A Proof of Proposition 1

The goal of Proposition 1 is to bound the approximation error of the linear surrogate model on the fairness constraint manifold. Specifically, we show that with high probability, the linear proxy value $\mathbf{w}_\phi^\top \mathbf{h}$ remains close to the reference value ϕ_{proxy} when $\mathbf{h}$ is sampled from the level set H_ϕ.

Proof. The proof relies on establishing an integral identity along a path connecting the reference point $\mathbf{h}_\phi$ to a target point $\mathbf{h}$ on the manifold, and then bounding the error using statistical properties of the domain geometry and gradient misalignment.

Step 1: Establishing the Gradient Integral Identity. Let $H_\phi = \{\mathbf{h} \in \mathcal{H} : \varphi(\mathbf{h}) = \phi\}$ be the constraint set defined in the proposition. Consider an arbitrary distribution shift vector $\mathbf{h} \in H_\phi$ and the fixed reference point $\mathbf{h}_\phi \in H_\phi$. We define a straight-line path $l(t)$ connecting them:

$$l(t) = \mathbf{h}_\phi + t(\mathbf{h} - \mathbf{h}_\phi), \quad t \in [0, 1].$$

According to the *Fundamental Theorem of Calculus for Line Integrals*, the difference in the fairness evaluation φ between the endpoints can be expressed as the integral of the gradient along the path:

$$\begin{aligned}\varphi(\mathbf{h}) - \varphi(\mathbf{h}_\phi) &= \int_0^1 \nabla\varphi(l(t))^\top \frac{d}{dt} l(t)\, dt \\ &= \int_0^1 \nabla\varphi(l(t))^\top (\mathbf{h} - \mathbf{h}_\phi)\, dt.\end{aligned}$$

Since both $\mathbf{h}$ and $\mathbf{h}_\phi$ lie on the same level set H_ϕ, we have $\varphi(\mathbf{h}) = \phi$ and $\varphi(\mathbf{h}_\phi) = \phi$. Consequently, the left-hand side is exactly zero, yielding the identity:

$$0 = \int_0^1 \nabla\varphi(l(t))^\top (\mathbf{h} - \mathbf{h}_\phi)\, dt. \tag{10}$$

Step 2: Introducing the Linear Surrogate and Error Bound. We define the approximation error of the linear surrogate $\mathbf{w}_\phi$ relative to the proxy reference $\phi_{\text{proxy}} := \mathbf{w}_\phi^\top \mathbf{h}_\phi$ as:

$$Y(\mathbf{h}) := |\mathbf{w}_\phi^\top \mathbf{h} - \phi_{\text{proxy}}| = |\mathbf{w}_\phi^\top (\mathbf{h} - \mathbf{h}_\phi)|.$$

Using the property that $\int_0^1 dt = 1$, we rewrite the linear term $\mathbf{w}_\phi^\top (\mathbf{h} - \mathbf{h}_\phi)$ as an integral $\int_0^1 \mathbf{w}_\phi^\top (\mathbf{h} - \mathbf{h}_\phi)\, dt$. Subtracting the gradient identity (10) (which equals 0) from this expression allows us to isolate the misalignment between the surrogate and the true gradient:

$$\begin{aligned}\mathbf{w}_\phi^\top (\mathbf{h} - \mathbf{h}_\phi) - 0 &= \int_0^1 \mathbf{w}_\phi^\top (\mathbf{h} - \mathbf{h}_\phi)\, dt \\ &\quad - \int_0^1 \nabla\varphi(l(t))^\top (\mathbf{h} - \mathbf{h}_\phi)\, dt \\ &= \int_0^1 \left(\mathbf{w}_\phi - \nabla\varphi(l(t))\right)^\top (\mathbf{h} - \mathbf{h}_\phi)\, dt.\end{aligned}$$

Taking the absolute value and applying the *Cauchy-Schwarz inequality* to the integrand:

$$\begin{aligned}Y(\mathbf{h}) &= \left| \int_0^1 \left(\mathbf{w}_\phi - \nabla\varphi(l(t))\right)^\top (\mathbf{h} - \mathbf{h}_\phi)\, dt \right| \\ &\leq \int_0^1 \|\mathbf{w}_\phi - \nabla\varphi(l(t))\| \cdot \|\mathbf{h} - \mathbf{h}_\phi\|\, dt. \end{aligned} \tag{11}$$

Since the term $\|\mathbf{h} - \mathbf{h}_\phi\|$ is independent of t, we can separate the geometric distance from the gradient deviation:

$$Y(\mathbf{h}) \leq \|\mathbf{h} - \mathbf{h}_\phi\| \cdot \underbrace{\int_0^1 \|\mathbf{w}_\phi - \nabla\varphi(\mathbf{h}_\phi + t(\mathbf{h} - \mathbf{h}_\phi))\|\, dt}_{\text{Gradient Misalignment along path}}. \tag{12}$$

Step 3: Bounding the Expected Error. We now consider the expectation over the distribution of $\mathbf{h}$ on the manifold H_ϕ. Taking the expectation $\mathbb{E}_{\mathbf{h} \in H_\phi}$ on both sides of (12) and applying the *Cauchy-Schwarz inequality for random variables* ($\mathbb{E}[XY] \leq \sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]}$). Let the path integral residual be denoted by $R = \int_0^1 \|\mathbf{w}_\phi - \nabla\varphi(l(t))\|\, dt$. Then we have:

$$\mathbb{E}[Y(\mathbf{h})] \leq \sqrt{\mathbb{E}\left[\|\mathbf{h} - \mathbf{h}_\phi\|^2\right]}\sqrt{\mathbb{E}\left[R^2\right]}.$$

Comparing this with the definitions given in Proposition 1, we identify the two terms on the right-hand side as A and B:

- $A = \sqrt{\mathbb{E}[\|\mathbf{h} - \mathbf{h}_\phi\|^2]}$ represents the effective diameter.
- $B = \sqrt{\mathbb{E}\left[\left(\int_0^1 \|\mathbf{w}_\phi - \nabla\varphi(\mathbf{h}_\phi + t(\mathbf{h} - \mathbf{h}_\phi))\|\, dt\right)^2\right]}$ quantifies the gradient misalignment.

Thus, the bound on the expected error is:

$$\mathbb{E}[Y(\mathbf{h})] \leq A \cdot B. \tag{13}$$

Step 4: Deriving the Probabilistic Concentration Bound. Finally, we apply *Markov's Inequality* to the non-negative random variable $Y(\mathbf{h})$. For any threshold $\delta > 0$:

$$P(Y(\mathbf{h}) \geq \delta) \leq \frac{\mathbb{E}[Y(\mathbf{h})]}{\delta} \leq \frac{A \cdot B}{\delta}.$$

Equivalently, the probability that the error remains within δ is:

$$P(Y(\mathbf{h}) \leq \delta) \geq 1 - \frac{A \cdot B}{\delta}.$$

To ensure a confidence level of at least $1 - \beta$ (where $\beta \in (0, 1)$), we set the failure probability bound equal to β, which implies $\frac{A \cdot B}{\delta} = \beta$, or $\delta = \frac{A \cdot B}{\beta}$. Substituting this δ back yields the final result:

$$P\left(|\mathbf{w}_\phi^\top \mathbf{h} - \phi_{\text{proxy}}| \leq \frac{A \cdot B}{\beta}\right) \geq 1 - \beta.$$

□

B Details of the Experiment in RQ1

In this section, we provide a detailed description of the experimental setup for RQ1, designed to verify the estimation accuracy of CoRe against an analytically solvable ground truth.

B.1 Dataset Construction

To calculate the exact consistency radius, we constructed a synthetic environment based on a trinomial distribution variable $X \in \{0, 1, 2\}$. The dataset was generated through the following steps:

1. **Performance metric function (M):** We randomly sampled 10 performance vectors $\mathbf{m} \in [0, 1]^3$, where each component m_i represents the performance of the model on the data category i.
2. **Audit Distributions (P):** We randomly sampled 10 probability vectors $\mathbf{p} \in \mathbb{R}^3$ from a Dirichlet distribution, ensuring $\sum p_i = 1$. This represents the initial audit data distribution.

The Cartesian product of these sets yields 100 distinct auditing configurations (pairs of M and P).

B.2 Feasible Threshold Selection

A critical parameter in our problem is the fairness tolerance threshold ϵ. Randomly choosing ϵ is problematic because certain fairness levels may be mathematically impossible for a given model M, regardless of the distribution shift.

To ensure valid and comprehensive testing, we first derived the **feasible range** of fairness values for each model. The theoretical boundaries of achievable fairness φ on the probability simplex are determined by the roots of the following quadratic equation:

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}, \tag{14}$$

where the coefficients depend on the model performance vector $\mathbf{m}$: $a = -1$, $b = m_0 + m_1 - 2m_2$, and $c = (m_0 - m_2)(m_2 - m_1)$.

For each of the 100 configurations, we calculated this feasible range and then uniformly sampled 10 values of ϵ within it. This sampling strategy generates a total of 1,000 test instances, covering the full spectrum of auditing scenarios—from loose constraints (easy to satisfy) to extremely tight constraints (borderline feasibility).

B.3 Microscopic Analysis: Accuracy Decay on Corner Cases

While the main text reports the aggregated error metrics, we provide here a fine-grained analysis of how the estimation behavior changes with the strictness of the fairness constraint.

Figure 5 visualizes the relationship between the ground-truth radius (Solid Line) and the radius estimated by CoRe (Points) for a representative (M, P) pair. The x-axis represents the chosen tolerance ϵ.

Observation: The results show a clear pattern: CoRe achieves high precision in the central region of the feasible range. However, the estimation error tends to increase as ϵ approaches the boundaries (indicated by the shaded regions in the figure).

Analysis: This phenomenon reveals the geometric characteristics of the consistency radius problem. The boundary regions correspond to extreme fairness values that are only achievable by highly skewed probability distributions. In these "corner cases," the geometry of the fairness constraint manifold becomes highly curved. Since CoRe relies on a linear surrogate vector $\mathbf{w}^*$ to approximate this non-convex constraint locally, the approximation error naturally grows when the local curvature is sharp.

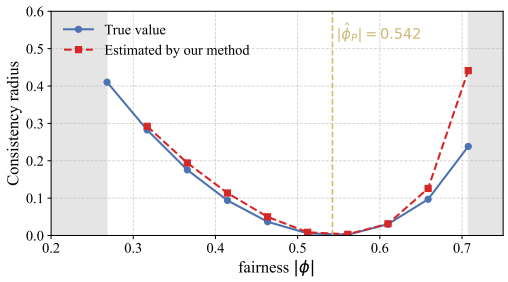

Figure 5: Visual comparison of the Ground-Truth Consistency Radius vs. CoRe Estimated Radius for a single configuration (M, P) across the feasible ϵ spectrum. The shaded regions denote the boundaries of the feasible fairness range.

Nevertheless, this analysis confirms that for the vast majority of practical auditing scenarios (the non-extreme regions), our convex relaxation method provides a reliable and tight estimation. The deviations at the boundaries also explain the outliers observed in the aggregated error statistics.

C Details of the Experiment in RQ2

This section details the model architectures, training protocols, and fairness intervention strategies for the three real-world datasets used in RQ2.

C.1 Dataset and Model Settings

MovieLens-1M (Interaction Data). The dataset contains approximately 6,000 users and 4,000 movies. We preprocessed the data by removing users and items with fewer than five interactions and re-indexing the remaining IDs. We employed a leave-one-out splitting strategy: for each user, the most recent interaction was used for testing, while the preceding interactions formed the training set.

We trained a LightGCN model, a state-of-the-art collaborative filtering method. The model consists of user and item embedding layers with a dimension of 64. Following the standard LightGCN implementation, we simplified the graph convolution by removing feature transformation and nonlinear activation. The model was trained using the Adam optimizer with a learning rate of 0.005 and a batch size of 1024. We used the Bayesian Personalized Ranking (BPR) loss, sampling one negative item for each positive instance. Early stopping was triggered if the validation NDCG@10 did not improve for 5 consecutive epochs.

Adult (Structured Data). We trained a Multilayer Perceptron (MLP) on the Adult dataset to predict income levels ($\leq$50K vs. $>$50K). Missing values were removed, and categorical features were encoded using `LabelEncoder`, while numerical features were standardized using `StandardScaler`. The dataset was split into

80% training and 20% testing, with 20% of the training data reserved for validation.

The MLP consists of two hidden layers with 64 and 32 units, respectively, using ReLU activation and Dropout (rate=0.2). The model was trained using the Adam optimizer (learning rate 1×10^{-3}) and binary cross-entropy loss. We applied early stopping with a patience of 5 epochs based on validation loss.

CelebA (Image Data). For the image classification task, we used the CelebA dataset to predict the "Smiling" attribute. We center-cropped the images to 178×178 and resized them to 128×128. The pixel values were normalized to the range $[0, 1]$. We utilized a ResNet-18 architecture. To simulate a standard auditing scenario where models are often trained from scratch on specific datasets, we trained the ResNet-18 without pre-trained weights. The model was optimized using Adam with a learning rate of 1×10^{-4} and a batch size of 64 for 20 epochs. The sensitive attribute was defined as "Male" (Gender).

D Details of the Experiment for Addressing practitioners' Queries

In this section, we provide the detailed experimental setup and data generation processes for the exploratory analyses presented in Section 6.

D.1 Simulation Setup for Radius and Fairness Correlation

To analyze the relationship between the consistency radius and the difference between the fairness evaluation result $|\varphi(D)|$ and the threshold ϵ, we conducted a simulation study using synthetic data.

Data Generation. We randomly constructed 3,000 synthetic datasets. For each dataset, we generated a metric vector $\mathbf{m}$ and a group attribute vector $\mathbf{g}$ as follows:

- **Sample Size:** Each synthetic dataset consists of $N = 100$ samples.
- **Metric Values ($\mathbf{m}$):** The metric values (e.g., prediction scores or losses) were sampled independently from a Uniform distribution over the interval $[0, 1]$, i.e., $m_i \sim \mathcal{U}(0, 1)$.
- **Group Attribute ($\mathbf{g}$):** We initialized binary group labels $\{0, 1\}$ randomly. To eliminate the confounding effect of group imbalance, we enforced a strictly balanced group ratio. Specifically, we adjusted the counts such that exactly 50 samples belonged to the protected group ($g = 0$) and 50 samples to the non-protected group ($g = 1$).

Correlation Analysis. For each dataset, we computed the consistency radius given a fixed fairness threshold ϵ (e.g., $\epsilon = 0$ or $\epsilon = 0.1$). We then calculated the absolute difference $|\hat{\phi}_P| - \epsilon$ and quantified its correlation with the consistency radius using the Kendall rank correlation coefficient (τ), utilizing the `scipy.stats` library.

D.2 Simulation Setup for Dataset Merging

To investigate the impact of merging audit datasets on the consistency radius, we analyzed pairs of datasets based on the assumption of identical group proportions.

Dataset Construction. We constructed 3,000 pairs of datasets $(\mathcal{D}_1, \mathcal{D}_2)$. Each dataset in the pair was generated independently using the same procedure described above ($N = 100$, $\mathbf{m} \sim \mathcal{U}(0, 1)$, and strictly balanced groups). Since both $\mathcal{D}_1$ and $\mathcal{D}_2$ are strictly balanced (50% group 0, 50% group 1), merging them naturally satisfies the requirement that the group proportions remain consistent across the individual and merged datasets.

Evaluation. For each pair, we created a merged dataset $\mathcal{D}_{\text{concat}} = \mathcal{D}_1 \cup \mathcal{D}_2$ by concatenating the metric and group vectors. We then computed the consistency radius for $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_{\text{concat}}$ separately under the same threshold ϵ. We compared the radius of the merged dataset against the minimum and maximum radii of the individual datasets to observe the trends shown in Figure 4b.

E CoRe Algorithm Implementation Details

The CoRe algorithm was implemented in Python. For the inner-level convex optimization problem (finding the optimal shift $\mathbf{h}$ under a linear constraint), we utilized the `CVXPY` library with the `ECOS` solver, which is efficient for handling entropy-based objectives and linear constraints.

The binary search in the outer loop was configured with a convergence tolerance of $\texttt{tol} = 1 \times 10^{-4}$ and a maximum of 30 iterations. The regularization parameter λ for stabilizing group scales was set to 1.0 by default. All experiments were conducted on a Linux server equipped with an Intel Xeon Gold 6240 CPU and an NVIDIA GeForce RTX 3090 GPU. The estimation of the consistency radius typically converges within seconds for audit datasets of size $n = 1,000$, confirming the computational efficiency of the method.