k^{(g,\tilde{y});\lambda}}{N^{(g,\tilde{y})}+K} \right) \right] \tag{31}$$

$$= \left( \frac{N^{(g,\tilde{y})} + K}{N^{(g,\tilde{y})}} - \frac{N^{(g,\tilde{y})}}{N^{(g,\tilde{y})} + K} \right) \tag{32}$$

$$= \frac{2N^{(g,\tilde{y})}K + K^2}{{N^{(g,\tilde{y})}}^2 + N^{(g,\tilde{y})}K} \leq c \tag{33}$$

Thus, we have the interval width, and we want to bound it by $c$ as in (33) $< c$. Solving for the quadratic form and taking the positive root, we arrive at,

$$N^{(g,\tilde{y})} \geq \frac{K}{2c} \left( 2 - c(N+K) + \sqrt{c^2(N+K)^2 + 4} \right) \tag{34}$$

While (34) is a sufficient quantity for the minimal count, a tight overbound of (34) (for more interpretable bounds) follows as,

$$\frac{K}{2c} \left( 2 - c(N+K) + \sqrt{c^2(N+K)^2 + 4} \right) \tag{35}$$

$$\leq \frac{K}{2c} \left( 2 - c(N+K) + \sqrt{c^2(N+K)^2 + 4c(N+K) + 4} \right) \tag{36}$$

$$= \frac{K}{2c} \left( 2 - c(N+K) + 2 + c(N+K) \right) \tag{37}$$

$$= \frac{2K}{c} \leq N^{(g,\tilde{y})} \tag{38}$$

□

### D.3 Tightness of Theorem 3.1

To characterize the tightness of Theorem 3.1, we can consider the interval width (Equation (25)) presented in Theorem 3.2. In our analysis, we assume the number of data points remains the same in the centralized setting. Then, for notational simplicity define,

$$P := \sum_{k=1}^{K} \alpha_k^{(g,\tilde{y});\lambda} \tag{39}$$

$$N := N^{(g,\tilde{y})} \tag{40}$$

Then, from the interval width expansion in Equation 29 we substitute Equations 39-40,

$$\text{IW}_K = \frac{P+K}{N} - \frac{P}{N+K}, \tag{41}$$

where the interval width (IW) is indexed by $K$ as $P$ and $N$ are fixed. Next, to quantify the tightness of the bound we compute the ratio of the $\text{IW}_K$ over $\text{IW}_1$ (i.e., compared to the centralized setting where $K = 1$). Then,

$$\frac{\text{IW}_K}{\text{IW}_1} = \left(\frac{P+K}{N} - \frac{P}{N+K}\right) / \left(\frac{P+1}{N} - \frac{P}{N+1}\right) \tag{42}$$

$$= \frac{K(N+1)(P+N+K)}{(N+K)(P+N+1)} \tag{43}$$

$$\leq K, \tag{44}$$

where the last step arises from recognizing $P \geq 0$ and $N > 0$. Hence the tightness of Theorem 3.1 (interval width) increases by less than a factor $K$, the number of clients.

# E Additional Experiment Details

## E.1 Datasets

We present a summary of common dataset statistics in Table E.1 and go into more details on each dataset in the following sections. Additionally, the Folktables and Fitzpatrick datasets are licensed under CC BY 4.0. The Pokec dataset in the SNAP repository is licensed under the BSD license.

Table E.1: Dataset Statistics. T refers to Tabular, G refers to Graph, and V refers to vision.
*ACS datasets have six (6) groups if using the *continental* split schemes (see Section E.2).
ˆ Number of inputs after removing those with unknown group information

| Name | Type | Size | # Labeled | # Groups | # Classes |
|---|---|---|---|---|---|
| ACSIncome | T | $1,664,500$ | ALL | race$(9)^*$ | 4 |
| ACSEducation | T | $1,664,500$ | ALL | race$(9)^*$ | 6 |
| Fitzpatrick | V | $16,012$ˆ | ALL | skin type$(6)$ | 9 |
| **Name** | **Type** | $(\lvert\mathcal{V}\rvert, \lvert\mathcal{E}\rvert)$ | **# Labeled** | **# Groups** | **# Classes** |
| Pokec-{n, z} | G | $(133,138,\ 1,458,258)$ | $17,594$ | region$(2)$, gender$(2)$ | 4 |

## E.2 Folktables Datasets

In the fairness space, the American Community Services (ACS) datasets from the `Folktables` library are a widely used set of tabular data [Ding et al., 2021]. The data is taken across the 51 U.S. states and territories. For our federated setup and each dataset below, we consider the following 6 partitioning schemes:

(1.) **All:** We consider each U.S. state and territory to be its own client

(2.) **Large:** We follow the Bureau of Economic Analysis's division of the U.S. into New England, the Mideast, the Great Lakes, the Plains, the Southeast, the Southwest, the Rocky Mountain, and the Far West.

(3.) **Small:** We follow the U.S. Census Bureau's division of the U.S. into the Northeast, the Midwest, the South, and the West

(4-6.) **Continental All, Continental Large, Continental Small:** The same as 1 to 3, but we only consider the *continental* U.S.–removing Alaska, Hawaii, and Puerto Rico.

All Folktable datasets have a race attribute. When we partition the data using all the states and territories, we use the full version of race, which has 9 groups. However, when partitioning just with *continental* U.S., we combine some demographic groups–primarily those from Alaska, Hawaii, and Puerto Rico–into the appropriate 'Other' categories, resulting in a total of 6 groups.

**ACSIncome:** We used the standard ACSIncome dataset from Folktables; however, we divided the targets into four classes by evenly splitting the income into 4 brackets. The sensitive attribute in this case is race, resulting in either 9 or 6 groups.

**ACSEducation:** This is a custom dataset. We used the ACSTravelTime data and selected Education Level as the target. The education level was divided into 6 groups: {did not complete high school, has a high school diploma, has a GED, started an undergrad program, completed an undergrad program, and completed graduate or professional school}. ACSEducation also uses race as a sensitive attribute.

### E.3 Non-Tabular Datasets

**Pokec-{n,z}:** The Pokec-{n, z} dataset [Takac and Zabovsky, 2012] is a social network graph dataset collected from Pokec, a popular social network in Slovakia. Since several rows in the dataset are missing features, two commonly used subgraphs are the Pokec-z and Pokec-n datasets. The graphs have four labels corresponding to the fieldwork and two sensitive attributes: gender (2 groups) and region (2 groups). Our experiments consider each attribute individually as well as intersectional fairness by creating an attribute with $4$ groups. For our federated setup, we use each subgraph as a single client, resulting in 2 clients.

**Fitzpatrick:** The Fitzpatrick dataset [Groh et al., 2021] contains clinical images classified based on the depicted skin condition. There are several levels of granularity regarding the skin condition label. We use a version with 9 skin conditions: {inflammatory, malignant epidermal, genodermatoses, benign dermal, benign epidermal, malignant melanoma, benign melanocyte, malignant cutaneous lymphoma, malignant dermal}. There are 6 demographic groups based on the Fitzpatrick skin type. For our federated setup, we use a Dirichlet partitioner to split the data into $K \in \{2, 4, 8\}$ clients.

### E.4 Hyperparameters and Implementation

To promote reproducibility, the source code for FedCF is provided in the supplementary material, along with the configuration files containing the hyperparameters used. All experiments were conducted on at least three seeds. The error bars were all too small to plot.

The project was written using the Flower AI Federated Learning framework [Beutel et al., 2020] for both base model training and the FedCF framework.

**Compute Resources:** Model training with the ACS and Pokec datasets was run on 1 A6000 GPU, and FedCF was run on 8 AMD EPYC 7313 CPUs. For the Fitzpatrick experiments, model training was done on 1 V100 GPU, and FedCF was run on 8 Dual Intel Xeon 8268 CPUs.

### E.5 Non-Conformity Scores

**Adaptive Prediction Sets (APS)** The most popular CP method for classification problems is APS [Romano et al., 2020c]. The scoring function first sorts the softmax logits in descending order and accumulates the class probabilities until the correct class is included. For tighter prediction sets, randomization is introduced through a uniform random variable.

Formally, let $\hat{\pi}$ be a trained classification model with softmaxed output. If $\hat{\pi}(\boldsymbol{x})_{(1)} \geq \hat{\pi}(\boldsymbol{x})_{(2)} \geq \cdots \geq \hat{\pi}(\boldsymbol{x})_{(C)}$, $u \sim U(0, 1)$, and $r_y$ is the rank of the correct label, then

$$s(\boldsymbol{x}, y) = \left[ \sum_{i=1}^{r_y} \hat{\pi}(\boldsymbol{x})_{(i)} \right] - u\hat{\pi}(\boldsymbol{x})_y.$$

APS has two major drawbacks that have led to it being surpassed by other methods in recent CP literature. First, APS tends to produce large (less efficient) prediction sets. Second, it does not account for structure in its formulation. To address these issues, alternatives like RAPS and DAPS have emerged[4].

**Regularized Adaptive Prediction Sets (RAPS)** Angelopoulos et al. [2022] introduces a regularization approach for APS. Given the same setup and notation as APS, define $o(\boldsymbol{x}, y) = |\{c \in \mathcal{Y} : \hat{\pi}(\boldsymbol{x})_c \geq \hat{\pi}(\boldsymbol{x})_y\}|$. Then,

[4]RAPS and DAPS have hyperparameters typically tuned on separate held-out data, but we fix them *a priori* to preserve data for calibration and evaluation as well as to be consistent with what prior federated conformal prediction works have done.

$$s(\boldsymbol{x}, y) = \left[\sum_{i=1}^{r_y} \hat{\pi}(\boldsymbol{x})_{(i)}\right] - u\hat{\pi}(\boldsymbol{x})_y + \nu \cdot \max\{(o(\boldsymbol{x}, y) - k_{reg}), 0\},$$

where $\nu$ and $k_{reg} \geq 0$ are regularization hyperparameters.

**Diffusion Adaptive Prediction Sets (DAPS)** Graphs are rich with neighborhood information, with nodes often exhibiting homophily. This suggests that the non-conformity scores of connected nodes are likely to be related. To leverage this insight, DAPS H. Zargarbashi et al. [2023] incorporates a one-step diffusion update on the non-conformity scores. Formally, if $s(\boldsymbol{x}, y)$ is a point-wise score function (e.g., APS), then the diffusion step yields a new score function

$$\hat{s}(\boldsymbol{v}, \boldsymbol{x}_v, y) = (1 - \delta)s(\boldsymbol{x}_v, y) + \frac{\delta}{|\mathcal{N}_{\boldsymbol{v}}|} \sum_{\boldsymbol{u} \in \mathcal{N}_{\boldsymbol{v}}} s(\boldsymbol{x}_u, y),$$

where $\delta \in [0, 1]$ is a diffusion hyperparamter and $\mathcal{N}_{\boldsymbol{v}}$ is the 1-hop neighborhood of $\boldsymbol{v}$.

# F On Using Group Information

**(Group-Classwise)-Formulation** A user can alter FedCF to use group information, such that $\lambda_{\text{opt}} \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{G}|}$. Recall from Section 3.2.1, for a given label $y$ the threshold vector is $\lambda_{\text{opt};(y)} \in \mathbb{R}^{|\mathcal{G}|}$ and $\Lambda^{|\mathcal{G}|}$ defined as $|\mathcal{G}|$ cartesian products of $\Lambda$, the problem becomes,

$$\min_{\lambda \in \Lambda^{|\mathcal{G}|}} f(\lambda) \quad \text{subject to} \quad \text{cg}(\lambda, F_M, \tilde{y}, \mathcal{G}) - c \leq 0, \tag{45}$$

where $f$ is the scalar objective, and we redfine cg as,

$$\text{cg}(\lambda, F_M, \tilde{y}, \mathcal{G}) := \max_{g_a \in \mathcal{G}} \{U_{\text{cov}}(\lambda_{(g_a)}, F_M, g_a, \tilde{y})\} - \min_{g_b \in \mathcal{G}} \{L_{\text{cov}}(\lambda_{(g_b)}, F_M, g_b, \tilde{y})\}. \tag{46}$$

to accommodate different thresholds for each group, such that $\lambda_{(g)}$ corresponds to the threshold for group $g$ for the fixed label $\tilde{y}$. For the results presented in Table 3, approximately solve problem (45) when using $f(\lambda) = \lambda^\top 1$–the element-wise sum of the $\lambda$ vector (this is discussed below in Section 3.2.1).

We first note that the feasible set for the classwise formulation (Equation 7) is a subset of the feasible set for the group-classwise formulation in Equation 45. We then further assume $f$ is elementwise monotone and separable, e.g. $f(\lambda) := f_{g_1}(\lambda_{g_1}) + \cdots + f_{g_{|\mathcal{G}|}}(\lambda_{g_{|\mathcal{G}|}})$ for $g_i \in \mathcal{G}$ where $f_{g_i}$ is a non-decreasing function($f(\lambda) = \lambda^\top 1$ trivially satisfies this assumption). Under this assumption, we observe that an optimizer for the classwise formulation will have an objective cost greater than or equal to the group-classwise formulation. Formally,

**Lemma F.1.** *For a fixed label $\tilde{y}$, fairness metric $F_m$, and set of groups $\mathcal{G}$, let $\lambda^L_{opt} \in \mathbb{R}$ be the optimizer of the classwise formulation in Equation 7 and $\lambda^G_{opt} \in \mathbb{R}^{|G|}$. Then,*

$$f(\lambda^G_{opt}) \leq f(\underbrace{[\lambda^L_{opt}, \ldots, \lambda^L_{opt}]}_{|\mathcal{G}|}) \tag{47}$$

*Proof.* By definition, the optimizer of the objective $f$, $\lambda^G_{\text{opt}}$, must have a smaller objective value than any other feasible solution, i.e., $\underbrace{[\lambda^L_{\text{opt}}, \ldots, \lambda^L_{\text{opt}}]}_{|\mathcal{G}|}$. □

**Remark F.2.** Lemma F.1 does not necessarily imply $\lambda^G_{\text{opt};(g)} \leq \lambda^L_{\text{opt}}$ (where $\lambda^G_{\text{opt};(g)}$ is the threshold for class $g$). However, for the class of problems we consider, this is a simplifying assumption that is often true since $\lambda^L_{\text{opt}}$ is increased to cover the worst-case group, $g$. Thus, we instead solve the following problem,

$$\min_{\lambda \in \Lambda^{|\mathcal{G}|}} f(\lambda) \tag{48a}$$

$$\text{s.t} \quad \text{cg}(\lambda, F_M, \tilde{y}, \mathcal{G}) - c \leq 0 \tag{48b}$$

$$\lambda_{(g)} \leq \lambda^L_{\text{opt}} \quad \forall g \in \mathcal{G}. \tag{48c}$$

To solve the problem in Equation 48 we first run the FedCF optimization (Algorithm C.1) described in to solve $\lambda^L_{\text{opt}}$, for $N$ rounds (num_rounds in the Algorithm). Following this, we run the Algorithm for $R$ iterations, where we admit a new solution as optimal if $\lambda^{(t+1)}_{(g)} \leq \lambda^{(t)}_{(g)}\ \forall g \in \mathcal{G}$ (using the element-wise montone property of $f$) and $\lambda$ satisfies the coverage gap requirement. While this leads to local minima, we use random restarts to escape them and find a better solution. By taking the described approach, we can more efficiently search in a smaller space of $\Lambda$ for the first $N$ steps and then perform $M$ searches in the larger space $\Lambda^{|\mathcal{G}|}$. For our implementation, we set $N$ to 50, $R$ to 50, and recall that $f(\lambda) = \lambda^\top 1$.

### F.1 Quantifying the Efficiency Improvement

In the following section, we bound the efficiency gap between the classwise and (group, class)-wise formulation when solving the problem in Equation 48:

**Theorem F.3.** *Let $\lambda^L \in \mathbb{R}^{|\mathcal{Y}|}$ and $\lambda^G \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{G}|}$ be optimizers for the class-wise and (group, class)-wise formulations of (Fed)CF, respectively. Then let $x_{test}$ represent an unseen covariate, and $\boldsymbol{g}(x_{test})$ map the covariate to its group. Lastly, let $J_{(y,g)}(\lambda) = \Pr[s(x_{test}, y) \leq \lambda | \boldsymbol{g}(x_{test}) = g]$ be the CDF function of the conditional score distribution and Lipschitz continuous, with Lipschitz constant $M_{(y,g)}$ over the interval $[\lambda_0, \lambda_{max}]$—that is for all $v, w \in [\lambda_0, \lambda_{max}]$, $|J_{(y,g)}(v) - J_{(y,g)}(w)| \leq M_{(y,g)}|v - w|$. Then, the efficiency gap between the two (Fed)CF formulations, $\Delta_{\text{eff}}$, satisfies,*

$$\Delta_{\text{eff}} \leq \sum_{y\in\mathcal{Y}^+}\sum_{g\in\mathcal{G}} M_{(y,g)} \left|(\lambda^L_{(y)} - \lambda^G_{(y,g)})\right| \Pr[\boldsymbol{g}(x_{test}) = g]. \tag{49}$$

*Proof.* Observe from the analysis of [Maneriker et al., 2025], the efficiency difference of two prediction sets is quantified as the difference in likelihood of a non-conformity score being less than the threshold for each label. That is,

$$\Delta_{\text{eff}} = \left| \mathop{\mathbb{E}}_{x_{\text{test}}\sim\mathcal{X}} \left[ \sum_{y\in\mathcal{Y}} \left( \mathbf{1}[s(x_{\text{test}}, y) \leq \lambda^L_{(y)}] - \mathbf{1}[s(x_{\text{test}}, y) \leq \lambda^G_{(y,\boldsymbol{g}_{\text{test}})}] \right) \right] \right|, \tag{50}$$

where $\lambda^L_{(y)}$ is the component for class $y$ in $\lambda^L$ and $\lambda^G_{y,g}$ is the component for $(y, g)$ in $\lambda^G$. For notational brevity, let $\boldsymbol{g}_{\text{test}} \coloneqq \boldsymbol{g}(x_{\text{test}})$. Then,

$$\begin{aligned}
\Delta_{\text{eff}} &= \left| \sum_{y\in\mathcal{Y}} \left( \Pr[s(x_{\text{test}}, y) \leq \lambda^L_{(y)}] - \Pr[s(x_{\text{test}}, y) \leq \lambda^G_{(y,\boldsymbol{g}_{\text{test}})}] \right) \right| && (51)\\
&= \left| \sum_{y\in\mathcal{Y}}\sum_{g\in\mathcal{G}} \left( \Pr[s(x_{\text{test}}, y) \leq \lambda^L_{(y)} \mid \boldsymbol{g}_{\text{test}} = g] - \Pr[s(x_{\text{test}}, y) \leq \lambda^G_{(y,\boldsymbol{g}_{\text{test}})} \mid \boldsymbol{g}_{\text{test}} = g] \right) \Pr[\boldsymbol{g}_{\text{test}} = g] \right| && (52)\\
&= \left| \sum_{y\in\mathcal{Y}}\sum_{g\in\mathcal{G}} \left( \Pr[s(x_{\text{test}}, y) \leq \lambda^L_{(y)} \mid \boldsymbol{g}_{\text{test}} = g] - \Pr[s(x_{\text{test}}, y) \leq \lambda^G_{(y,\boldsymbol{g}_{\text{test}})} \mid \boldsymbol{g}_{\text{test}} = g] \right) \Pr[\boldsymbol{g}_{\text{test}} = g] \right| && (53)\\
&= \left| \sum_{y\in\mathcal{Y}}\sum_{g\in\mathcal{G}} \left( J_{(y,g)}(\lambda^L_{(y)}) - J_{(y,g)}(\lambda^G_{(y,g)}) \right) \Pr[\boldsymbol{g}_{\text{test}} = g] \right| && \text{(Definition of CDF)} \quad (54)\\
&\leq \sum_{y\in\mathcal{Y}}\sum_{g\in\mathcal{G}} \left| J_{(y,g)}(\lambda^L_{(y)}) - J_{(y,g)}(\lambda^G_{(y,g)}) \right| \Pr[\boldsymbol{g}_{\text{test}} = g] && \text{(Triangle Inequality)} \quad (55)\\
&\leq \sum_{y\in\mathcal{Y}}\sum_{g\in\mathcal{G}} M_{(y,g)} \left| (\lambda^L_{(y)} - \lambda^G_{(y,g)}) \right| \Pr[\boldsymbol{g}_{\text{test}} = g]. && \text{(Lipschitz Continuity of } J_{(y,g)}) \quad (56)
\end{aligned}$$

Lastly, we note that since $\lambda^L$ and $\lambda^G$ are optimizers to their respective (Fed)CF formulation, for all $y \notin \mathcal{Y}^+$, $\lambda^L_{(y)} = \lambda^G_{(y,\boldsymbol{g}_{\text{test}})} = \lambda_0$. Thus, the bound reduces to,

$$\Delta_{\text{eff}} \leq \sum_{y\in\mathcal{Y}^+}\sum_{g\in\mathcal{G}} M_{(y,g)} \left| (\lambda^L_{(y)} - \lambda^G_{(y,g)}) \right| \Pr[\boldsymbol{g}_{\text{test}} = g] \tag{57}$$

□

# G FedCF vs FairFed + FCP

FedCF takes a (black-box) federated machine learning model and applies the ideas from Conformal Fairness to produce fair federated conformal predictors; however, it is natural to ask whether using an underlying fair ML model would suffice. We conduct an experiment comparing FedCF vs FCP with FairFed [Ezzeldin et al., 2023]. We report the worst-case fairness violations using Fitzpatrick with 8 clients for Demographic Parity. For consistency, we use a Demographic Parity-based weight-scheme for FairFed and set $\beta \in \{0.5, 2.0\}$ in Table G.1, where $\beta$ is a parameter that controls the fairness budget in the weight update. If $\beta = 0$, then FairFed is equivalent to FedAvg. We observe that using FCP with FairFed is insufficient to achieve the desired fairness guarantee, demonstrating the utility and need for FedCF.

Table G.1: **Fitzpatrick, 8 clients, RAPS** We report the **fairness disparity** for different $c$ values and observe that simply using a fair base model is insufficient for controlling the fairness-specific coverage gap between groups.

| $\beta$ | $c = 0.1$ | | $c = 0.15$ | | $c = 0.2$ | |
|---|---|---|---|---|---|---|
| | FCP | FedCF | FCP | FedCF | FCP | FedCF |
| 0.5 | 0.302 | **0.045** | 0.302 | **0.140** | 0.303 | **0.197** |
| 2.0 | 0.247 | **0.024** | 0.247 | **0.122** | 0.247 | **0.122** |

# H FedCF with Enhanced Privacy

Preserving data privacy is a fundamental pillar of FL mechanisms, as they typically interact with sensitive client data. In this vein, we formulate an *enhanced privacy* version of FedCF.

## H.1 Enhanced Privacy

To better preserve privacy (compared to the *communication efficient* approach), we can offload more of the computation to the client-side, making it harder for the server-side to reverse-engineer or infer distributional information from the sent quantities. Expanding Equation 13, we get

$$U_{\text{cov}}(\lambda, F_M, g_a, \tilde{y}) - L_{\text{cov}}(\lambda, F_M, g_b, \tilde{y}) = \sum_{k=1}^{K} \gamma_k \underbrace{\left\{ \frac{(\alpha_k^{(g_a,\tilde{y});\lambda} + 1)}{(n_k+1)L^{(g_a,\tilde{y})}} - \frac{\alpha_k^{(g_b,\tilde{y});\lambda}}{(n_k+1)U^{(g_b,\tilde{y})}} \right\}}_{\text{Returned by the Client}}. \tag{58}$$

In this formulation, the client sends back the summand for each group, positive label pair except for $g_a \neq g_b$, making the space complexity of the client's message $\mathcal{O}\big(|\mathcal{G}|^2|\mathcal{Y}^+|\big)$–quadratic with respect to the number of groups and linear with respect to positive labels. Observe that for $g_a = g_b$, we are computing the interval width (Equation 25) for which we have a tighter overbound from the left hand side of Equation 33 which can be computed using the prior terms sent to compute $L^{(g,\tilde{y})}$ and $U^{(g,\tilde{y})}$. This allows us to mitigate reconstruction attacks, which would enable one to recover $\alpha_k^{(g_a,\tilde{y})}$ if the pairwise gap between $g_a$ and $g_b$ were sent. While this can lead to an overapproximation of the coverage gap, the bounds in Equation 35 establish when this estimate is less than $c$ (i.e., $\frac{2K}{c} \leq N^{(g,\tilde{y})}$), thus not hurting our approaches' formulation to find the smallest valid $\lambda_{\text{opt};(y)}$ for each label $y$. As with the *communication-efficient* protocol, the total computational time complexity for the *ehanced privacy* over $T$ server rounds is $\mathcal{O}\Big(T|\mathcal{Y}^+|\Big(K + \max_k \sum_{g\in\mathcal{G}} n_k^{(g,\tilde{y})}\Big)\Big)$.

The data privacy improves with this approach compared to the *communication efficient* version, since the data sent to the server is the difference of client-level summary statistics, which obfuscates individual distribution information from the server. However, unlike the *communication efficient* approach, the upper-coverage term ($U_{\text{cov}}$) is not separable from the aggregated sum, thus preventing us from enforcing $U_{\text{cov}} < 1$. In limited data settings, this results in more conservative coverage gap estimates, which increases the prediction set size when using the *enhanced privacy* approach.

**Unknown $\gamma_k$.** In our work, we take $\gamma_k \propto (n_k + 1)$; however, if $\gamma_k$ is unknown, we can still apply FedCF using the Enhanced Privacy variant just discussed. This is because we can factor our $\gamma_k$ in Equation 58 and apply the analysis done in Lu et al. [2023, Appendix B].

Below we provide a side-by-side comparison of the *communication efficient* and *enhanced privacy* algorithms for computing the federated coverage gap on the client side (Figure H.1). A unified server-side algorithm can be found in Algorithm H.3.

Algorithm H.1 More Communication Efficient Client-Side Computation for Coverage Gap

**function** CLIENTCG_COMM_EFFICIENT($k$, $\lambda$, $F_M$, $\tilde{y}$, $\mathcal{G}$)
  $l_k = [0]_{\mathcal{G}}$
  $u_k = [0]_{\mathcal{G}}$
  **for** $g \in \mathcal{G}$ **do**
    **if** `use_mle` **then**
      $l_k[g] \leftarrow \frac{\alpha_k^{(g,\tilde{y});\lambda}}{n_k}$
      $u_k[g] \leftarrow \frac{\alpha_k^{(g,\tilde{y});\lambda}}{n_k}$
    **else**
      $l_k[g] \leftarrow \frac{\alpha_k^{(g,\tilde{y});\lambda}}{((n_k+1))}$
      $u_k[g] \leftarrow \frac{\alpha_k^{(g,\tilde{y});\lambda}+1}{(n_k+1)}$
    **end if**
  **end for**
  **return** $l_k, u_k, n_k$
**end function**

Algorithm H.2 Client-Side Computation for Coverage Gap with Enhanced Privacy

**function** CLIENTCG_PRIVATE($k$, $\lambda$, $F_M$, $\tilde{y}$, $\mathcal{G}$)
  $l_k = [0]_{\mathcal{G}}$
  $u_k = [0]_{\mathcal{G}}$
  **for** $g \in \mathcal{G}$ **do**
    **if** `use_mle` **then**
      `global` $\pi^{(g,\tilde{y})}$
      $l_k[g] \leftarrow \frac{\alpha_k^{(g,\tilde{y});\lambda}}{\left(n_k \cdot \pi^{(g,\tilde{y})}\right)}$
      $u_k[g] \leftarrow \frac{\alpha_k^{(g,\tilde{y});\lambda}}{\left(n_k \cdot \pi^{(g,\tilde{y})}\right)}$
    **else**
      `global` $L^{(g,\tilde{y})}$
      `global` $U^{(g,\tilde{y})}$
      $l_k[g] \leftarrow \frac{\alpha_k^{(g,\tilde{y});\lambda}}{\left(\cdot(n_k+1) \cdot U^{(g,\tilde{y})}\right)}$
      $u_k[g] \leftarrow \frac{\alpha_k^{(g,\tilde{y});\lambda}+1}{\left((n_k+1) \cdot L^{(g,\tilde{y})}\right)}$
    **end if**
  **end for**
  $pw_cg_k = [0]_{\mathcal{G}\times\mathcal{G}}$
  // Pairwise coverage gap
  **for** $(g_a, g_b) \in \mathcal{G} \times \mathcal{G}$ **do**
    $pw_cg_k[g_a, g_b] \leftarrow u_k[g_a] - l_k[g_b]$
  **end for**
  **return** $pw_cg_k, n_k$
**end function**

Figure H.1: **Pseudocode for the two client-side protocols to compute the coverage gap.** The *enhanced privacy* version (on the right) includes the pairwise computation step, which results in a larger space complexity compared to the more *communication efficient* version (on the left). The `global` is used to convey that the priors are precomputed and available when needed.

**Algorithm H.3** Full Server-side Aggregation for Coverage Gap

**function** SERVERCG($\lambda_0, F_M, \tilde{y}, \mathcal{G}$, `formulations`)
  $n_list = [0]_{\mathcal{K}}$
  $l_list = [0]_{\mathcal{K}\times\mathcal{G}}$, $u_list = [0]_{\mathcal{K}\times\mathcal{G}}$ // Used for comm. efficient formulations
  $pw_cg_list = [0]_{\mathcal{K}\times\mathcal{G}\times\mathcal{G}}$ // Used for private formulations
  **for** client $k \in \mathcal{K}$ **in parallel do**
    **if** `formulations` == `COMM_EFFICIENT` **then**
      Receive $(l_k, u_k, n_k) =$ CLIENTCG_COMM_EFFICIENT$(k, \lambda_0, F_M, \tilde{y}, \mathcal{G})$
      $l_list[k] \leftarrow l_k$, $u_list[k] \leftarrow u_k$
    **else**
      Receive $(pw_cg_k, n_k) =$ CLIENTCG_PRIVATE$(k, \lambda_0, F_M, \tilde{y}, \mathcal{G})$
      $pw_cg_list[k] \leftarrow pw_cg_k$
    **end if**
    $n_list[k] \leftarrow n_k$
  **end for**

  // Initialize final coverage variables
  $N = \sum_{k\in\mathcal{K}} n_list[k]$, $K = |\mathcal{K}|$, $U_{\text{cov}} = [0]_{\mathcal{G}}$, $L_{\text{cov}} = [0]_{\mathcal{G}}$, $PW_{\text{cov}} = [0]_{\mathcal{G}\times\mathcal{G}}$
  `all_comm_efficient` = `all`(`formulations`$[k]$ == `COMM_EFFICIENT`)
  **for** client $k \in \mathcal{K}$ **do**
    $\gamma_k = ((n_list[k] + 1)/(N + K))$
    **if** `all_comm_efficient` **then**
      **if** `use_mle` **then**
        $U_{\text{cov}}$ += $\left(\gamma_k / \pi^{(g,\tilde{y})}\right) \cdot u_list[k]$ // Standard operations are element-wise
        $L_{\text{cov}}$ += $\left(\gamma_k / \pi^{(g,\tilde{y})}\right) \cdot l_list[k]$
      **else**
        $U_{\text{cov}}$ += $\left(\gamma_k / L^{(g,\tilde{y})}\right) \cdot u_list[k]$ // Standard operations are element-wise
        $L_{\text{cov}}$ += $\left(\gamma_k / U^{(g,\tilde{y})}\right) \cdot l_list[k]$
      **end if**
    **else**
      **if** `formulations`$[k]$ == `COMM_EFFICIENT` **then**
        $PW_{\text{cov}}$ += $\gamma_k \cdot \left(u_list[k] \ominus l_list[k]^\top\right)$ // $\ominus$ is pairwise differences between two vectors.
      **else**
        $PW_{\text{cov}}$ += $\gamma_k \cdot pw_cg_list[k]$
      **end if**
    **end if**
  **end for**

  **if** `all_comm_efficient` **then**
    $U_{\text{cov}} =$ element_wise_min$(U_{\text{cov}}, [1]_{\mathcal{G}})$ // Limit upper coverage prior to coverage gap calculation
    cov_gap $= \max_{g\in\mathcal{G}} U_{\text{cov}}[g] - \min_{g\in\mathcal{G}} L_{\text{cov}}[g]$
  **else**
    cov_gap $= \min\left\{\max_{g_a,g_b\in\mathcal{G}} PW_{\text{cov}}[g_a, g_b], 1\right\}$ // Limit Coverage Gap to 1
  **end if**
  **return** cov_gap
**end function**

### H.2 Hybrid

In real-world scenarios, clients often have varying privacy and communication requirements. For example, clients in resource-constrained areas may not have the network bandwidth to send the necessary packets to the centralized server. In our proposed *hybrid* approach, a client may elect to be *communication efficient*, without preventing the remaining clients from using the *enhanced privacy* protocol. We present the full server-side algorithm, which combines the *communication efficient*, *enhanced privacy*, and *hybrid* protocols for the federated coverage gap, in Algorithm H.3.

### H.3 Empirical Comparison

We conduct two experiments using the Fitzpatrick dataset and $8$ clients, as well as the larger ACSIncome dataset with the *continental_all* partition scheme–$48$ clients–to test the *communication efficient*, *enhanced privacy*, and *hybrid* protocols. For the *hybrid* protocol, we randomly assign half the clients to each protocol. From Table H.1, we observe that all configurations control the fairness disparity within the closeness criterion; however, if all clients agree upon the *communication efficient* protocol, FedCF achieves a better efficiency with a slightly worse fairness disparity, albeit still within the closeness criterion. Though with more data, we observe that the efficiency gaps are smaller as seen in Table H.2.

With these empirical results, note that under the hybrid setting, clients that optimize for communication efficiency still benefit from the fact that they can operate over a limited bandwidth network connection. The required bandwidth for a particular client undergoes a factor of $\approx \frac{|\mathcal{G}|}{2}$ reduction–i.e. $\mathcal{O}(|\mathcal{G}|^2|\mathcal{Y}^+) \rightarrow \mathcal{O}(2 \cdot |\mathcal{G}||\mathcal{Y}^+)$, when a client selects the *communication efficient* protocol while ensuring the remaining clients benefit from the *enhanced privacy* protocol. For Fitzpatrick, this results in the communication overhead (in bytes) being reduced by a factor of three. ($|\mathcal{G}|/2 = 3$, for Fitzpatrick). For the ACS datasets using the small, large, or all client assignments, this reduction corresponds to $|\mathcal{G}|/2 = 4.5$. In our implementation, with additional message overhead (to transmit the data structure), the number of bytes transmitted for the ACSIncome (Continental Large) dataset increases from 614.9 bytes for the *communication-efficient* protocol to 897.2 bytes for the *enhanced-privacy* protocol (Table H.3).

Table H.1: **Fitzpatrick, 8 clients, RAPS.** Each entry is of the form, **efficiency/fairness disparity**. We bold the lower fairness disparity value for each comparison. We observe that the *communication efficient* approach produces the most efficient prediction sets, while having a similar or higher fairness disparity. The *enhanced privacy* approach and *hybrid* approach have similar performance (w.r.t efficiency and fairness disparity), with minor differences stemming from the stochasticity of FedCF, as they default to the same coverage-gap aggregation protocol (see Algorithm H.3). All methods improve upon the baseline fairness disparity and control for the closeness criterion.

(a) Enhanced Privacy

| Metric | $c = 0.1$ | | $c = 0.15$ | | $c = 0.2$ | |
|---|---|---|---|---|---|---|
| | Base | Ours | Base | Ours | Base | Ours |
| **Dem_Parity** | 3.504 / 0.164 | 5.314 / **0.089** | 3.500 / 0.166 | 3.880 / **0.089** | 3.501 / 0.166 | 3.641 / **0.146** |
| **Pred_Eq** | 3.500 / 0.170 | 5.328 / **0.082** | 3.502 / 0.169 | 3.944 / **0.087** | 3.502 / 0.169 | 3.525 / **0.150** |

(b) Hybrid (50-50)

| Metric | $c = 0.1$ | | $c = 0.15$ | | $c = 0.2$ | |
|---|---|---|---|---|---|---|
| | Base | Ours | Base | Ours | Base | Ours |
| **Dem_Parity** | 3.501 / 0.164 | 5.313 / **0.088** | 3.502 / 0.166 | 3.881 / **0.089** | 3.503 / 0.162 | 3.586 / **0.146** |
| **Pred_Eq** | 3.499 / 0.170 | 5.328 / **0.082** | 3.506 / 0.164 | 3.946 / **0.087** | 3.502 / 0.171 | 3.524 / **0.150** |

(c) Communication Efficient

| Metric | $c = 0.1$ | | $c = 0.15$ | | $c = 0.2$ | |
|---|---|---|---|---|---|---|
| | Base | Ours | Base | Ours | Base | Ours |
| **Dem_Parity** | 3.504 / 0.164 | 5.241 / **0.089** | 3.505 / 0.164 | 3.863 / **0.087** | 3.504 / 0.162 | 3.586 / **0.146** |
| **Pred_Eq** | 3.502 / 0.169 | 5.293 / **0.083** | 3.500 / 0.168 | 3.870 / **0.087** | 3.503 / 0.169 | 3.525 / **0.150** |

Table H.2: **ACSIncome, Continental All, RAPS.** Each entry is of the form, **efficiency/fairness disparity**. We observe that with sufficient data, each protocol performs at a similar efficiency, and they all decrease the baseline fairness disparity and control it within the closeness criterion. Our fairness disparity values are bolded.

(a) Enhanced Privacy

| Metric | $c = 0.1$ | | $c = 0.15$ | | $c = 0.2$ | |
|---|---|---|---|---|---|---|
| | Base | Ours | Base | Ours | Base | Ours |
| **Dem_Parity** | 2.611 / 0.148 | 3.046 / **0.086** | 2.609 / 0.148 | 2.633 / **0.138** | 2.613 / 0.148 | 2.613 / **0.148** |
| **Pred_Eq** | 2.611 / 0.161 | 3.307 / **0.074** | 2.611 / 0.161 | 2.662 / **0.138** | 2.612 / 0.161 | 2.612 / **0.161** |

(b) Hybrid (50-50)

| Metric | $c = 0.1$ | | $c = 0.15$ | | $c = 0.2$ | |
|---|---|---|---|---|---|---|
| | Base | Ours | Base | Ours | Base | Ours |
| **Dem_Parity** | 2.612 / 0.148 | 3.040 / **0.086** | 2.608 / 0.148 | 2.632 / **0.138** | 2.609 / 0.148 | 2.609 / **0.148** |
| **Pred_Eq** | 2.607 / 0.160 | 3.304 / **0.071** | 2.609 / 0.161 | 2.660 / **0.138** | 2.611 / 0.161 | 2.611 / **0.161** |

(c) Communication Efficient

| Metric | $c = 0.1$ | | $c = 0.15$ | | $c = 0.2$ | |
|---|---|---|---|---|---|---|
| | Base | Ours | Base | Ours | Base | Ours |
| **Dem_Parity** | 2.611 / 0.148 | 3.043 / **0.086** | 2.608 / 0.148 | 2.632 / **0.138** | 2.610 / 0.148 | 2.610 / **0.148** |
| **Pred_Eq** | 2.609 / 0.161 | 3.315 / **0.073** | 2.612 / 0.161 | 2.663 / **0.138** | 2.609 / 0.161 | 2.609 / **0.161** |

Table H.3: **ACSIncome, Continental Large.** Average bytes transmitted by each client under the communication-efficient and enhanced-privacy protocols across 100 communication rounds. The communication-efficient protocol incurs lower communication overhead than the enhanced-privacy protocol.

| **Client** | **Communication Efficient** | **Enhanced Privacy** |
|---|---|---|
| Client 0 | 619.5 | 885.2 |
| Client 1 | 612.8 | 896.5 |
| Client 2 | 605.0 | 881.1 |
| Client 3 | 609.0 | 907.8 |
| Client 4 | 603.2 | 896.5 |
| Client 5 | 636.8 | 917.8 |
| Client 6 | 610.4 | 904.8 |
| Client 7 | 622.5 | 888.4 |
| **Average** | **614.9** | **897.2** |

# I Differential Privacy in FedCF

As both the Communication-Efficient and Enhanced Privacy approaches can be targeted by reconstruction attacks from a malicious server, we extend FedCF to formally consider $(\epsilon, \delta)$-differential privacy (DP) to maintain robustness under adversarial conditions. DP is a mathematically rigorous framework for data privacy [Dwork, 2006], where $\delta$ is the probability that $\epsilon$-DP is violated. We can embed DP within our framework via client shuffling and additive noise approaches. Client shuffling is a global DP approach that is performed after the client sends data. Before the server receives the data, it goes through a trusted, centralized shuffler to anonymize which client has sent what data [Erlingsson et al., 2019]. Our framework can accommodate client shuffling due to its parallelism with client-side computation and its additive aggregation approach.

For additive noise, we propose augmenting the values each client sends back with Gaussian noise [Dwork et al., 2014, Dong et al., 2022], such that a client returns,

$$h = \frac{(\alpha_k^{(g_a,\tilde{y});\lambda} + 1)}{(n_k+1)L^{(g_a,\tilde{y})}} - \frac{\alpha_k^{(g_b,\tilde{y});\lambda}}{(n_k+1)U^{(g_b,\tilde{y})}} + X, \tag{59}$$

where $X$ is a Gaussian random variable (R.V). For the *communication efficient* approach, one would add a Gaussian R.V. to the upper coverage and lower coverage terms returned by the client. To ensure $(\epsilon, \delta)$-DP, we make $X \sim \mathcal{N}\left(0, \frac{2\ln(1.25/\delta)(\Delta h)^2}{\epsilon^2}\right)$, where $\Delta h$ is the sensitivty of $h$–or how much $h$ can change if one of the points in the client's dataset changes. For FedCF, $h$ can be affected by data changes in the covariates (or non-conformity scores), labels, and group memberships.

## I.1 Example: Differential Privacy Bounds for Enhanced Privacy Protocol

Observe using the *enhanced privacy* approach, $\Delta h \leq \frac{1}{n_k}\left(\frac{1}{L^{(g_a,\tilde{y})}} + \frac{1}{U^{(g_b,\tilde{y})}}\right)$. For the *communication efficient* approach $\Delta h \leq \frac{1}{n_k}$ for the components of the upper and lower coverage terms sent by the client. The server will know each client's sensitivity and their choice of $\epsilon$ and $\delta$.

To demonstrate how the server can estimate the coverage gap, we will consider an example using the *enhanced privacy* approach. The result from server aggregation is,

$$\begin{aligned}&\text{cov_gap_est}(\lambda, F_M, g_a, g_b, \tilde{y}) \\ &\quad := \sum_{k=1}^{K} \gamma_k \underbrace{\left\{ \frac{(\alpha_k^{(g_a,\tilde{y});\lambda} + 1)}{(n_k+1)L^{(g_a,\tilde{y})}} - \frac{\alpha_k^{(g_b,\tilde{y});\lambda}}{(n_k+1)U^{(g_b,\tilde{y})}} + X_k^{(g_a,g_b,\tilde{y})} \right\}}_{\text{Returned by the Client}},\end{aligned} \tag{60}$$

where $X_k^{(g_a,g_b,\tilde{y})} \sim \mathcal{N}(0, {\sigma^2}_{k;(g_a,g_b,\tilde{y})})$ such that ${\sigma^2}_{k;(g_a,g_b,\tilde{y})}$ provides $(\epsilon_k, \delta_k)$-DP for the client. Then observe,

$$\begin{aligned}&\text{cov_gap_est}(\lambda, F_M, g_a, g_b, \tilde{y}) \\ &\quad = \underbrace{\sum_{k=1}^{K} \gamma_k \left\{ \frac{(\alpha_k^{(g_a,\tilde{y});\lambda} + 1)}{(n_k+1)L^{(g_a,\tilde{y})}} - \frac{\alpha_k^{(g_b,\tilde{y});\lambda}}{(n_k+1)U^{(g_b,\tilde{y})}} \right\}}_{\text{true coverage gap}} + \underbrace{\sum_{k=1}^{K} \gamma_k X_k^{(g_a,g_b,\tilde{y})}}_{\text{Guassian R.V}} \qquad (61) \\ &\quad = \text{cov_gap}(\lambda, F_M, g_a, g_b, \tilde{y}) + X, \;\; X \sim \mathcal{N}\left(0, \sum_{k=1}^{K} \gamma_k^2 {\sigma^2}_{k;(g_a,g_b,\tilde{y})}\right) \qquad (62)\end{aligned}$$

Using a prespecified probability $\beta$ we can accept or reject the statement $\text{cov_gap_est}(\lambda, F_M, g_a, g_b, \tilde{y}) \leq c$. In other words, we can check whether,

$$\text{cov_gap}(\lambda, F_M, g_a, g_b, \tilde{y}) + X \leq c \implies X \leq c - \text{cov_gap}(\lambda, F_M, g_a, g_b, \tilde{y})$$

$$\implies \boxed{\frac{X}{\underbrace{\sqrt{\sum_{k=1}^{K} \gamma_k X_k^{(g_a, g_b, \tilde{y})}}}_{\text{Standard Normal RV}}} \leq \frac{c - \text{cov_gap}(\lambda, F_M, g_a, g_b, \tilde{y})}{\sqrt{\sum_{k=1}^{K} \gamma_k X_k^{(g_a, g_b, \tilde{y})}}}}.$$

Then, if $\Phi\left(\frac{c - \text{cov_gap}(\lambda, F_M, g_a, g_b, \tilde{y})}{\sqrt{\sum_{k=1}^{K} \gamma_k X_k^{(g_a, g_b, \tilde{y})}}}\right) > \beta$, where $\Phi$ is the CDF of the standard normal distribution, we can accept the coverage gap as being less than $c$. In other words, with probability $\beta$, the closeness criterion is satisfied with $\lambda$.

While using Gaussian noise results in a PAC-style guarantee, one can instead keep the strict (non-PAC) guarantee of FedCF by adding non-negative noise to the upper coverage terms and subtracting it from the lower coverage terms. The server then can only overestimate the coverage gap, so if the noisy coverage gap satisfies the closeness criterion, so does the exact coverage gap. We discuss this approach in Appendix I.2.

### I.2 Strict Differential Privacy

To preserve the strict (non-PAC) guarantee of FedCF, the noise added by each client must never decrease the estimated coverage gap. In this vein, each client adds non-negative noise to its upper coverage terms and subtracts it from its lower coverage terms, so that the server can only overestimate the coverage gap. We use the truncated Laplace mechanism [Geng et al., 2020], where $X \sim \text{Laplace}(0, \Delta h/\epsilon)$ is truncated to $[-A, A]$ with $A = \frac{\Delta h}{\epsilon} \ln\left(1 + \frac{e^\epsilon - 1}{2\delta}\right)$, and shifted by $A$ to be non-negative. Observe that a non-zero $\delta$ is necessary. Since the noisy coverage gap is never below the exact coverage gap, observing it rules out with certainty every dataset with a larger coverage gap, which $(\epsilon, 0)$-DP does not allow [Humphries et al., 2021]. Note that the strict guarantee can also be satisfied trivially: if the noisy coverage gap exceeds $c$ for every $\lambda < \lambda_{\max}$, FedCF returns the degenerate solution $\lambda_{\text{opt};(\tilde{y})} = \lambda_{\max}$ (cf. Theorem 3.2), which has a fairness disparity of 0 but includes $\tilde{y}$ in every prediction set.

#### I.2.1 Per-Round Noise

Adding noise to the coverage terms returned in every round of the communication efficient approach requires splitting the privacy budget across the $T$ server rounds [Dwork et al., 2010b]. Moreover, after server aggregation, the noise is scaled by $1/N^{(g,\tilde{y})}$, making it largest for small groups. We observe that on ACSIncome (continental large, 8 clients) using RAPS for Demographic Parity, with $T = 100$ and $\delta = 10^{-5}$, FedCF returns the degenerate solution for every $\epsilon \leq 100$ and only recovers for $\epsilon \geq 10^3$, as the smallest group has only 2,960 calibration points.

#### I.2.2 One-Shot Noisy Histogram

To avoid splitting the budget across rounds, each client instead sends a noisy summary of its calibration scores once, before the threshold search begins. For each $(g, \tilde{y})$ pair, the client counts the scores in $\mathcal{S}_k^{(g,\tilde{y})}$ that fall in each of $B$ fixed bins. For a threshold $\lambda$ on a bin edge, $\alpha_k^{(g,\tilde{y});\lambda}$ is then the sum of the counts of all bins up to $\lambda$ (i.e., a prefix sum of the histogram). Since every $\alpha_k^{(g,\tilde{y});\lambda}$ needed during the search is computed from the noisy histogram, the entire search is post-processing of a single $(\epsilon, \delta)$-DP output, and there is no composition over rounds.

However, adding noise to each bin would make a prefix sum accumulate up to $B$ noise terms. Following the tree mechanism [Dwork et al., 2010a, Chan et al., 2011], the client instead stores the histogram as a binary tree, where each leaf is a bin, each internal node holds the total count of its two children, and every node receives noise. Any prefix sum is then the sum of at most $\log_2 B$ nodes. For example, with $B = 8$, the total count of bins 1–6 is the count stored at the node for bins 1–4 plus the

count stored at the node for bins 5–6, i.e., 2 noisy values instead of 6. In turn, each calibration point is counted in one node per level (e.g., a score in bin 6 is counted in the nodes for bin 6, bins 5–6, bins 5–8, and bins 1–8), i.e., $\log_2 B + 1$ nodes, so the budget is split over these nodes.

Since the noise of each node lies in $[-A, A]$, a prefix sum over $j$ nodes is within $jA$ of its exact value. Thus, the client adds $jA$ to the upper coverage term and subtracts $jA$ from the lower coverage term, so the server still only overestimates the coverage gap. These bounds require $\lambda$ to lie on a bin edge; otherwise, the bin containing $\lambda$ would have to be counted entirely in the upper term and excluded entirely from the lower term, inflating the coverage gap. Thus, the server rounds every queried $\lambda$ up to the nearest bin edge before sending it to the clients. This preserves the strict guarantee, since the coverage gap is evaluated at the rounded $\lambda$, which is also the threshold used to construct the test prediction sets. Rounding up only enlarges the prediction sets, by at most one bin width, so it costs some efficiency but never validity. Note that a larger $B$ results in finer thresholds, but more noise per prefix sum.

The noise added to each count is independent of the group size; however, the coverage terms divide the counts by $N^{(g,\tilde{y})}$, so the noise has the largest effect on small groups. In ACSIncome, with the six race groups, the smallest group (American Indian) has only 2,960 calibration points across all clients, and the noise alone would exceed $c$ for moderate $\epsilon$. Thus, we merge the race groups into White and non-White ($N^{(g,\tilde{y})} = 321{,}976$ and $89{,}062$). Note that merging changes the fairness specification, as closeness is then only guaranteed between the merged groups. This mirrors Theorem 3.2, which already requires sufficient calibration data for every group–adding privacy raises this requirement.

We report the results in Table I.1. We observe that FedCF is degenerate for $\epsilon \leq 2$, but for $\epsilon \in \{5, 10\}$, FedCF controls for the closeness criterion and improves upon the fairness disparity of FCP, at a moderate cost in efficiency. In contrast, adding noise in every round is still degenerate at $\epsilon = 5$ for the same two groups. These values of $\epsilon$ are within the range commonly recommended for machine learning [Ponomareva et al., 2023]. Lastly, the group counts $n_k^{(g,\tilde{y})}$ and the initial FCP quantile are sent without noise, but can be privatized with standard mechanisms.

Table I.1: **ACSIncome (continental large, 8 clients, 2 groups) using RAPS for Demographic Parity with one-shot noisy histograms.** Each entry is of the form, **efficiency/fairness disparity**, averaged over 3 seeds, with $c = 0.1$ and $\delta = 10^{-5}$. We observe that for $\epsilon \geq 5$, FedCF controls for the closeness criterion and improves upon the fairness disparity of FCP. FCP and FedCF without noise do not depend on the number of bins $B$.

| Method | $B = 32$ | $B = 64$ |
|---|---|---|
| FCP | 2.359 / 0.135 | |
| FedCF (no noise) | 2.527 / 0.103 | |
| FedCF, $\epsilon = 1$ | 3.405 / 0.000 | 3.406 / 0.000 |
| FedCF, $\epsilon = 2$ | 3.406 / 0.000 | 3.406 / 0.000 |
| FedCF, $\epsilon = 5$ | 2.803 / 0.067 | 2.874 / 0.057 |
| FedCF, $\epsilon = 10$ | 2.803 / 0.067 | 2.668 / 0.067 |

# J FedCF for Auditing

Auditing tools are vital for regulatory bodies to ensure ML models comply with fairness and safety standards [Maneriker et al., 2023]. In this regard, we present how FedCF can be used to determine if a federated conformal predictor is *fair* according to the regulator's specification of fairness and closeness criterion, $c$ [U.S. Equal Employment Opportunity Commission, 1979, New York City Council, 2021, 2023, European Parliament and Council of the European Union, 2024].

To assess compliance, FedCF can use the global threshold ($\lambda$) values used by the previously trained conformal-predictor and provide it to each client. Then, the client should send the sufficient values calculated via Algorithm H.1 (or Algorithm H.2) to compute the federated coverage gap. The server would aggregate these values using Algorithm H.3. If the calculated coverage gap is below $c$, then the server can assert that the conformal predictor is fair.

Our auditing approach does not require all clients to provide data for auditing. As discussed in Section 3.2, our guarantees hold assuming that the test-point, $(\boldsymbol{x}_{\text{test}}, y_{\text{test}}) \sim \sum_{k=1}^{K} \gamma_k P_k$, is sampled from a mixture of client distributions where $\gamma_k$ is the probability the test point is sampled from $P_k$, or equivilantly is exchangeable with data from client $k$. Thus, if a subset of clients used to train the original federated conformal predictor provides auditing data, then the audit guarantees will hold assuming that $(\boldsymbol{x}_{\text{test}}, y_{\text{test}})$ are sampled from a mixture consisting of the subset of clients used for auditing. This result allows clients to *independently* decide if they would like to submit data for auditing.

The auditing tool provided by FedCF can also be used to ascertain the *marginal* fairness with respect to each client. Using the auditing procedure described above with data from one client, FedCF can determine if the global, federated conformal predictor maintains fairness with respect to data from a single client. If the computed coverage gap is less than $c$, then the fairness guarantees hold with regard to $(\boldsymbol{x}_{\text{test}}, y_{\text{test}}) \sim P_k$, i.e., the client's marginal distribution.

# K More Results

Here, we provide additional results for the ACS and Pokec-{n,z} datasets. Recall, in each figure, we use a **solid** line to represent the *average* efficiency of the **base federated conformal predictors** across different thresholds and a **dashed** line to represent the corresponding *average* worst-case fairness disparity. The bar plot shows the efficiency and worst-case fairness disparity using FedCF, while the **dots** indicate the *desired* fairness disparity. We report the average base performance for clarity and readability

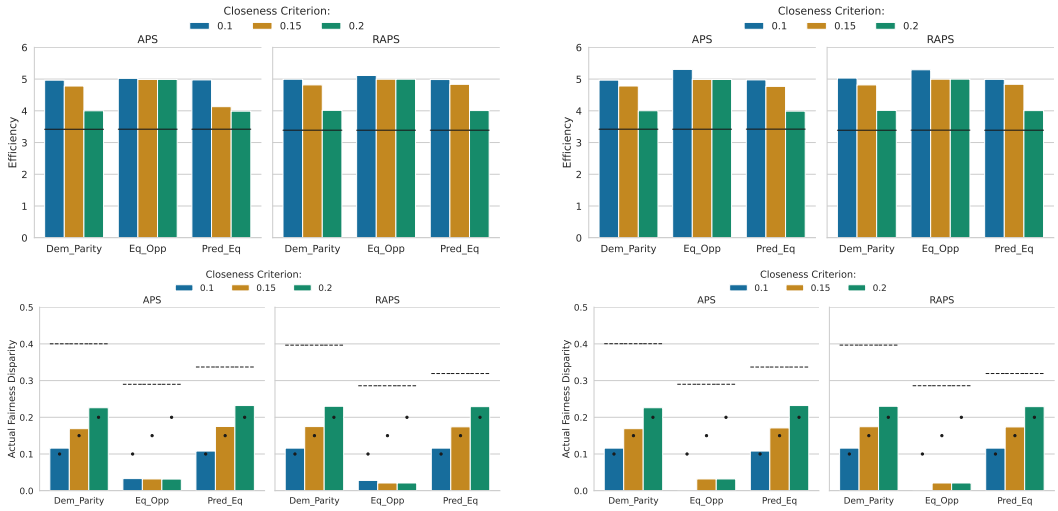
## K.1 Coverage Results

A key property of FedCF is that it ensures the base coverage guarantee is satisfied by bounding the threshold search space by FCP's conformal quantile, $\hat{q}(\alpha)$ from (3). In every case, FedCF's coverage will be greater than or equal to FCP's coverage.

Table K.1: Various datasets using Demographic Parity. We compare the marginal coverage of the Base (FCP) approach against FedCF over different closeness criteria ($c$).

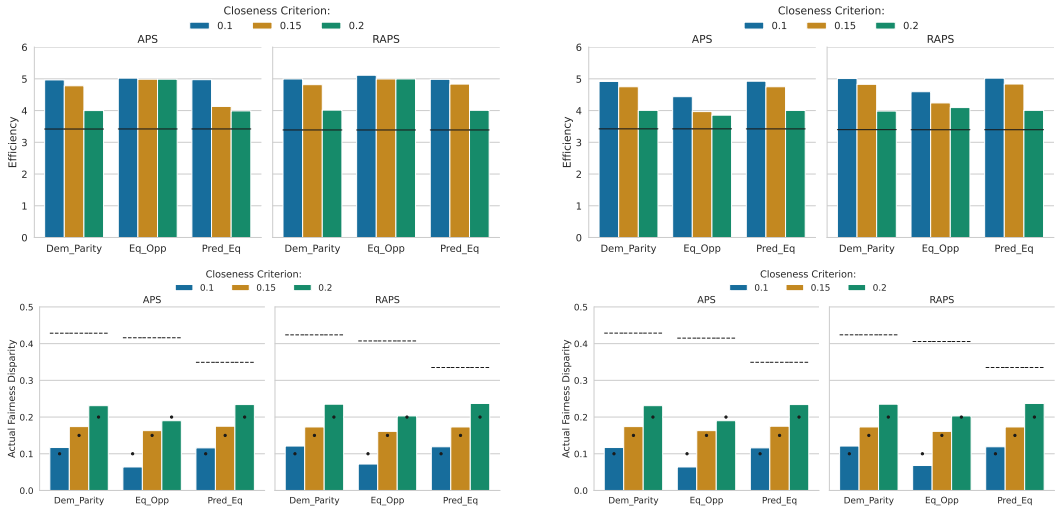
| | Pokec-$\{n, z\}$ (DAPS) | | | Fitzpatrick (RAPS) | | | ACSEducation (RAPS) | | |
|---|---|---|---|---|---|---|---|---|---|
| $c$ | 0.1 | 0.15 | 0.2 | 0.1 | 0.15 | 0.2 | 0.1 | 0.15 | 0.2 |
| Base (FedCP) | 0.892 | 0.892 | 0.892 | 0.895 | 0.895 | 0.895 | 0.932 | 0.932 | 0.932 |
| FedCF | 0.898 | 0.894 | 0.892 | 0.927 | 0.909 | 0.899 | 0.977 | 0.973 | 0.952 |

## K.2 Impact of Data Heterogeneity on ACSEducation: US vs Continental US



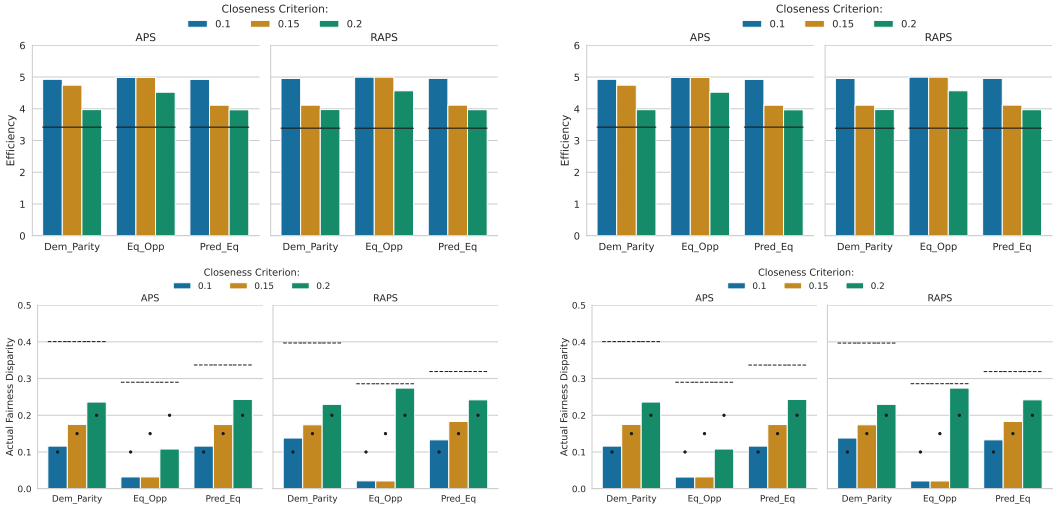

(a) **Communication Efficient.** (b) **Enhanced Privacy.**

Figure K.1: **ACSEducation, Small, Interval Bounds.** The plots in the top row indicate the efficiency with the corresponding fairness disparity plots in the bottom row. We observe that when all US states are included (and Puerto Rico), the closeness criterion is satisfied. However, the efficiency for Equal Opportunity is high for all closeness criterion values, especially compared to the continental US version of ACSEducation in Figure K.2. This result stems from a conservative coverage gap estimate during calibration due to limited covariate representation for some groups.



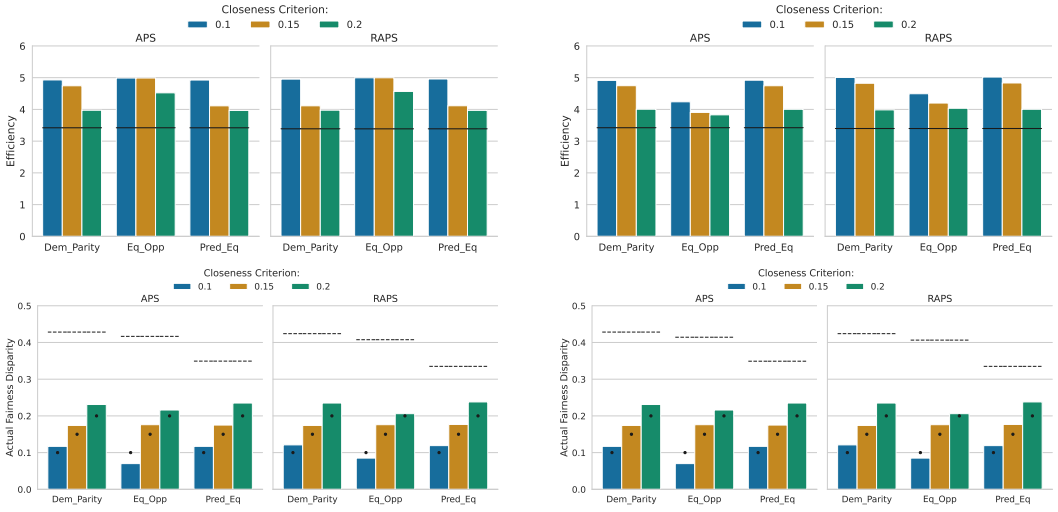

(a) **Communication Efficient.** (b) **Enhanced Privacy.**

Figure K.2: **ACSEducation, Continental Small, Interval Bounds.** The top row demonstrates the efficiency of FedCF when using the continental version of ACSEducation, and its fairness disparity on the bottom row. Compared to Figure K.1, the efficiencies improved (particularly for Equal Opportunity using RAPS), due to increased covariate representation for all sensitive groups.




(a) **Communication Efficient.** (b) **Enhanced Privacy.**

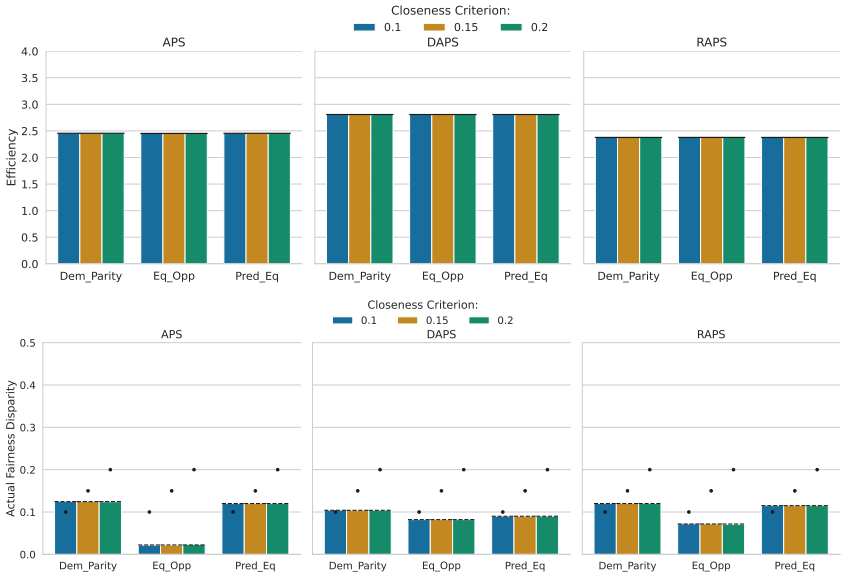
Figure K.3: **ACSEducation, Small, Point Estimates** The plots in the top row indicate the efficiency with the corresponding fairness disparity plots in the bottom row. We observe that using point estimates will result in a similar or lower efficiency than using the interval bounds approach in Figure K.1, at the cost of a similar or higher fairness violation. Because the MLE does not provide a finite sample guarantee, the violation can exceed the desired closeness criterion, but will be lower than the baseline federated conformal predictor.



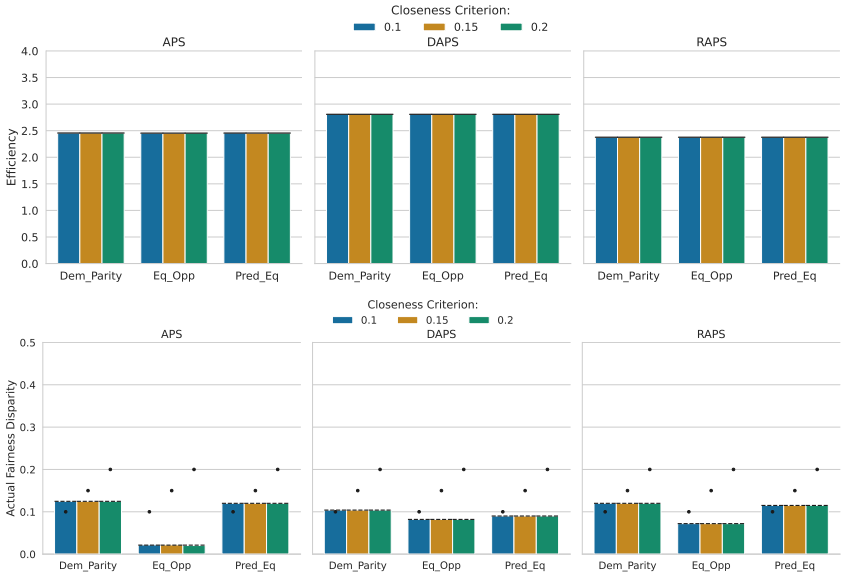

(a) **Communication Efficient.** (b) **Enhanced Privacy.**

Figure K.4: **ACSEducation, Continental Small, Point Estimates.** The plots in the top row indicate the efficiency with the corresponding fairness disparity plots in the bottom row. We observe that using point estimates will result in a similar or lower efficiency than using the interval bounds approach in Figure K.2, at the cost of a similar or higher fairness violation. Because the MLE does not provide a finite sample guarantee, the violation can exceed the desired closeness criterion, but will be lower than the baseline federated conformal predictor.

## K.3 Impact of different sensitive attributes for Pokec-{n,z}




(a) **Communication Efficient.**

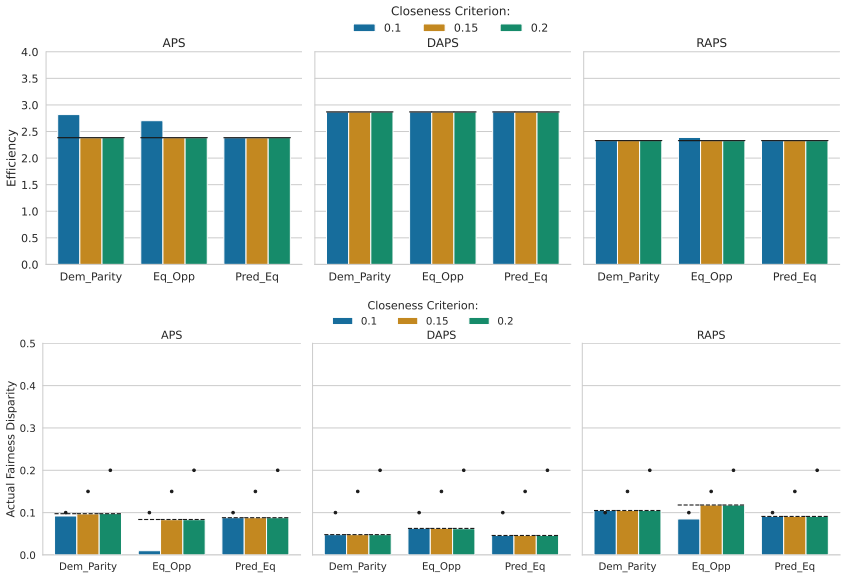



(b) **Enhanced Privacy.**

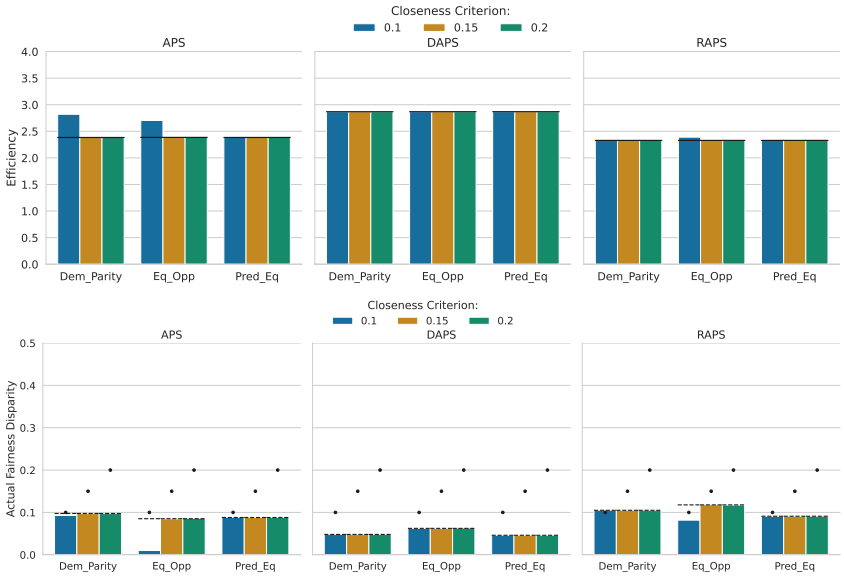
Figure K.5: **Pokec-{n,z}, *gender*.** For each plot (a) and (b), the top plots are for the efficiency, and the bottom plots are for the fairness disparity. The baseline disparity is within the closeness criterion, so we see no changes in efficiency when using FedCF. This is the case when using either the *communication efficient* and *enhanced privacy* protocols.

## K.4 Efficiency vs. Closeness Criterion

Figure K.8 shows how the average prediction set size changes as the closeness criterion $c$ is relaxed on ACSIncome (continental large partition). Efficiency improves quickly up to the elbow at $c = 0.125$, where the efficiency drops from 3.32 at $c = 0.025$ to 2.77, a reduction of 0.55 over a 0.1 increase in $c$. Beyond it, relaxing $c$ further results in a smaller improvement to efficiency and approaches the fairness-agnostic base predictor. Practitioners can use a similar curve to choose the smallest $c$ before



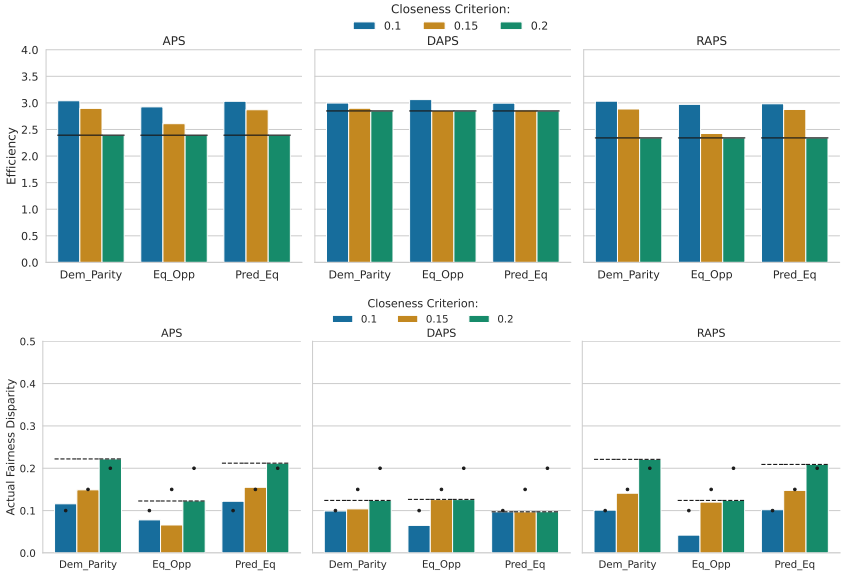

(a) **Communication Efficient.**



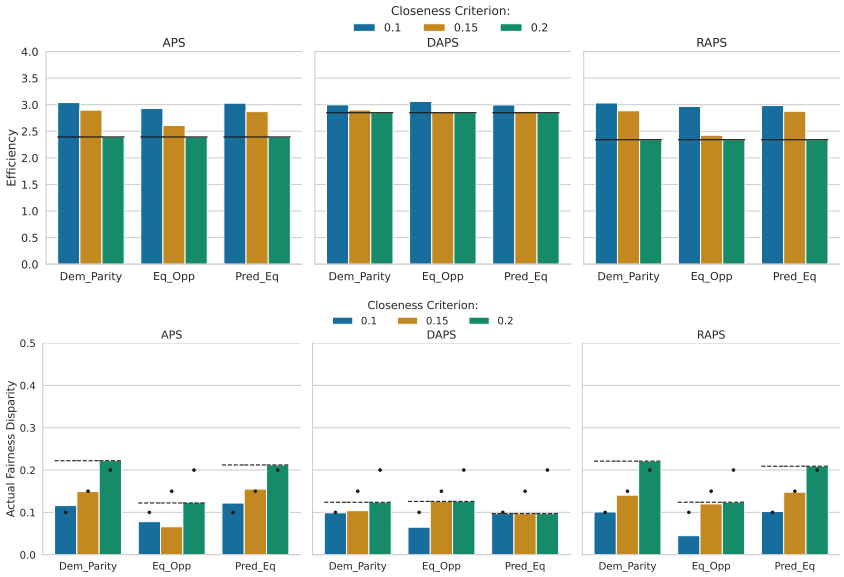

(b) **Enhanced Privacy.**

Figure K.6: **Pokec-{n,z},** ***region*****.** For each plot (a) and (b), the top plots are for the efficiency, and the bottom plots are for the fairness disparity. Note that while the baseline disparity is within the closeness criterion for the test set, the finite-sample guarantee from using the interval bounds ensures FedCF looks for a better threshold, resulting in a smaller violation with a small cost to efficiency. This is the case when using either the *communication efficient* and *enhanced privacy* protocols.

efficiency degrades sharply. The measured test-set disparity (right) closely tracks the target $y = c$, exceeding it by at most 0.015.




(a) **Communication Efficient.**




(b) **Enhanced Privacy.**

Figure K.7: **Pokec-{n,z}, *region* and *gender*.** For each plot (a) and (b), the top plots are for the efficiency, and the bottom plots are for the fairness disparity. In the case of intersectional fairness, since there are more groups, the violation will be worse than considering a single sensitive attribute. We observe that in all cases, FedCF produces a threshold that satisfies the closeness criterion, at a slight cost to efficiency. This is the case when using either the *communication efficient* and *enhanced privacy* protocols.

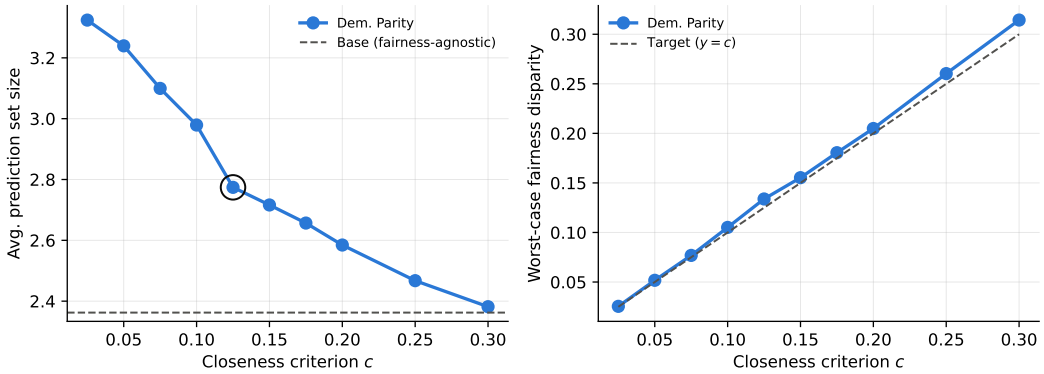



Figure K.8: ACSIncome (continental large, 8 clients) using RAPS for Demographic Parity. Left: average prediction set size vs. closeness criterion $c$; the dashed line is the fairness-agnostic base predictor and the circle marks the elbow. Right: worst-case fairness disparity vs. $c$; the dashed line is $y = c$.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We provide the theoretical justification and implementation details in Section 3 and empirical results in Section 4. Further theory and proofs are available in Appendix C, D, and F with more results in Appendix K.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have a Limitations paragraph in the conclusion.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theorems that are presented in Section 3 have proofs in Appendix D and F, with assumptions explicilty detailed.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All details needed to reproduce results are presented in Section 4 and Appendix E. The code used for the experiments is available in the supplemental material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code is provided in the supplementary material, while all datasets are available publicly (details on procuring data can be found in the code).

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: This information is provided in Section 4 and Appendix E.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We ran the experiments for 3 different seeds for random data splitting of the calibration and test sets, and found the deviations to be too small to be legibly plotted in the figures.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include this information in Appendix E in Section E.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research adheres to the code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the impacts in Section 7.3.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper and described methods do not have such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We provide citations and licenses for the datasets we used in Appendix E.1, and adhered to their respective licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We have not released any assets. Once our code is made public, we will release it under the Apache-2.0 license.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: N/A

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: N/A

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were **not** used during development.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.