

Federated Computation of ROC and PR Curves

Xuefeng Xu
University of Warwick
Coventry, United Kingdom
xuefeng.xu@warwick.ac.uk

Graham Cormode
University of Oxford
Oxford, United Kingdom
graham.cormode@cs.ox.ac.uk

ABSTRACT

Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves are fundamental analytical tools for evaluating classification models, providing a comprehensive view of performance across decision thresholds. In modern data management settings, such evaluation must often be performed over data that is distributed across multiple parties and subject to strict privacy constraints. This setting arises naturally in federated learning (FL) and collaborative data platforms, where raw prediction scores and labels cannot be centrally collected. We study the problem of computing ROC and PR curves over distributed data under formal privacy guarantees. The key challenge is that exact computation requires access to all prediction scores and labels, leading to linear communication costs and privacy risks. To address this, we propose a novel framework for approximating ROC and PR curves using quantiles of the score distribution, which can be computed efficiently under secure aggregation and distributed differential privacy. Our approach requires only $O(Q)$ communication, where Q is the number of quantiles, and avoids sharing raw data entirely. We provide theoretical guarantees on the approximation quality by bounding the Area Error (AE) between the true and estimated curves, demonstrating principled trade-offs among approximation accuracy, privacy protection, and communication cost. Our method is robust to data heterogeneity and skew, making it suitable for real-world distributed data management scenarios. Experiments on real-world datasets show that our approach achieves high accuracy with low communication overhead under strong privacy constraints. Overall, this work provides a practical and theoretically grounded solution for privacy-preserving evaluation of classification models over distributed data.

PVLDB Reference Format:

Xuefeng Xu and Graham Cormode. Federated Computation of ROC and PR Curves. PVLDB, 20(2): 325 - 337, 2026.
doi:10.14778/3849423.3849443

PVLDB Artifact Availability:

The source code, data, and/or other artifacts have been made available at <https://github.com/xuefeng-xu/fedcurve>.

1 INTRODUCTION

Modern data management systems increasingly operate in decentralized environments where data is distributed across multiple parties and cannot be directly shared due to privacy constraints.

This paradigm appears in federated learning (FL) [25], cross-silo analytics, and collaborative data platforms. In such settings, a central coordinator must support global analytics and monitoring tasks without direct access to raw data, raising fundamental challenges in distributed query processing under privacy constraints. Developing effective solutions for managing these distributed workloads requires novel systems and algorithmic solutions.

A critical yet underexplored task in this setting is the evaluation of classification models. In practice, model evaluation is not a one-time procedure: it is required during training for monitoring convergence; before deployment for model selection; and after deployment for continuous performance tracking. While existing FL systems typically support simple aggregate metrics such as accuracy or loss (e.g., Flower [6]), these metrics are insufficient for understanding model behavior. Accuracy, in particular, is unreliable on imbalanced datasets [33], where a classifier may appear effective simply by predicting the majority class. Loss values are often not interpretable in practice and provide limited insight for deployment decisions. Moreover, one model may achieve a lower loss value, but have worse ROC/PR curves compared to another. Without computing these curves in the federated setting, such critical differences in model behavior would remain hidden. This fundamental limitation prevents a full assessment of model performance, hindering informed decisions about model deployment and improvement.

In contrast, Receiver Operating Characteristic (ROC) [33] and Precision-Recall (PR) [35] curves are standard analytical tools that characterize model performance across all decision thresholds. These curves provide richer insights into trade-offs such as true positive versus false positive rates, and precision versus recall. However, computing these curves in a distributed and privacy-preserving setting poses significant challenges. First, exact computation of ROC and PR curves requires access to all prediction scores and labels, which conflicts with the fundamental constraints of distributed data management systems where raw data cannot be centralized. Second, even if data sharing were allowed, the communication cost of transmitting all scores scales linearly with the dataset size, making it impractical for large-scale systems. Third, prior work has shown that even aggregate outputs such as ROC curves may leak sensitive information [29], and access to prediction scores can enable membership inference attacks [37]. These challenges highlight the need for principled methods that support complex analytical queries over distributed data while providing strong privacy guarantees. We therefore seek solutions that are accurate, private, computationally and communication-efficient, suitable for implementation in emerging data management systems.

In this paper, we formulate ROC and PR curve computation as a distributed analytics problem and propose a novel approximation framework based on quantiles of the prediction score distribution. Our methods combine techniques from data privacy [16] and data

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit <https://creativecommons.org/licenses/by-nc-nd/4.0/> to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.
Proceedings of the VLDB Endowment, Vol. 20, No. 2 ISSN 2150-8097.
doi:10.14778/3849423.3849443

sketching [26]. Our key insight is that ROC and PR curves can be approximated using a small number of quantile points, which can be computed efficiently in a distributed manner without exposing raw data. We design protocols that compute these quantiles using histogram aggregation, enabling accurate and efficient federated estimation regardless of data heterogeneity. Each client locally bins its prediction scores and sends histograms to the server, which reconstructs the global score distribution without accessing raw data. For privacy protection, we integrate distributed differential privacy (DDP) [22], where each client adds independent noise to their local histograms before aggregation. The aggregated noisy histograms are then used to extract quantiles and construct approximate ROC and PR curves via monotone interpolation. This design provides end-to-end privacy guarantees while maintaining communication efficiency and robustness to heterogeneous client data distributions.

We analyze the accuracy of our approach using the Area Error (AE) (Definition 5.1), which measures the deviation between the true and approximate curves. Under mild assumptions on the score distribution, i.e., that they are $O(1)$ -Lipschitz (Definition 5.3), we show that the AE is $O(1/Q)$ for the ROC curve, where Q is the number of quantiles, which governs the communication cost. The AE bound is $\tilde{O}(1/Q)$ ⁱ for the PR curve under mild class imbalance, where the class ratio $r = n^+/n^- \geq 0.1$, with n^+ and n^- representing the number of positive and negative examples, respectively. Under extreme imbalance, i.e., $r \ll 1$, the AE is bounded by $\tilde{O}(1/(Qr))$. This bound is extended to $\tilde{O}(1/Q + 1/(n^+\epsilon) + 1/(n^-\epsilon))$ ⁱⁱ under distributed differential privacy [16], where n^+ and n^- are the number of positive and negative examples and ϵ is the standard privacy budget. Notably, our method is agnostic to data heterogeneity and skew across clients, and requires only $O(Q)$ communication, providing an efficient accuracy-privacy-communication trade-off.

Empirically, we show that with $Q \approx 10^2$, the AE of ROC is often close to 10^{-3} , the AE of PR is below 10^{-2} , and these remain low even under strong privacy settings (e.g., $\epsilon \leq 1$). While the well-behavedness condition is instrumental in proving the theoretical error bounds, our experiments demonstrate that the method achieves consistently low area error even when this condition is violated in practice (see the score distribution in Figure 4), indicating robustness across diverse datasets and classifiers. By implementing our method within an existing distributed system (Flower), we confirm that the communication cost is small in practice, and that other computational costs (time and memory usage) are also low.

Our work bridges techniques from distributed data management, approximate query processing, and differential privacy to enable a new class of analytical queries, i.e., performance evaluation of trade-off curves, over sensitive, distributed data. We present and evaluate novel algorithms for providing characterizations of classifier efficacy with guaranteed privacy and accuracy. The proposed approach is simple to implement, communication-efficient, and compatible with existing federated data-processing systems.

Contributions. Our main contributions are as follows:

- (1) We formulate the computation of ROC and PR curves through distributed queries over the prediction score distributions. We show how both curves can be reconstructed from compact approximations of the distributions.
- (2) We propose a quantile-based approach for approximating the empirical cumulative distribution functions (ECDFs) and construct the ROC and PR curves in the federated setting with $O(Q)$ communication and no raw data sharing, independent of the number of examples and data heterogeneity across clients.
- (3) We introduce Area Error as a measure of full-curve approximation and establish an $O(1/Q)$ bound for the ROC curve and $\tilde{O}(1/Q)$ for the PR curve under a mild smoothness condition of score distribution.
- (4) We present a privacy-preserving variant of our protocol and extend the analysis of the Area Error to secure aggregation and distributed differential privacy. The resulting bounds explicitly capture the trade-offs between accuracy, privacy, and communication cost.
- (5) We perform empirical validation showing high approximation accuracy and low communication overhead under strong privacy guarantees. We implement the protocol in Flower using SecAgg+ protocol [8] and demonstrate its practical feasibility.

While several components of our solution are standard techniques in the literature (secure aggregation, hierarchical histograms, private noise addition), the resulting protocol for ROC and PR curve materialization is novel. In particular, the error analysis in Section 5 provides the first accuracy guarantee for this problem.

Outline. Section 2 discusses related work. Section 3 reviews the relevant background. Section 4 describes our method for federated computation of ROC and PR curves via quantiles. Section 5 presents theoretical bounds on the Area Error under both exact and approximate quantiles, as well as under privacy constraints. Section 6 reports empirical validation on multiple datasets and compares against the prior work. Section 7 discusses further extensions and limitations, and Section 8 concludes the paper.

2 RELATED WORK

Several problems related to evaluation of classifier performance have been studied across data management and machine learning in recent years. Fawcett [17] proposed heuristic methods for merging ROC curves across split datasets, either by averaging true positive rates (TPRs) at fixed false positive rates (FPRs), or by averaging FPR-TPR pairs at fixed thresholds. However, these methods lack any formal error guarantees. Moreover, these averaging schemes do not directly recover the global ROC curve under heterogeneous client data. Chen et al. [10] investigated differentially private ROC computation in centralized settings using median threshold selection and range queries, followed by post-processing for monotonicity, but did not provide formal error bounds. It is also unclear how these methods could be adapted to a distributed setting.

More recent work [4] addresses AUC computation, which is a scalar summary of the ROC curve, under Local DP using hierarchical histograms. The error bound was improved by Cormode and Markov [13] using similar measurements, with relaxed assumptions. However, the magnitude of error due to local DP is large, and so we seek a better tradeoff between privacy and accuracy. Sun et al.

ⁱAs usual, $\tilde{O}$ notation suppresses logarithmic factors.

ⁱⁱIn the extreme class imbalance case for the PR curve, the error bound becomes $\tilde{O}(1/(Qr) + 1/(n\epsilon))$; throughout the paper, our analysis focuses on mild class imbalance, where the error is $\tilde{O}(1/Q)$.

[38] explored DP-based AUC computation in a distributed setting. Their approach is to compute the FPRs and TPRs locally at each client, add noise and then combine them. While this guarantees privacy, it does not provide any bound on the result accuracy.

While AUC provides a scalar performance summary, we argue that the full ROC and PR curves offer a more detailed and hence valuable view, requiring different techniques for computation and error analysis. In parallel with the development of this work, Barczewski et al. [2] recently proposed differentially private computation of ROC curve in FL. Their proposed approach implements a uniformly spaced points that (approximately) define the ROC curve. For privacy and security, the algorithm adds differentially private noise to the gathered statistics, and uses function secret sharing to securely evaluate them. Finally, post-processing is applied to smooth the noisy curves produced. Their analysis bounds the *Squared Point Error* at the selected query thresholds. Because their guarantee and ours differ in the error metric, evaluation domain, and underlying assumptions, they are not directly comparable. Their approach also incurs greater computational and communication overhead, while our empirical evaluation shows that our method achieves lower Area Error in most tested settings.

3 PRELIMINARIES

This section reviews the key concepts underlying our approach, including ROC/PR curves, differential privacy in distributed settings, federated quantile estimation, and monotone interpolation.

3.1 The ROC and PR Curves

The **ROC Curve** evaluates classifiers by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR). The x-axis represents the FPR, defined as $F = FP/n^-$, where FP counts false positives (as the score threshold varies) and n^- is the total number of negative examples. The y-axis represents the TPR, defined as $T = TP/n^+$, where TP counts true positives and n^+ is the number of positive examples. Since both FPR and TPR are independent of the class distribution, the ROC curve remains invariant to class skew. An efficient algorithm for constructing the ROC curve is outlined by Fawcett [17]. It involves sorting prediction scores in descending order, using each score as a threshold, and computing the corresponding FPRs and TPRs. Linear interpolation is applied between points. The ROC curve is non-decreasing, progressing from (0,0) (score threshold $s = 1$) to (1,1) ($s = 0$). A random classifier lies along the diagonal $y = x$, while curves bending toward the top-left indicate better performance. The area under the ROC curve (AUC-ROC) is a common summary metric, computed via the trapezoidal rule.

The **PR Curve** plots precision against recall. The x-axis represents recall, equals TPR, while the y-axis represents precision, defined as $P = TP/(TP+FP)$. Unlike the ROC curve, the PR curve is sensitive to class imbalance since precision depends on both classes. This property makes the PR curve more informative when positive instances are rare [36]. The PR curve is non-monotonic, spanning from (0,1) ($s = 1$) to (1, n^+/n) ($s = 0$), as precision may decrease with increasing recall. A random classifier yields a horizontal line at $y = n^+/n$, while curves bending toward the top-right are better.

There exists a one-to-one correspondence between ROC and PR curves [15]. In ROC space, linear interpolation is the standard

method for connecting points. However, in PR space, linear interpolation can lead to overestimation of performance because precision is a non-linear function of TPR and FPR. In *Scikit-learn* [32], right-end constant interpolation is used to ensure consistency with average precision, which is a scalar summary of the PR curve and differs from the trapezoidal rule used for AUC-ROC.

3.2 Differential Privacy and Federated Quantiles

Differential Privacy (DP) [16] guarantees that the output of a function f is insensitive to the inclusion or exclusion of any single data point. Formally, a mechanism $\mathcal{M}$ satisfies ϵ -DP if for any two neighboring datasets D and D' differing by one record, the output distributions $\mathcal{M}(D)$ and $\mathcal{M}(D')$ are close, with probability ratio bounded by e^ϵ , i.e., $\Pr[\mathcal{M}(D) \in S] \leq e^\epsilon \Pr[\mathcal{M}(D') \in S]$ for all subsets S of the output space. A smaller ϵ implies stronger privacy but requires more noise, which may lead to less accurate results. The noise magnitude also depends on the sensitivity of the function, defined as $\Delta f = \max_{D, D'} \|f(D) - f(D')\|$. One common way to achieve DP is to add Laplace noise to the output of f , where the noise scale is proportional to $\Delta f/\epsilon$, e.g., $\text{Lap}(0, \Delta f/\epsilon)$. When an ϵ -DP mechanism is applied T times on overlapping inputs, *composition bounds* show that the privacy guarantee varies as $T\epsilon$.

Secure Aggregation (SA) [8] is a cryptographic protocol that enables a server to compute the aggregate (sum) of client inputs without revealing any individual contribution. Each client masks its local input using random values that cancel out only after aggregation. The server receives only the masked inputs and can recover the total sum, while learning nothing about any client's individual data beyond what is implied by the final aggregate. Modern SA implementations can tolerate client dropouts and communication failures, making them practical for large-scale federated systems. SA has become a standard building block for federated learning and privacy-preserving distributed analytics, where many computations, such as vector sums, averages, histograms, and other aggregate statistics, can be expressed as additive operations.

Distributed DP (DDP) [22] allows clients to collaboratively add noise to ensure privacy. For example, each client adds local noise for which the scale is proportional to $1/m$ of the total noise required in the centralized setting, where m is the number of clients. Therefore, after aggregating noise from all clients, the total noise scale is the same as in the centralized setting. To achieve ϵ -DDP, continuous noise can be sampled from the difference of two independent Gamma distributions, while discrete noise can be sampled from the difference of two independent Pólya distributions. This also extends to (ϵ, δ) -DP mechanisms such as the Skellam Mechanism [1] and the Distributed Discrete Gaussian Mechanism [24]. Generally, DDP can be achieved with Secure Aggregation [8].

Federated Quantile Estimation can be achieved using histogram-based methods. Each client bins local data and transmits a noisy histogram. The server aggregates these to form a global histogram from which quantiles are estimated. Hierarchical histograms [12, 23, 34] are preferred over flat histograms, as they provide improved accuracy and enable consistency enforcement via post-processing, making them well-suited for federated and differentially private settings. Specifically, a hierarchical histogram of height h divides the score range into b^i bins at level i , where b is the branching factor.

Each client splits its privacy budget ϵ equally across the h layers and adds independent noise to each bin. The server aggregates histograms via simple summation. While there are other approaches to DP quantile estimation (e.g., [21]), these are not well-suited to the federated setting; hence we adopt the histogram approach.

3.3 Monotone Interpolation

Linear interpolation is a simple and efficient method for joining up a monotone series of points but may lack smoothness and fail to capture underlying trends of the data. Polynomial interpolation improves smoothness but may introduce oscillations, potentially violating monotonicity. Piecewise-cubic Hermite interpolation (PCHIP) [19, 20, 30] provides a balance: it constructs piecewise cubics that preserve monotonicity and avoid overshoot via carefully selected derivatives at each point. Although higher-order methods like quintic interpolation [14, 27] can offer smoother results, they are less efficient. In this work, we make use of both linear and PCHIP interpolation for reconstructing curves from quantiles.

4 FEDERATED ROC AND PR CURVES

In this section, we present our method for approximating ROC and PR curves in the federated setting. We first describe how to construct curves from score quantiles and then explain how to estimate those quantiles using histograms under both SA and DDP settings. We assume that all parties are honest-but-curious; that is, the server follows the protocol but may attempt to infer sensitive information from the received data from clients.

4.1 Curve Approximation via Quantiles

Constructing ROC and PR curves uses prediction scores and corresponding class labels. In federated settings, these raw values cannot be shared due to privacy constraints. Instead, since ROC and PR curves are determined by the distributions of scores for positive and negative classes, we will estimate the empirical cumulative distribution function (ECDF) $\Phi(s)$ of prediction scores using quantiles, computed separately for the positive and negative classes. Figure 1 shows a concrete example where the ECDF is divided into five evenly spaced quantile intervals, each with a width of 0.2. PCHIP interpolation is used to estimate the ECDF between quantile points. We can then compute TPR/FPR (for ROC) and Precision/Recall (for PR) at arbitrary thresholds using the interpolated ECDFs.

For the ROC curve, the FPR depends only on the negative score distribution, while TPR depends only on the positive score distribution. They are computed as:

$$F(s) = \Pr[S > s \mid y = 0] = 1 - \Phi^-(s), \quad (1)$$

$$T(s) = \Pr[S > s \mid y = 1] = 1 - \Phi^+(s). \quad (2)$$

For the PR curve, the recall equals $T(s)$, and precision $P(s)$ is computed using $T(s)$ and $F(s)$ as:

$$P(s) = \frac{T(s)n^+}{T(s)n^+ + F(s)n^-}. \quad (3)$$

Figure 2 shows the resulting approximated ROC and PR curves derived from the estimated ECDFs. In this example, the classifier used is XGBoost [9] on the Adult dataset [3], although our approach is agnostic to the classifier and data used. For exposition, we assume

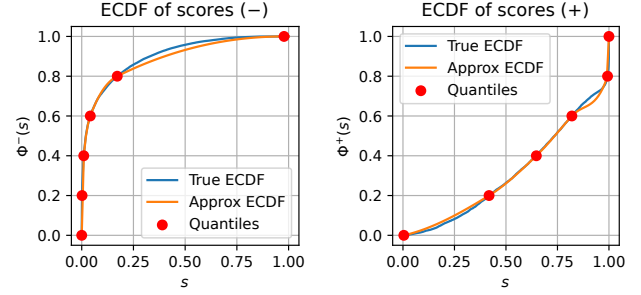


Figure 1: ECDF approximation using six quantiles for prediction scores of negative and positive classes.

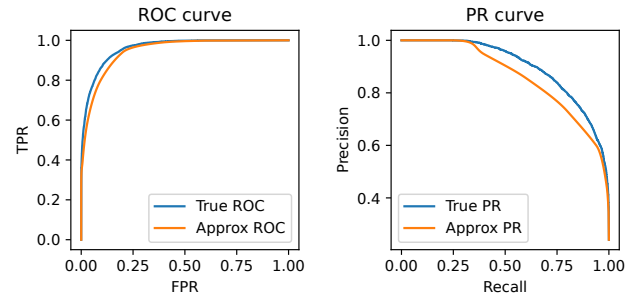


Figure 2: Approximating ROC and PR curves constructed from approximated ECDFs.

that prediction scores lie in $[0, 1]$ (and so may be thought of as probabilities), but the approach applies to arbitrary score ranges.

4.2 Quantile Estimation via Histograms

Quantiles are estimated via hierarchical histograms of height h . At each level i ($1 \leq i \leq h$), the prediction score range is divided into b^i equal-width bins, where b is the branching factor. Each client builds separate histograms for positive and negative examples and sends them to the server. Under DP, each client splits its privacy budget ϵ equally across the h layers and adds independent noise to each bin [23, 34]. The histograms are aggregated at the server, enabling quantile estimation without accessing raw scores. The server uses this information to compute a (noisy) ECDF of the score distribution, and then derives the FPR and TPR values to plot the ROC and PR curves. The full protocol is described in Algorithm 1. Highlighting indicates communication between clients and the server, i.e., the histograms aggregated via secure aggregation.

Privacy Guarantee. Due to the addition of suitably scaled private noise, the histograms received by the server achieve ϵ -differential privacy (Section 3.2). Moreover, since this is the only access to sensitive client data, the resulting algorithm meets the ϵ -DP guarantee. If applied repeatedly T times, the guarantee is $T\epsilon$ -DP (Section 3.2).

Data Heterogeneity. A positive property of the histogram-based quantile estimation is that it is agnostic to data heterogeneity and

Algorithm 1 Federated Construction of ROC and PR Curves

Client Part:

- 1: **Input:** Prediction scores for positives s^+ and negatives s^-
- 2: Build histograms h^+ , h^- over s^+ , s^-
- 3: **if** Differential Privacy is enabled **then**
- 4: Add i.i.d. privacy noise to every histogram bin
- 5: Send histograms h^+ and h^- to the server

Server Part:

- 1: Aggregate client histograms to obtain global H^+ and H^-
 - 2: **if** Differential Privacy is enabled **then**
 - 3: Apply post-processing to noisy H^+ and H^-
 - 4: Extract total positive/negative counts $\hat{n}^+$, $\hat{n}^-$ from H^+ , H^-
 - 5: Extract a set of quantiles $\hat{q}^+$, $\hat{q}^-$ from H^+ , H^-
 - 6: Interpolate $\hat{q}^+$, $\hat{q}^-$ to build ECDFs $\hat{\Phi}^+(s)$, $\hat{\Phi}^-(s)$
 - 7: **for** each evaluation threshold s **do**
 - 8: $\hat{F}(s) \leftarrow 1 - \hat{\Phi}^-(s)$ ▷ FPR
 - 9: $\hat{T}(s) \leftarrow 1 - \hat{\Phi}^+(s)$ ▷ TPR / Recall
 - 10: $\hat{P}(s) \leftarrow \frac{\hat{T}(s) \hat{n}^+}{\hat{T}(s) \hat{n}^+ + \hat{F}(s) \hat{n}^-}$ ▷ Precision
 - 11: Plot ROC curve: $(\hat{F}(s), \hat{T}(s))$; Plot PR curve: $(\hat{T}(s), \hat{P}(s))$
-

skew. Since each client locally bins scores and the server aggregates counts additively, the resulting global histogram matches the centralized case, regardless of how data is distributed across clients.

Client Dropout. The secure aggregation protocol is designed to tolerate client dropout, as the server can still compute the aggregated result provided the number of participating clients is above a certain threshold. This threshold represents the minimum number of surviving clients needed to reconstruct the secret masks.

Communication Cost. The communication cost is linear in the number of leaf bins: $O(b^h)$. Under SA, each client sends only the b^h leaf counts. Under DP, a client sends $b + b^2 + \dots + b^h = \frac{b(b^h - 1)}{b - 1}$ bins, which is also $O(b^h)$. To ensure accurate quantile estimation, we require $b^h > Q$, where Q is the number of quantiles. To balance communication and accuracy, we set $h = \lceil \log_b Q \rceil + c$, where c is a small constant (typically 2 or 3). This keeps the communication cost at $O(Q)$, which is much smaller than the dataset size. In our experiments, we use $b = 2$ and $c = 2$. When $Q = 1024$, this is sufficient to ensure low Area Error (Definition 5.1). Under this setting, each client transmits approximately 2^{13} bins per histogram, totaling around 8K integers (32-bit), i.e., 32K Bytes. Without privacy constraints, there is only a single communication round; using Secure Aggregation [5, 8], this increases to four rounds or fewer.

5 ANALYSIS OF AREA ERROR

In this section, we define the Area Error (AE) metric for quantifying the discrepancy between true and approximated ROC and PR curves, and provide theoretical bounds on AE when using quantiles for curve approximation. We first analyze the case of exact quantiles, and then extend the analysis to the federated setting with approximate quantiles and privacy noise.

Shorthand Notation. Throughout this section, we adopt the following shorthand notation. For function composition involving T , F and their inverses, we write $T(f) := T(F^{-1}(f))$ to denote TPR as a function of FPR f , and $F(t) := F(T^{-1}(t))$ to denote FPR as a function of TPR t . Similarly, $\hat{T}(f) := \hat{T}(\hat{F}^{-1}(f))$ and $\hat{F}(t) := \hat{F}(\hat{T}^{-1}(t))$ denote their estimated counterparts. Other mixed compositions, such as $T(\hat{F}^{-1}(f))$ or $\hat{T}(F^{-1}(f))$, are stated explicitly when needed. For PR curves, we use $P(t) := P(T^{-1}(t))$ for precision as a function of recall, and $\hat{P}(t) := \hat{P}(\hat{T}^{-1}(t))$ for the estimated precision.

5.1 Area Error Definition

We use Area Error (AE) to quantify the discrepancy between the true and approximated ROC or PR curves, similar in spirit to the notion of the L_1 metric from Clemenccon and Vayatis [11]. Formally:

Definition 5.1 (AE). The Area Error is defined as the integral of the absolute difference between the true and estimated curves over the full domain, i.e., FPR for ROC and recall for PR from 0 to 1:

$$\text{AE}_{\text{ROC}} = \int_0^1 |T(f) - \hat{T}(f)| df, \quad \text{AE}_{\text{PR}} = \int_0^1 |P(t) - \hat{P}(t)| dt. \quad (4)$$

Here, $T(f)$ and $\hat{T}(f)$ denote the true and estimated TPR at FPR f , while $P(t)$ and $\hat{P}(t)$ denote the true and estimated precision at recall t (since recall is equivalent to TPR).

AE is a stricter error measure than absolute AUC difference (which computes $|\int_0^1 T(f) df - \int_0^1 \hat{T}(f) df|$ for ROC). A small AE implies a small AUC error, but not vice versa. This is because AUC aggregates performance into a single scalar, while AE captures curve discrepancies over the full domain. Another weak alternative is the Point Error [2], which averages the absolute or squared difference at the score thresholds. However, this neglects cumulative errors and lacks robustness across the full curve.

5.2 Bounding Area Error Using Exact Quantiles

We first analyze the case where **exact** quantiles are used for curve approximation. This serves as a theoretical baseline for understanding the error introduced by quantile-based approximation, without the additional complications of estimation error and privacy noise. We consider the case where quantiles are evenly spaced. Let f and t denote the exact FPR and TPR at arbitrary score threshold s , and let $\hat{f}$ and $\hat{t}$ be their estimations. If Q^- and Q^+ exact quantiles are used for the negative and positive classes respectively, then:

$$|f - \hat{f}| \leq \frac{1}{2(Q^- - 1)}, \quad |t - \hat{t}| \leq \frac{1}{2(Q^+ - 1)}. \quad (5)$$

The bound arises because the maximum quantile error corresponds to half the width of a quantile interval. Using this, we can now bound the AE for the ROC curve.

THEOREM 5.2 (ROC-AE-EXACTQ). *If Q exact quantiles are used for both positive and negative examples, then the Area Error between the true and estimated ROC curves is bounded by $O(1/Q)$.*

PROOF. The high-level idea is that when querying the distribution of the negative class at the exact quantile points, the false positive rates (FPRs) are found precisely; when probing elsewhere in the distribution, the FPR absolute error is guaranteed to be at

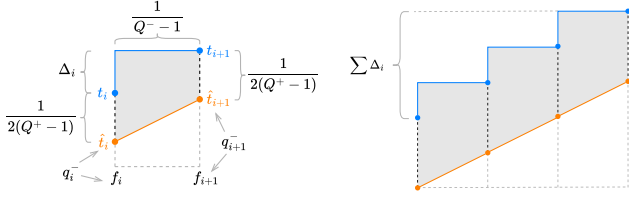


Figure 3: Worst-case area error with quantiles on negative examples only.

most half the quantile width (the distance between successive quantile measurements). The same holds for true positive rates (TPRs) at quantile points for the positive class. As a result, we evenly divide the ROC space into $Q - 1$ vertical strips, where each strip contributes an error of at most $O(1/Q^2 + \Delta_i/Q)$, with the first term accounting for the bounded deviation in TPR/FPR and the second term corresponding to a region of uncertain area. Since the ROC space is bounded and $\sum_i \Delta_i \leq 1$, the overall Area Error is bounded by $O(1/Q)$ when summing over all $Q - 1$ strips. In the remainder of this proof, we formally establish this result.

Since TPRs and FPRs are computed independently from separate class distributions, we consider two cases, expanded below: (1) using Q^- exact quantiles for negative examples to precisely determine FPRs, while TPRs remain approximated; (2) additionally using Q^+ exact quantiles for positive examples to tighten the TPR estimates. To formalize this bound, we analyze worst-case error by considering how TPR and FPR may deviate between quantile points.

(1) *Using Q^- Quantiles for Negative Examples.* Dividing the FPR axis using Q^- quantiles results in $Q^- - 1$ vertical strips. For any pair of consecutive negative quantiles, q_i^- and q_{i+1}^- , the corresponding FPR values $f_i = F(q_i^-)$ and $f_{i+1} = F(q_{i+1}^-)$ are accurately estimated. Let $t_i = T(q_i^-)$ and $\hat{t}_i = \hat{T}(q_i^-)$ denote the true and estimated TPRs, respectively. In order to show a bound on the AE, we consider the (deterministic) worst possible behavior of the ROC curve.

To maximize AE under the constraints of the measured quantiles, we can set $t_i - \hat{t}_i = t_{i+1} - \hat{t}_{i+1} = \frac{1}{2(Q^+ - 1)}$, the largest possible gap while meeting the quantile definition. Define $\Delta_i = t_{i+1} - t_i$ as the change in TPRs between consecutive points. This scenario is illustrated in the left panel of Figure 3: it corresponds to the TPR suddenly “leaping” up to t_{i+1} at f_i and remaining constant until f_{i+1} (blue line in the figure). The opposite extreme, the TPR remaining steady from f_i to f_{i+1} then leaping up to t_{i+1} , would also contribute a symmetric (negative) area error (similar to the dashed grey line), so it suffices to focus on the case of positive area error.

The total AE is computed as the sum of the areas of trapezoids, as illustrated in the right panel of Figure 3. We apply the standard trapezoid area formula, $\frac{1}{2}(a + b)h$, where a and b are the lengths of parallel sides, and h is the (perpendicular) ‘height’. Since $\sum_i \Delta_i \leq 1$, the worst-case error is bounded by $O(1/Q)$:

$$\text{AE}_{\text{ROC}} \leq \sum_{i=1}^{Q^- - 1} \frac{\frac{1}{2(Q^+ - 1)} + \frac{1}{2(Q^+ - 1)} + \Delta_i}{2} \cdot \frac{1}{Q^- - 1} \quad (6)$$

$$= \sum_{i=1}^{Q^- - 1} \frac{1}{2(Q^+ - 1)(Q^- - 1)} + \frac{\Delta_i}{2(Q^- - 1)} \quad (7)$$

$$\leq \frac{1}{2(Q^+ - 1)} + \frac{1}{2(Q^- - 1)} = O(1/Q) \quad (8)$$

The proof for the case of (2) adding Q^+ quantiles for positive examples is analogous. It respects the same bound but is more complicated; we provide the full details in the technical report [40]. $\square$

The corresponding area error bound for the PR curve given **exact** quantiles follows a similar outline, but has more details due to the non-linear relationship between precision and TPR/FPR. Moreover, since obtaining exact quantiles is often impractical in practice, we focus our study on reporting PR curves using approximate quantiles in the federated setting, which is more relevant for real-world applications. In Section 5.3, we introduce the notion of well-behaved score distributions and establish lemmas that bound the error in TPR and FPR. These lemmas are crucial for analyzing the Area Error in the federated setting with approximate quantiles and privacy noise, which we discuss in Section 5.4 and Section 5.5.

5.3 Well-Behaved Score Distributions and Lemmas for Approximate Quantiles

In the federated setting, approximate quantiles introduce quantile estimation errors bounded by the width of the finest histogram bin. A key challenge is that these errors can propagate to large errors in TPR/FPR when score distributions may exhibit pathological behavior (e.g., spikes where many examples share the same score). To address this, we consider a Lipschitz smoothness condition on score distributions to control how quickly ECDFs may change. This condition, formalized below as “well-behaved” score distributions, is similar to assumptions in prior work [13] and lets us bound Area Error by relating quantile estimation errors to errors in T and F .

Definition 5.3 (Well-behaved score distribution). Let $\Phi^+(s)$ and $\Phi^-(s)$ denote the cumulative distribution functions of the score distribution for positive and negative examples, respectively. We say the distributions are ℓ -well-behaved if both are ℓ -Lipschitz:

$$|\Phi^+(s_1) - \Phi^+(s_2)| \leq \ell|s_1 - s_2| \text{ and } |\Phi^-(s_1) - \Phi^-(s_2)| \leq \ell|s_1 - s_2| \quad (9)$$

With well-behaved score distributions, we can bound the absolute error of the TPR and FPR using the quantile estimation error α . These bounds allow us to prove the Area Error of the ROC and PR curves in the federated setting. The following lemmas formalize these bounds. Lemma 5.4 bounds the error in TPR and FPR at a given score s , which is essentially the ECDFs of the score distributions, since the variable is the score s ; Lemma 5.5 bounds the error in TPR at a given FPR f or FPR at a given TPR t , which is more directly relevant for analyzing the Area Error of the curves. In particular, $T(f) = T(F^{-1}(f))$ yields the ROC curve, and the variable is the FPR f . This bound is used in Theorem 5.6, the Area Error analysis of the ROC curve. Although $F(t) = F(T^{-1}(t))$ is not the PR curve, it is used in Theorem 5.7, the Area Error analysis of the PR curve, and the variable is the TPR t .

Remark. The well-behaved condition is used in our subsequent theoretical analysis to ensure that quantile estimation errors translate to bounded errors in TPR and FPR. We note that this condition may not hold universally in practice, particularly for score distributions with spikes. Despite this, our empirical results demonstrate robustness across diverse datasets. For instance, in Figure 4, the

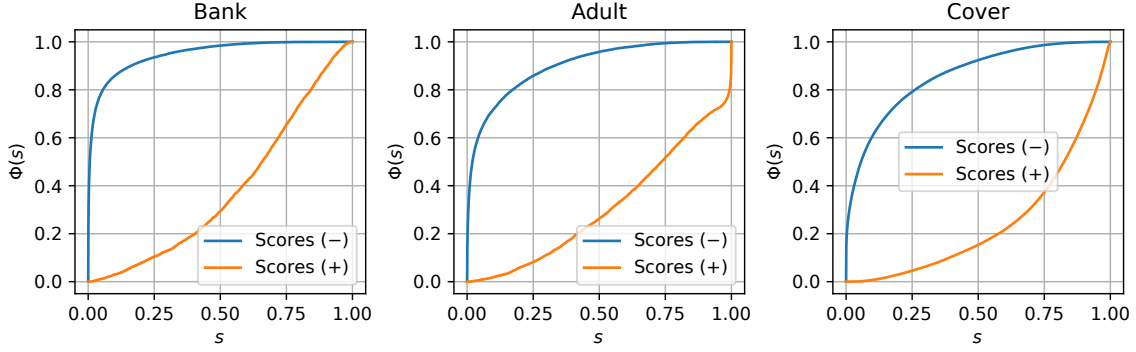


Figure 4: ECDF of prediction scores (XGBoost).

Adult dataset exhibits clustering at extremes: for negative examples, 40% of scores fall below 0.01 and 80% below 0.17; for positive examples, 20% exceed 0.993 and 40% exceed 0.82. This clustering violates the Lipschitz condition locally, yet our method still achieves strong empirical accuracy (Section 6). This robustness may arise because the clustering is only confined to narrow score intervals and has a limited effect on error of the overall curve. Thus, while the well-behaved assumption is used to prove our theoretical guarantees, the method performs reliably in practice even when this condition is not met. Meanwhile, the privacy guarantees need no such assumption, and hold regardless.

LEMMA 5.4. *Using the histogram approach for quantile estimation in the federated setting, let the additive error of quantile estimation be bounded by α , assuming the score distributions are $O(1)$ -well-behaved, then we have the following bounds:*

$$|T(s) - \hat{T}(s)| \leq O(\alpha) \quad \text{and} \quad |F(s) - \hat{F}(s)| \leq O(\alpha) \quad (10)$$

PROOF. Since the histogram approach guarantees that the quantile estimation error is bounded:

$$|T^{-1}(t) - \hat{T}^{-1}(t)| \leq \alpha \quad \text{and} \quad |F^{-1}(f) - \hat{F}^{-1}(f)| \leq \alpha \quad (11)$$

Since $T(s) = 1 - \Phi^+(s)$ and $F(s) = 1 - \Phi^-(s)$, by $O(1)$ -well-behaved distributions, we have:

$$|T(s_1) - T(s_2)| \leq O(|s_1 - s_2|) \quad \text{and} \quad |F(s_1) - F(s_2)| \leq O(|s_1 - s_2|) \quad (12)$$

Then the absolute error can be bounded as follows:

$$|T(s) - \hat{T}(s)| = |T(s) - T(T^{-1}(\hat{T}(s)))| \quad (13)$$

$$\leq O(|s - T^{-1}(\hat{T}(s))|) \quad (14)$$

$$= O(|\hat{T}^{-1}(\hat{T}(s)) - T^{-1}(\hat{T}(s))|) \leq O(\alpha) \quad (15)$$

Similarly, we have $|F(s) - \hat{F}(s)| \leq O(\alpha)$. $\square$

Next, we can use Lemma 5.4 to bound the error of $T(F^{-1}(f))$ and $F(T^{-1}(t))$, which we write as $T(f)$ and $F(t)$ for shorthand.

LEMMA 5.5. *Using the histogram approach for quantile estimation in the federated setting, let the additive error of quantile estimation be bounded by α , and assuming the score distributions are $O(1)$ -well-behaved, then we have the following bounds:*

$$|T(f) - \hat{T}(f)| \leq O(\alpha) \quad \text{and} \quad |F(t) - \hat{F}(t)| \leq O(\alpha) \quad (16)$$

PROOF. By $O(1)$ -well-behaved distributions and additive error of quantile estimation, we have:

$$|T(F^{-1}(f)) - T(\hat{F}^{-1}(f))| \leq O(|F^{-1}(f) - \hat{F}^{-1}(f)|) \leq O(\alpha) \quad (17)$$

Now consider the absolute error:

$$\begin{aligned} |T(f) - \hat{T}(f)| &= |T(F^{-1}(f)) - \hat{T}(\hat{F}^{-1}(f))| \end{aligned} \quad (18)$$

$$= |T(F^{-1}(f)) - T(\hat{F}^{-1}(f)) + T(\hat{F}^{-1}(f)) - \hat{T}(\hat{F}^{-1}(f))| \quad (19)$$

$$\leq |T(F^{-1}(f)) - T(\hat{F}^{-1}(f))| + |T(\hat{F}^{-1}(f)) - \hat{T}(\hat{F}^{-1}(f))| \quad (20)$$

$$\leq O(\alpha) + O(\alpha) = O(\alpha) \quad (21)$$

Similarly, we have $|F(t) - \hat{F}(t)| \leq O(\alpha)$. $\square$

Remark. We note that our analysis (here, and subsequently) only requires that the interpolation of the ECDF is monotone. Hence, both linear interpolation and PCHIP satisfy this property. Consequently, they yield the same asymptotic Area Error bounds. Our recommendation is for PCHIP, as we observe that this leads to smoother curves and improved area error in practice.

5.4 Bounding the Area Error Using Approximate Quantiles

Next, we analyze the Area Error of the ROC and PR curves when using approximate quantiles. In the federated setting, we use the histogram approach to estimate quantiles, which introduces an additive error α in quantile estimation. This error propagates to errors in TPR and FPR, which in turn affect the Area Error of the ROC and PR curves. Using the bounds established in Lemma 5.5, we can derive bounds on the Area Error for both curves under the assumption of well-behaved score distributions.

THEOREM 5.6 (ROC-AE-APPROXQ). *If Q approximate quantiles are used for both positive and negative examples, under the assumption of $O(1)$ -well-behaved score distributions, the Area Error between the true and estimated ROC curves is bounded by $O(1/Q)$.*

PROOF. Let the quantile estimation error be $\alpha = O(1/Q)$, which is the width of the finest histogram bin. By well-behaved score

distributions and Lemma 5.5, we have $|T(f) - \hat{T}(f)| \leq \alpha$. Therefore, the total area error is the integral of the absolute error:

$$\text{AE}_{\text{ROC}} = \int_0^1 |T(f) - \hat{T}(f)| df \leq \alpha = O(1/Q) \quad (22)$$

□

The analysis for the PR curve is similar in outline, but requires additional care. Precision depends on both $T(s)$ and $F(s)$ via Equation (3), and is also affected by class imbalance [39]. In the worst case (when t is close to 0), the absolute error $|P(t) - \hat{P}(t)|$ can be close to 1. When t is close to 1, the absolute error is as high as $O(1/Q)$. However, we can bound the AE by integrating the absolute error over the entire range of TPR (Recall).

THEOREM 5.7 (PR-AE-APPROXQ). *If Q approximate quantiles are used for both positive and negative examples, under the assumption of $O(1)$ -well-behaved score distributions and mild class imbalance ($r = n^+/n^- \geq 0.1$), the Area Error between the true and estimated PR curves is bounded by $\tilde{O}(1/Q)$. Under extreme class imbalance ($r \ll 1$), the AE is bounded by $\tilde{O}(1/(Qr))$.*

PROOF. AE is the integral of $|P(t) - \hat{P}(t)|$ over $t \in [0, 1]$. By bounding $|P - \hat{P}|$ as a function of t , and integrating, we obtain a worst-case Area Error. However, under extreme class imbalance, the bound is also dependent on the ratio r . See full details below.

Since precision is dependent on TPR, we can bound the absolute error in precision $|P(t) - \hat{P}(t)|$ as a function of t , then compute the area error by integrating over the TPR range. Given Equation (3), we reformulate the precision P as a function of t :

$$P(t) = \frac{t \cdot n^+}{t \cdot n^+ + F(t) \cdot n^-} \quad (23)$$

We first analyze the perfectly balanced setting (Case 1) and then extend to imbalanced settings (Case 2).

Case 1: Perfectly Balanced Class Distribution ($r = 1$). To build an intuition, we first consider the case where the number of positive examples is equal to the number of negative examples. This gives the pattern of the proof that we subsequently extend to arbitrary ratios of positive to negative examples. Since $n^+ = n^-$, we have:

$$P(t) = \frac{t}{t + F(t)} \quad \text{and} \quad \hat{P}(t) = \frac{t}{t + \hat{F}(t)} \quad (24)$$

Let the absolute error in FPR at given t be bounded, i.e., $|F(t) - \hat{F}(t)| \leq \alpha$, where $\alpha = O(1/Q)$ (Lemma 5.5). To maximize the error $P(t) - \hat{P}(t)$, we minimize $\hat{P}(t)$ by maximizing $\hat{F}(t)$. We set $\hat{F}(t) = F(t) + \alpha$. This gives:

$$P(t) - \hat{P}(t) \leq \frac{t}{t + F(t)} - \frac{t}{t + F(t) + \alpha} \quad (25)$$

$$= \frac{t \cdot \alpha}{(t + F(t))(t + F(t) + \alpha)} \quad (26)$$

For the other direction, to maximize $\hat{P}(t) - P(t)$, we maximize $\hat{P}(t)$ by minimizing $\hat{F}(t)$. We set $\hat{F}(t) = F(t) - \alpha$. This gives:

$$\hat{P}(t) - P(t) \leq \frac{t}{t + F(t) - \alpha} - \frac{t}{t + F(t)} \quad (27)$$

$$= \frac{t \cdot \alpha}{(t + F(t))(t + F(t) - \alpha)} \quad (28)$$

Therefore, the absolute error in precision is bounded by the larger of (26) and (28), which is (28):

$$|\hat{P}(t) - P(t)| \leq \frac{t \cdot \alpha}{(t + F(t))(t + F(t) - \alpha)} \quad (29)$$

$$= \frac{\alpha}{(1 + F(t)/t)(t + F(t) - \alpha)} \quad (30)$$

$$\leq \frac{\alpha}{t + F(t) - \alpha} \leq \frac{\alpha}{t - \alpha} \quad (31)$$

Note that the precision is in the range of $[0, 1]$, and the maximum absolute error is 1. Therefore, for $t \leq 2\alpha$, the absolute error is bounded by 1; for $t > 2\alpha$, the absolute error is bounded by $\alpha/(t - \alpha)$, which is smaller than 1. Integrating this bound over t yields the total area error:

$$\text{AE}_{\text{PR}} = \int_0^1 |P(t) - \hat{P}(t)| dt \quad (32)$$

$$\leq 2\alpha + \int_{2\alpha}^1 \frac{\alpha}{t - \alpha} dt \quad (33)$$

$$= 2\alpha + \alpha \log(1/\alpha - 1) \quad (34)$$

$$\leq 2\alpha + \alpha \log(1/\alpha) \quad (35)$$

$$= O(\log Q/Q) = \tilde{O}(1/Q) \quad (36)$$

For the last term, since $\alpha \log(1/\alpha) = \log Q/Q$ and $\log Q$ grows much more slowly than Q , we may treat it as a constant, and think of this term as being dominated by $1/Q$. Pragmatically, setting $Q = 10^4$ gives $\text{AE} \approx 10^{-3}$, which is sufficient for most applications.

Case 2: Imbalanced Class Distribution ($r \neq 1$). Under the imbalanced class distribution, we consider the case to maximize $\hat{P}(t) - P(t)$ by maximizing $\hat{P}(t)$. For the other direction, the results are similar so we omit for brevity. The absolute error in precision is bounded by:

$$\hat{P}(t) - P(t) = \frac{t \cdot n^+}{t \cdot n^+ + \hat{F}(t) \cdot n^-} - \frac{t \cdot n^+}{t \cdot n^+ + F(t) \cdot n^-} \quad (37)$$

$$\leq \frac{t \cdot n^+}{t \cdot n^+ + (F(t) - \alpha) \cdot n^-} - \frac{t \cdot n^+}{t \cdot n^+ + F(t) \cdot n^-} \quad (38)$$

$$= \frac{n^+ n^- \cdot t}{t \cdot n^+ + F(t) \cdot n^-} \cdot \frac{\alpha}{t \cdot n^+ + F(t) \cdot n^- - \alpha \cdot n^-} \quad (39)$$

$$= \frac{n^+ n^-}{n^+ + (F(t)/t) \cdot n^-} \cdot \frac{\alpha}{t \cdot n^+ + F(t) \cdot n^- - \alpha \cdot n^-} \quad (40)$$

Let $r = n^+/n^-$. We distinguish two sub-cases:

Subcase (a) ($r > 1$): The absolute error is bounded by:

$$|\hat{P}(t) - P(t)| \leq \frac{n^-}{n^+ + (F(t)/t) \cdot n^-} \cdot \frac{n^+ \cdot \alpha}{t \cdot n^+ + F(t) \cdot n^- - \alpha \cdot n^-} \quad (41)$$

$$= \frac{1}{r + F(t)/t} \cdot \frac{\alpha}{t + F(t)/r - \alpha/r} \quad (42)$$

$$\leq \frac{1}{r} \cdot \frac{\alpha}{t - \alpha/r} \leq \frac{\alpha}{t - \alpha/r} \leq \frac{\alpha}{t - \alpha} \quad (43)$$

This is the same as the perfectly balanced class distribution, the area error is bounded by $\tilde{O}(1/Q)$.

Subcase (b) ($r < 1$): The absolute error is bounded by:

$$|\hat{P}(t) - P(t)| \leq \frac{n^+}{n^+ + (F(t)/t) \cdot n^-} \cdot \frac{n^- \cdot \alpha}{t \cdot n^+ + F(t) \cdot n^- - \alpha \cdot n^-} \quad (44)$$

$$= \frac{1}{1 + F(t)/(t \cdot r)} \cdot \frac{\alpha}{t \cdot r + F(t) - \alpha} \leq \frac{\alpha}{t \cdot r - \alpha} \quad (45)$$

The Area Error is bounded by:

$$\text{AE}_{\text{PR}} = \int_0^1 |P(t) - \hat{P}(t)| dt \quad (46)$$

$$\leq 2\alpha/r + \int_{2\alpha/r}^1 \frac{\alpha}{t \cdot r - \alpha} dt \quad (47)$$

$$= 2\alpha/r + \alpha/r \cdot \log(r/\alpha - 1) \quad (48)$$

$$\leq 2\alpha/r + \alpha/r \cdot \log(r/\alpha) \quad (49)$$

Therefore, provided that Q is sufficiently large relative to $1/r$ (so that $2\alpha/r < 1$), the Area Error is bounded by $\tilde{O}(\alpha/r) = \tilde{O}(1/(Qr))$. For mild class imbalance with $r \geq 0.1$, this reduces to $\tilde{O}(1/Q)$. $\square$

Next, we extend the analysis to the federated setting with security and privacy constraints, which could introduce additional sources of error from privacy noise.

5.5 Error Analysis in the Federated Setting with Security and Privacy

We analyze the Area Error separately for the Secure Aggregation (SA) and Distributed Differential Privacy (DDP) settings. In SA, the main source of error is from the approximate quantile estimation due to histogram binning. In DDP, there is additional error from the privacy noise added to the histogram counts. We combine these errors to derive bounds on the Area Error for both the ROC and PR curves under these settings.

THEOREM 5.8 (AE-SA). *Under Secure Aggregation, assuming $O(1)$ -well-behaved score distributions, the Area Error between the true and estimated curves is bounded by $O(1/Q)$ for the ROC curve and $\tilde{O}(1/Q)$ for the PR curve.*

PROOF. Under SA without privacy noise, the quantile estimation error is bounded by the histogram bin width $\alpha = O(1/Q)$. Applying Theorems 5.6 and 5.7 directly yields the bounds. $\square$

THEOREM 5.9 (AE-DDP). *Under Distributed Differential Privacy with budget ϵ , assuming $O(1)$ -well-behaved score distributions, the Area Error between the true and estimated curves is bounded by $\tilde{O}(1/Q + 1/(n^+ \epsilon) + 1/(n^- \epsilon))$ for both the ROC and PR curves, where n^+ and n^- are the number of positive and negative examples, respectively.*

PROOF. Under DDP, the privacy budget ϵ is split equally across h layers of the hierarchical histogram. Each bin incurs Laplace noise with scale h/ϵ , yielding variance $O(h^2/\epsilon^2)$ per bin.

For quantile estimation, we traverse down the histogram tree, selecting the child bin that contains the quantile at each layer. This traversal requires at most b bins per layer and h layers, accumulating total noise variance $O(bh^3/\epsilon^2)$. The absolute error of the quantile estimate is thus $O(\sqrt{bh^3}/(n^* \epsilon))$ when normalized by the number of examples n^* , where n^* is equal to n^+ or n^- for positive or negative examples, respectively. For typical settings with $b = 2$, $h = O(\log Q)$,

Table 1: Dataset statistics.

Datasets	# Row	# Col	Pos-to-Neg Ratio
Bank	45K	16	0.132
Adult	33K	14	0.317
Cover*	581K	54	0.574

*Class 1 is treated as positive; others as negative.

and $n^* \gg \sqrt{bh^3}$, the expected quantile estimation error is bounded by $\alpha = \tilde{O}(1/Q + 1/(n^* \epsilon))$. For simplicity, we denote $\alpha^+ = \tilde{O}(1/Q + 1/(n^+ \epsilon))$ and $\alpha^- = \tilde{O}(1/Q + 1/(n^- \epsilon))$ for the quantile estimation errors of positive and negative examples, respectively.

Plugging the quantile error α^+ and α^- into Eq. (11) in Lemma 5.4, we have $|T(s) - \hat{T}(s)| \leq O(\alpha^+)$ and $|F(s) - \hat{F}(s)| \leq O(\alpha^-)$. Meanwhile, Eq. (17) in Lemma 5.5 will be bounded by $O(\alpha^-)$. Therefore, Eq. (21) in Lemma 5.5 will be bounded by $O(\alpha^+ + \alpha^-)$, i.e., $|T(f) - \hat{T}(f)| \leq O(\alpha^+ + \alpha^-)$. Similarly, we have $|F(t) - \hat{F}(t)| \leq O(\alpha^+ + \alpha^-)$. Since Theorem 5.6 only uses the bound of $|T(f) - \hat{T}(f)|$ and Theorem 5.7 only uses the bound of $|F(t) - \hat{F}(t)|$, we can replace the quantile error α with $O(\alpha^+ + \alpha^-)$, which yields the Area Error bounds for the ROC and PR curves under DDP.

The class counts n^+ and n^- are estimated from the noisy histograms with absolute error $O(h/\epsilon) = O(\log Q/\epsilon)$. For typical datasets and privacy budgets (e.g., $\epsilon \approx 1$), this error is negligible relative to n^+ and n^- . For the ROC curve, these counts are not needed. For the PR curve, the error is dominated by the quantile estimation error term and does not change the overall bound. $\square$

6 EMPIRICAL EVALUATION

We evaluate our method on three datasets: Bank [31], Adult [3], and Cover [7]. To account for the sensitivity of class imbalance, we select datasets with varying positive-to-negative ratios r . Dataset statistics are summarized in Table 1. We evaluate score functions produced by two classifiers: XGBoost and Logistic Regression (see the Appendices of [40] for more results); recall that our results apply to any score function, so the nature of the underlying classifier is not a major factor. We set the branching factor $b = 2$ and the height $h = \lceil \log_2 Q \rceil + 2$. For differential privacy, we use distributed discrete Laplace noise and apply post-processing by default.

6.1 Initial Experiments

Performance Evaluation. We integrated our method into the Flower federated learning framework [6] and evaluated its performance using the SecAgg+ protocol [8] for secure histogram aggregation. We measured the client-side execution time and memory consumption while varying the number of histogram bins. The results are shown in Figure 5. Since SecAgg+ consists of four phases, we exclude the setup and key sharing phases and report only the cost of the histogram aggregation phase, which is the component introduced by our method. Consistent with the communication analysis in Section 4.2, our method incurs very little communication overhead (less than 1 MB per client), while the overall execution time and memory usage are dominated by the underlying secure aggregation protocol. Across all configurations, the client-side aggregation takes approximately 1.5 seconds and consumes around

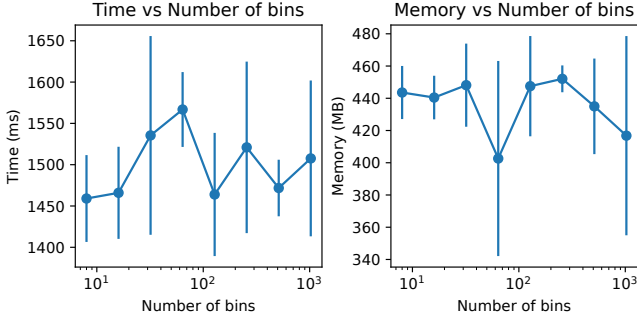


Figure 5: Time and memory performance.

450 MB of memory (mostly from the secure aggregation protocol and Flower itself), with both metrics remaining nearly constant as the number of histogram bins increases.

Interpolation Methods. We compare linear interpolation and piecewise cubic Hermite interpolation (PCHIP) for the interpolation step (Section 4.1). Results are shown in Figure 6, under Secure Aggregation (SA, no privacy noise), and Distributed Differential Privacy (DDP, $\epsilon = 1$ or 0.3). The PCHIP method consistently outperforms linear interpolation, though the margin is small in some cases. Since PCHIP yields smoother curves that more closely approximate the true ROC and PR curves, we recommend using it in practice. In Figure 6, we see that under SA the AE decreases as Q increases, whereas under DDP there is a plateau due to privacy noise, as predicted by our analysis (Theorem 5.9).

Privacy-Accuracy Trade-off. We evaluate performance under varying privacy levels $\epsilon \in \{0.1, 0.3, 1\}$. As shown in Figure 7, smaller ϵ results in higher area error. Nonetheless, the error remains low even under strong privacy guarantees ($\epsilon \leq 1$). Larger datasets such as Cover exhibit smaller error due to more accurate quantile estimation under DP (recall, our error bound scales with $1/(n\epsilon)$). Moreover, the error of the PR curve is slightly higher than that of the ROC curve, since the error includes a logarithmic factor in Q for the PR curve, while it is linear in $1/Q$ for the ROC curve (Theorems 5.6 and 5.7).

We also compare the Exact Quantiles (EQ), as discussed in Section 5.2 and Secure Aggregation of histograms (SA), as discussed in Section 5.5. In Figure 7, for Cover dataset the gap between the two is more noticeable, while it is negligible on datasets like Adult and Bank, especially when Q is large. This is because the quantile estimation error is smaller with finer histogram bins (larger Q).

6.2 Comparison with Prior Work

We compare our method with the range-query-based method of Barczewski et al. [2], which estimates TPR and FPR at N uniformly spaced thresholds using DP-noised range queries. In contrast, our method uses Q quantiles to adaptively select thresholds based on the score distribution. Uniform score thresholds may devote many queries to low-density regions while providing insufficient resolution where scores are concentrated. However, our quantile-based method ensures that consecutive representative points differ by approximately $1/Q$ in the ECDF, which is better for capturing the curve shape. To enforce monotonicity, their method introduces

smoothing variables v_i added to the TPR and FPR series. The values of v_i are optimized via ℓ_1 or ℓ_2 minimization. In contrast, our method uses post-processing to enforce consistency at the hierarchical histogram levels, which ensures that the estimated quantiles are monotonic. This approach is simpler and more efficient than solving an optimization problem, while still yielding accurate and consistent estimates. Moreover, we leverage smooth monotone interpolation (PCHIP) to construct ECDFs from the estimated quantiles. Our empirical ablation in Section 6.1 shows that PCHIP provides an additional but comparatively smaller improvement over linear interpolation. Their analysis bounds the *Squared Point Error* of the ECDF estimates at N uniformly spaced query thresholds by $O(\log^3 N / (n\epsilon)^2)$ ⁱⁱⁱ, which is $O(\log^{3/2} N / (n\epsilon))$ for the *Absolute Point Error* at those thresholds. But this guarantee does not control the error between consecutive thresholds. In contrast, under the assumption of $O(1)$ -well-behaved score distributions, our bound $O(1/Q + \log^{3/2} Q / (n\epsilon))$ applies at any score threshold (Lemma 5.4). When $Q \approx N$, the privacy-dependent terms have the same asymptotics; however, the guarantees are not directly comparable because our bound includes the term $1/Q$ and requires a Lipschitz condition, whereas their bound applies only at the selected thresholds.

Note that the point error fails to capture cumulative errors over the entire curve, thus we evaluate both methods using Area Error. Empirically, as shown in Figure 8, our method produces more stable area errors with increasing Q . In contrast, the Range method can show erratic AE results due to increased variance in each bin and added smoothing noise. The range method can perform better when $Q < 100$ (lower accuracy regime), especially on the Cover dataset, as it uses fixed thresholds for both TPR and FPR (for example, in Figure 8). Our approach, which estimates quantiles separately for positives and negatives, may diverge in threshold alignment, affecting accuracy at small Q . This discrepancy vanishes as Q increases.

6.3 Experiments on Class-Imbalanced Data

We evaluate the effect of class imbalance on PR curves by varying the positive-to-negative ratio r . Specifically, we fix the set of negative examples and randomly subsample the positive examples to achieve the desired ratio. These experiments use the Secure Aggregation method, with results shown in Figure 9.

In most cases, the AE increases as r decreases, which aligns with the theoretical area error bound of $\tilde{O}(1/(Qr))$. However, there are instances where the observed error is smaller than expected. This occurs because classifier performance degrades significantly on extremely imbalanced datasets, causing the PR curve to approximate that of a random classifier, as a horizontal line at $y = n^+/n \approx 0$. As this is nearly flat, even a poor approximation has low area error.

7 DISCUSSION

Extension to Multi-class Setting. Our method extends to multi-class classification by decomposing the problem into multiple binary tasks, such as one-vs-rest or one-vs-one schemes. Then, these binary tasks can be evaluated using the proposed quantile-based approach, with AE guarantees holding for each curve. For example, in a one-vs-rest scheme, we can compute a ROC and PR curve for

ⁱⁱⁱBarczewski et al. [2] omits the dependence on n in the statement of their error bound.

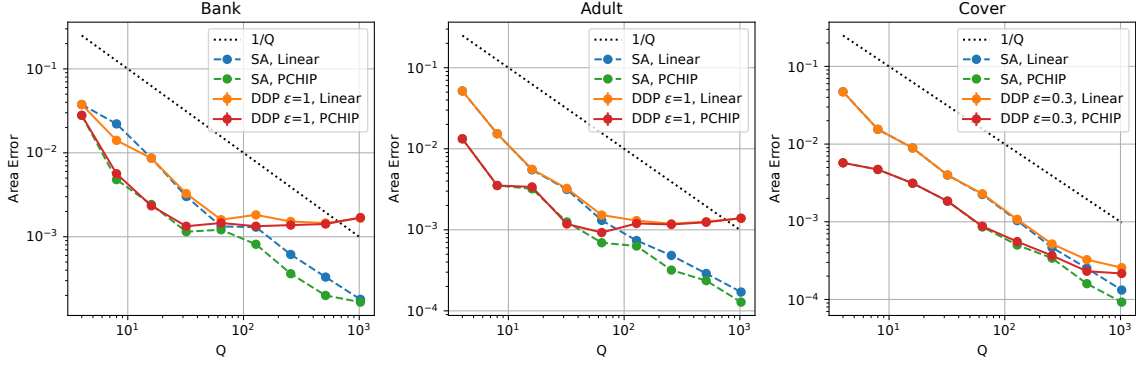


Figure 6: Interpolation method comparison (ROC, XGBoost).

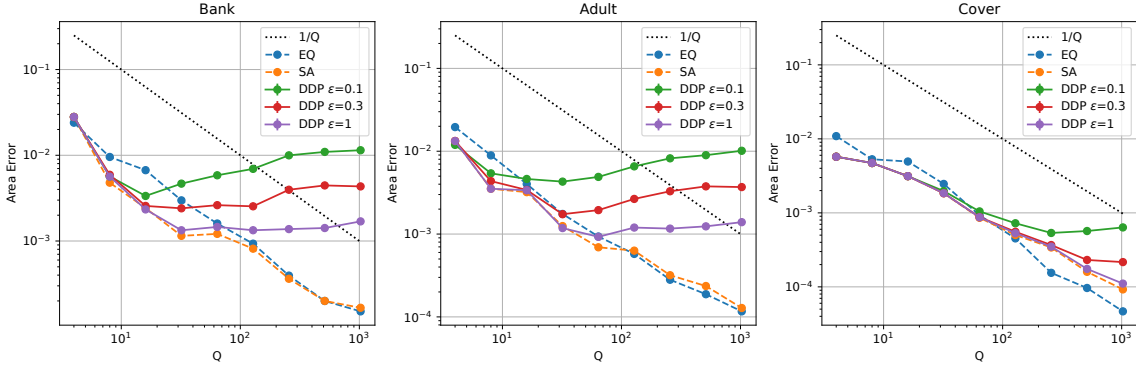


Figure 7: Effect of ε (ROC, XGBoost).

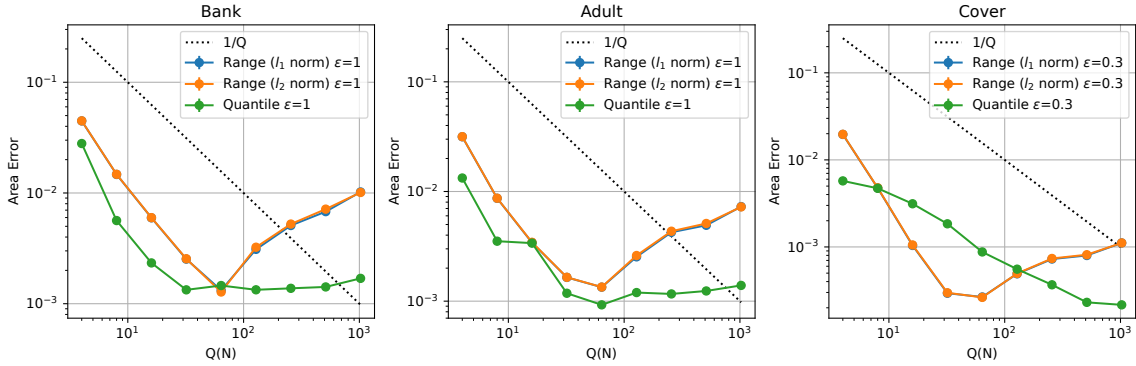


Figure 8: Comparison with range-query method (ROC, XGBoost).

each class against the rest, and then aggregate these curves to obtain multi-class curves such as macro- or weighted-average. The area error bounds carry over accordingly, ensuring that the overall evaluation remains accurate even in the multi-class setting.

Extension to Other Metrics. Our approach extends naturally to derived metrics such as AUC, where the area error directly bounds the absolute AUC difference by $O(1/Q)$, and the Mann-Whitney

U-statistic, of which AUC is a special case. It also applies to the Detection Error Tradeoff (DET) [28] curve, which plots the False Negative Rate (FNR) against the FPR on a log scale. Since $\text{FNR} = 1 - \text{TPR}$, we can compute DET similarly to ROC. Precision-Recall-Gain curves [18], which adjust precision and recall by the fraction of positive examples, are another promising direction for extending our approach. We leave their exploration for future work.

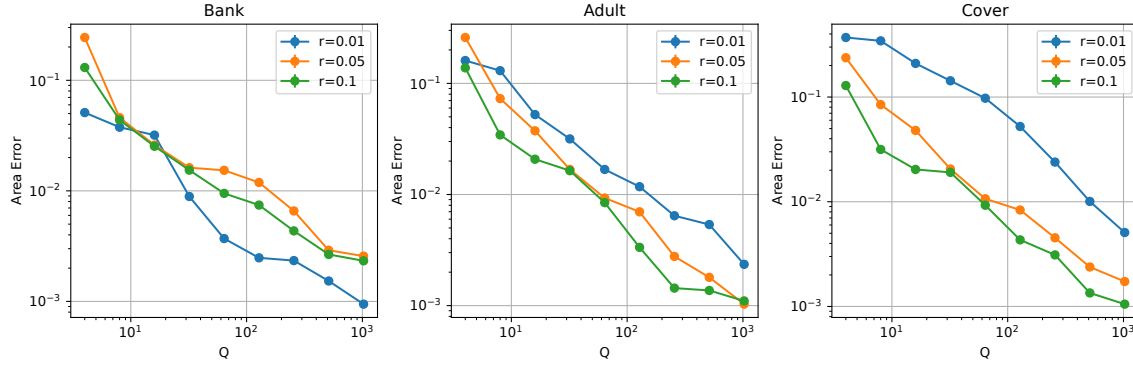


Figure 9: Effect of class imbalance on PR curve (XGBoost).

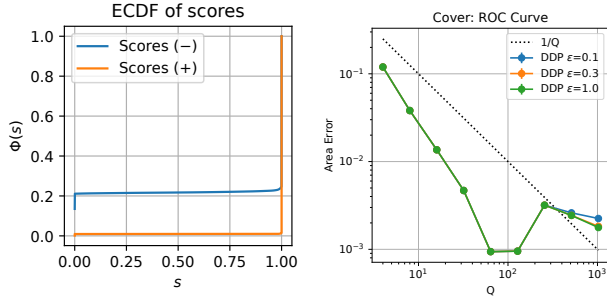


Figure 10: Results on Cover dataset, GaussianNB classifier.

$O(1)$ -well-behaved Score Distributions. Our analysis shows that this assumption is a sufficient condition for our theoretical analysis, ensuring that quantile estimation errors translate to bounded errors in the ECDFs. However, it does not hold universally in practice, particularly for score distributions with spikes as illustrated in Figure 4. Especially in the left panel of Figure 10, the score distribution is highly concentrated around 0 and 1. Despite this, our empirical results demonstrate robustness across diverse datasets, even in the extreme case as shown in the left panel of Figure 10. Hence the well-behaved assumption is sufficient but not necessary for practical performance; an open question is whether error guarantees can be proved under a weaker assumption.

Extension to Local Differential Privacy. Our method generalizes to the local DP setting, where each client adds noise independently to their local histogram. In this case, the total noise increases with the number of clients, which is much larger than the distributed DP setting. To simplify the analysis, we assume each client holds one example, which is common in local DP. The total number of clients is equal to the number of examples n . The variance of the count in each bin is $O(n/\epsilon^2)$, and the expected quantile estimation error is $(1/Q + \sqrt{bh}/(\sqrt{n}\epsilon))$. Therefore, the Area Error is bounded by $O(1/Q + \sqrt{h}/(\sqrt{n}\epsilon))$; see [12] for relevant local DP techniques.

Limitations. A limitation is that at small Q , our method may not outperform range-query-based methods, especially when aligned thresholds are critical. This can be mitigated by hybrid methods

combining quantiles and range queries. For the ROC curve, quantiles of all examples can be computed, and TPR/FPR can be derived using range queries, as in Cormode and Markov [13]. Monotonicity can be enforced using post-processing as in Barczewski et al. [2]. Under mild assumptions, this approach may yield stronger area error bounds for the ROC curve, though it introduces high variance when Q is large and lacks clear area error bounds for the PR curve.

8 CONCLUSION

We studied the problem of computing ROC and PR curves as distributed analytical queries over sensitive data. In federated settings, exact evaluation by centralizing data is disallowed due to privacy constraints and prohibitive communication costs. We proposed a quantile-based framework that approximates these curves using compact distribution summaries, enabling efficient and privacy-preserving computation without access to raw data. Our method provides provable guarantees on the Area Error, which scales as $O(1/Q)$ (or $\tilde{O}(1/Q)$), while communication cost grows linearly with the number of quantiles Q . This establishes a clear and tunable trade-off between accuracy and communication, a key requirement in large-scale distributed systems. The proposed design leverages histogram-based aggregation, making it robust to data heterogeneity and compatible with secure aggregation and distributed differential privacy mechanisms. Experimental results demonstrate that our approach achieves high accuracy under strong privacy constraints, highlighting its practicality for real-world deployments. For future work, we plan to extend the framework to additional evaluation metrics and monitoring tasks. Another important direction is to strengthen robustness against adversarial or malicious clients, since it is possible for clients to manipulate their local histograms to bias the global curve estimation. In such cases, additional mechanisms such as by bounding the influence of each client (via succinct non-interactive zero-knowledge proofs) so that each client’s contribution falls within certain limits.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable feedback. XX is supported by a scholarship from the Department of Computer Science, University of Warwick. GC is supported in part by EPSRC grant EP/V044621/1.

REFERENCES

- [1] Naman Agarwal, Peter Kairouz, and Ziyu Liu. 2021. The skellam mechanism for differentially private federated learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, Article 386, 13 pages.
- [2] Antoine Barczewski, Amal Mawass, and Jan Ramon. 2025. Differentially Private Empirical Cumulative Distribution Functions. arXiv:2502.06651 [cs.CR]
- [3] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5XW20>.
- [4] James Bell, Aurélien Bellet, Adria Gascon, and Tejas Kulkarni. 2020. Private Protocols for U-Statistics in the Local Model and Beyond. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Vol. 108. PMLR, Virtual, 1573–1583.
- [5] James Henry Bell, Kallista A. Bonawitz, Adria Gascón, Tancrède Lepoint, and Mariana Raykova. 2020. Secure Single-Server Aggregation with (Poly)Logarithmic Overhead. In *ACM SIGSAC Conference on Computer and Communications Security*. ACM, New York, NY, USA, 1253–1269.
- [6] Daniel J. Beutel, Taner Topal, Akhil Mathur, Xinchu Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwong Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, and Nicholas D. Lane. 2022. Flower: A Friendly Federated Learning Research Framework. arXiv:2007.14390 [cs.LG]
- [7] Jock Blackard. 1998. Covertype. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C50K5N>.
- [8] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical Secure Aggregation for Privacy-Preserving Machine Learning. In *ACM CCS* (Dallas, Texas, USA). ACM, New York, NY, USA, 1175–1191.
- [9] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA). ACM, New York, NY, USA, 785–794.
- [10] Yan Chen, Ashwin Machanavajjhala, Jerome P. Reiter, and Andres F. Barrientos. 2016. Differentially Private Regression Diagnostics. In *IEEE ICDM*. IEEE Computer Society, Los Alamitos, CA, USA, 81–90.
- [11] Stéphan Clemençon and Nicolas Vayatis. 2009. Tree-Based Ranking Methods. *IEEE Transactions on Information Theory* 55, 9 (2009), 4316–4336.
- [12] Graham Cormode, Tejas Kulkarni, and Divesh Srivastava. 2019. Answering Range Queries Under Local Differential Privacy. *Proc. VLDB Endow.* 12, 10 (2019), 1126–1138.
- [13] Graham Cormode and Igor L. Markov. 2023. Federated Calibration and Evaluation of Binary Classifiers. *Proc. VLDB Endow.* 16, 11 (2023), 3253–3265.
- [14] P. Costantini. 1997. Algorithm 770: BVSPIS—a package for computing boundary-valued shape-preserving interpolating splines. *ACM Trans. Math. Softw.* 23, 2 (June 1997), 252–254.
- [15] Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (Pittsburgh, Pennsylvania, USA). ACM, New York, NY, USA, 233–240.
- [16] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.* 9, 3-4 (2014), 211–407.
- [17] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern Recognit. Lett.* 27, 8 (2006), 861–874.
- [18] Peter A. Flach and Meelis Kull. 2015. Precision-Recall-Gain curves: PR analysis done right. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1* (Montreal, Canada). MIT Press, Cambridge, MA, USA, 838–846.
- [19] F. N. Fritsch and J. Butland. 1984. A Method for Constructing Local Monotone Piecewise Cubic Interpolants. *SIAM J. Sci. Statist. Comput.* 5, 2 (1984), 300–304.
- [20] F. N. Fritsch and R. E. Carlson. 1980. Monotone Piecewise Cubic Interpolation. *SIAM J. Numer. Anal.* 17, 2 (1980), 238–246.
- [21] Jennifer Gillenwater, Matthew Joseph, and Alex Kulesza. 2021. Differentially Private Quantiles. In *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139. PMLR, Virtual, 3713–3722.
- [22] Slawomir Goryczka and Li Xiong. 2017. A Comprehensive Comparison of Multiparty Secure Additions with Differential Privacy. *IEEE Transactions on Dependable and Secure Computing* 14, 05 (Sept. 2017), 463–477.
- [23] Michael Hay, Vibhor Rastogi, Gerome Miklau, and Dan Suciu. 2010. Boosting the Accuracy of Differentially Private Histograms Through Consistency. *Proc. VLDB Endow.* 3, 1 (2010), 1021–1032.
- [24] Peter Kairouz, Ziyu Liu, and Thomas Steinke. 2021. The Distributed Discrete Gaussian Mechanism for Federated Learning with Secure Aggregation. In *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139. PMLR, Virtual, 5201–5212.
- [25] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D’Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adria Gascón, Badhi Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. 2021. Advances and Open Problems in Federated Learning. *Foundations and Trends® in Machine Learning* 14, 1–2 (2021), 1–210.
- [26] Edo Liberty and Jelani Nelson. 2012. Data Mining in Streams. Tutorial at ACM SIGKDD. Tutorial presented on August 12, 2012.
- [27] Thomas Lux, Layne T. Watson, Tyler Chang, and William Thacker. 2023. Algorithm 1031: MQSI—Monotone Quintic Spline Interpolation. *ACM Trans. Math. Softw.* 49, 1, Article 6 (March 2023), 17 pages.
- [28] Alvin F. Martin, George R. Doddington, Terri Kamm, Mark Ordowski, and Mark A. Przybicki. 1997. The DET curve in assessment of detection task performance. In *5th European Conference on Speech Communication and Technology (Eurospeech 1997)*. ISCA, Rhodes, Greece, 1895–1898.
- [29] Gregory J. Matthews and Ofer Harel. 2013. An Examination of Data Confidentiality and Disclosure Issues Related to Publication of Empirical ROC Curves. *Academic Radiology* 20, 7 (2013), 889–896.
- [30] Cleve B. Moler. 2004. *Numerical Computing with Matlab*. Society for Industrial and Applied Mathematics, Philadelphia, USA.
- [31] S. Moro, P. Rita, and P. Cortez. 2012. Bank Marketing. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5K306>.
- [32] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12 (2011), 2825–2830.
- [33] Foster J. Provost, Tom Fawcett, and Ron Kohavi. 1998. The Case against Accuracy Estimation for Comparing Induction Algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 445–453.
- [34] Wahbeh H. Qardaji, Weining Yang, and Ninghui Li. 2013. Understanding Hierarchical Methods for Differentially Private Histograms. *Proc. VLDB Endow.* 6, 14 (2013), 1954–1965.
- [35] Vijay Raghavan, Peter Bollmann, and Gwang S. Jung. 1989. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Trans. Inf. Syst.* 7, 3 (July 1989), 205–229.
- [36] Takaya Saito and Marc Rehmsmeier. 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE* 10, 3 (03 2015), 1–21.
- [37] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. In *IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, Los Alamitos, CA, USA, 3–18.
- [38] Jiankai Sun, Xin Yang, Yuanshun Yao, Junyuan Xie, Di Wu, and Chong Wang. 2023. DPAUC: Differentially Private AUC Computation in Federated Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 12 (Jun. 2023), 15170–15178.
- [39] Christopher K. I. Williams. 2021. The Effect of Class Imbalance on Precision-Recall Curves. *Neural Comput.* 33, 4 (2021), 853–857.
- [40] Xuefeng Xu and Graham Cormode. 2025. Federated Computation of ROC and PR Curves. arXiv:2510.04979 [cs.LG]