Fairness and De-biasing

Understanding Conditional Parity and its Applications

July 5, 2025

Tutorial Outline

Machine Learning Fairness: The Core Problem

The Challenge

How do we ensure that machine learning models make decisions that are **independent of sensitive attributes** while preserving relevant information for accurate predictions?

What We Want to Prevent:

- Systematic discrimination
- Biased decision-making
- Disparate impact on protected groups
- Unfair treatment based on irrelevant characteristics

What We Want to Preserve:

- Accuracy and performance
- Merit-based decisions
- Legitimate risk assessment
- Relevant business considerations

The Solution: Conditional Independence

Make predictions independent of sensitive attributes, given relevant conditioning information.

Conditional Parity: The Universal Framework

General Definition

Model predictions should be independent of sensitive attributes, conditional on a chosen set of variables.

$$\boxed{\hat{Y} \perp A \mid Z} \tag{1}$$

Symbol	Meaning
$\hat{Y}$	Model prediction/decision
A	Sensitive attribute (race, gender, age)
Z	Conditioning variables (features we condition on)
$\perp$	Statistical independence

Key Insight

The choice of conditioning variables Z determines the type of fairness and when it can be applied. This is the **fundamental decision** in fairness design.

The Information Availability Spectrum

Forward-Looking Decisions

Examples:
Credit Scoring
Hiring
Insurance
College Admissions

Condition on:
Risk factors
Qualifications
Observable features

Mixed Scenarios

Examples:
Medical Diagnosis
Model Auditing
Performance Review
Post-hoc Analysis

Condition on:
True outcomes
Known results
Verified labels

Retrospective Analysis

Information Available at Decision Time

No Outcomes | Some Outcomes | Full Outcomes

The Fundamental Principle: What we can condition on depends on what information is available when we need to make the fairness determination.

The Fairness Taxonomy: Special Cases of Conditional Parity

Conditional Parity: $\hat{Y} \perp A \mid Z$

Unconditional ($Z = \emptyset$)
Demographic Parity
$P(\hat{Y} = 1 | A = a) = P(\hat{Y} = 1 | A = b)$

Risk Factors ($Z = X$)
Risk-Based Parity
$\hat{Y} \perp A \mid \{\text{credit score, income}\}$

True Negatives ($Z = \{Y = 0\}$)
Predictive Parity
Equal precision among negative cases

$$P(\hat{Y} = 1 | Y = 0, A = a) = P(\hat{Y} = 1 | Y = 0, A = b)$$

True Positives ($Z = \{Y = 1\}$)
Equal Opportunity
Equal recall among positive cases

$$P(\hat{Y} = 1 | Y = 1, A = a) = P(\hat{Y} = 1 | Y = 1, A = b)$$

All fairness criteria are conditional parity with different choices of Z

Forward-Looking Fairness: Risk-Based Parity

When to Use: When making decisions without knowing future outcomes - the most common real-world scenario.

$$\hat{Y} \perp A \mid X \quad \text{where } X = \text{observable risk factors} \tag{2}$$

$$P(\text{Approval}|\text{Credit Score} = c, A = \text{male}) = P(\text{Approval}|\text{Credit Score} = c, A = \text{female})$$

Credit Scoring Example:

- **Decision:** Approve/deny loan
- **Sensitive attribute** (A)**:** Race, gender
- **Risk factors** (X)**:** Credit score, income, DTI ratio
- **Unknown:** Will they actually repay?

Fairness Requirement: Equal approval rates for applicants with same risk profile.

Other Applications:

- **Hiring:** Equal selection rates given qualifications
- **College admissions:** Equal acceptance given academic metrics

Key Advantage: Can be enforced at decision time because all conditioning variables are observable.

Retrospective Fairness: Equal Opportunity

When to Use: When true outcomes are known and we want to audit model performance retrospectively.

$$\hat{Y} \perp A \mid Y = 1 \quad \text{(condition on true positive outcomes)} \tag{3}$$

Medical Diagnosis Example:

- **Decision:** Predict disease presence
- **Ground truth:** Test results available
- **Fairness:** Equal detection rates for patients who actually have disease

Credit Auditing Example:

- **Context:** 2 years after loan decisions
- **Now known:** Who actually repaid
- **Audit:** Were approval rates fair among those who repaid?

Why It Works Here:

- True outcomes Y are observable
- Can evaluate retrospectively
- Useful for model auditing
- Performance assessment

Why It Fails in Forward-Looking:

- Can't condition on unknown Y
- Decision must be made before outcome

Comprehensive Fairness: Equalized Odds

Definition: Condition on both positive and negative true outcomes - combines equal opportunity with equal false positive rates.

$$\hat{Y} \perp A \mid Y \quad \text{(condition on all true outcomes)} \tag{4}$$

This is equivalent to requiring both:

$$P(\hat{Y} = 1 | Y = 1, A = a) = P(\hat{Y} = 1 | Y = 1, A = b) \quad \text{(Equal Opportunity)} \tag{5}$$

$$P(\hat{Y} = 1 | Y = 0, A = a) = P(\hat{Y} = 1 | Y = 0, A = b) \quad \text{(Equal FPR)} \tag{6}$$

Applications:

- Criminal justice risk assessment
- Medical diagnosis auditing
- Comprehensive model evaluation
- High-stakes decision auditing

Characteristics:

- Most comprehensive fairness criterion
- Addresses both types of errors
- Requires known ground truth
- Primarily for retrospective analysis

Demographic Parity: The Baseline Case

Definition: No conditioning - require equal positive rates regardless of any other factors.

$$\hat{Y} \perp A \mid \emptyset \quad \Leftrightarrow \quad P(\hat{Y} = 1|A = a) = P(\hat{Y} = 1|A = b) \tag{7}$$

When Appropriate:

- Legal compliance requirements
- When background differences should be ignored
- Quota-based systems
- Representation goals

Advantages:

- Simple to understand and implement
- Clear legal interpretation
- Easy to measure and monitor

Serious Limitations:

- Ignores legitimate differences
- Can lead to unqualified selections
- May violate merit-based principles
- Often conflicts with accuracy

Risk Example: In credit scoring, demographic parity could require approval regardless of creditworthiness, leading to losses and potentially harming borrowers.

Why Credit Scoring Cannot Use Equal Opportunity

t_0: Decision Time

Available:
- Credit score: 680
- Income: \$55K
- DTI ratio: 28%
- Employment: 3 years

Unknown:
- Will they repay?
- Future circumstances
- Economic conditions
- $Y = ?$

THE PROBLEM:
Equal opportunity requires conditioning on $Y = 1$,
but Y is unknown at decision time!
$P(\hat{Y} = 1 | Y = 1, A) = ?$ when Y is unknown

t_1: 24 months later

Now Known:
- All previous info
- Actual repayment
- $Y = 1$ or $Y = 0$
- Economic events

Timeline

Risk-Based Parity

Equal Opportunity

Loan Application
Decision Point

Loan Outcome
Outcome Known

The Solution: Different Fairness for Different Times

Decision Time

Risk-Based Parity

$\hat{Y} \perp A \mid \{\text{credit score, income, DTI}\}$

Goal: Fair decisions given available information

Time $\rightarrow$

Audit Time

Equal Opportunity

$\hat{Y} \perp A \mid Y = 1$

Goal: Evaluate fairness among those who repaid

Examples:
Credit approval
Hiring decisions
Insurance underwriting

Examples:
Loan performance audits
Employee performance reviews
Claims analysis

Key Insight: Use risk-based parity for decisions, then audit with equal opportunity when outcomes are known. The same domain can use different fairness criteria depending on whether we're making decisions or conducting audits.

Adverse Impact Ratio: The Universal Metric

Definition

AIR measures the ratio of favorable outcomes between unprivileged and privileged groups - works across all fairness frameworks.

$$\text{AIR} = \frac{P(\hat{Y} = 1 | A = \text{unprivileged})}{P(\hat{Y} = 1 | A = \text{privileged})} \tag{8}$$

The 80% Rule:

- AIR ≥ 0.8: Generally fair
- AIR < 0.8: Potential disparate impact
- Based on EEOC guidelines

Practical Advantage:

- Easy to compute and interpret
- Legal and regulatory alignment
- Consistent across fairness types
- Clear threshold for action

Framework-Specific Metrics

Fairness Framework	**Key Metric**	**Formula**
Demographic Parity	Statistical Parity Diff	$P(\hat{Y} = 1 \mid A = a) - P(\hat{Y} = 1 \mid A = b)$
Risk-Based Parity	Conditional AIR	$\frac{P(\hat{Y}=1 \mid X, A=a)}{P(\hat{Y}=1 \mid X, A=b)}$
Equal Opportunity	TPR Difference	$P(\hat{Y} = 1 \mid Y = 1, A = a) - P(\hat{Y} = 1 \mid Y = 1, A = b)$
Equalized Odds	Max(TPR, FPR) Diff	$\max(\lvert \Delta \mathrm{TPR} \rvert, \lvert \Delta \mathrm{FPR} \rvert)$

When to Use:

- **Demographic Parity:** Baseline compliance
- **Risk-Based Parity:** Decision time
- **Equal Opportunity:** Auditing
- **Equalized Odds:** Comprehensive audit

Framework-Specific Metrics

MoDeVa Supported Metrics:

- **AIR**: Adverse Impact Ratio
- **Precision**: PPV disparity ratio
- **Recall**: TPR disparity ratio
- **SMD**: Standardized mean difference (regression)

Metric Selection Strategy:

- Start with AIR for all frameworks
- Add framework-specific metrics
- Monitor multiple metrics simultaneously
- Understand metric relationships and conflicts

Setting Up Data and Models in MoDeVa

```python
from modeva import DataSet
from modeva.models import MoLGBMClassifier, MoXGBClassifier
# Load and preprocess data
ds = DataSet(name="TaiwanCredit")
ds.load("TaiwanCredit")
ds.encode_categorical(method="ordinal")
ds.preprocess()
# Configure target and features
ds.set_target("FlagDefault")
# Keep sensitive attributes separate - don't use for modeling
ds.set_inactive_features(["SEX", "MARRIAGE", "AGE"])
ds.set_random_split()
# Train models
model_lgbm = MoLGBMClassifier(name="LGBM_model", max_depth=2, n_estimators=100)
model_xgb = MoXGBClassifier(name="XGB_model", max_depth=2, n_estimators=100)
model_lgbm.fit(ds.train_x, ds.train_y)
model_xgb.fit(ds.train_x, ds.train_y)
```

Implementing Risk-Based Parity

```python
from modeva import TestSuite
# Set up protected group data
ds.set_protected_data(ds.raw_data[["SEX", "MARRIAGE", "AGE"]])
# Define groups for risk-based parity analysis
group_config = {
    "Gender": {"feature": "SEX", "protected": 1.0, "reference": 2.0},
    "Marriage": {"feature": "MARRIAGE", "protected": 2.0, "reference": 1.0},
    "Age": {
        "feature": "AGE",
        "protected": {"lower": 60, "lower_inclusive": True},
        "reference": {"upper": 60, "upper_inclusive": False}
    }
}
# Create test suite and evaluate fairness
ts = TestSuite(ds, model_lgbm)
# Assess overall fairness using AIR
results = ts.diagnose_fairness(
    group_config=group_config, favorable_label=1,
    metric="AIR", threshold=0.8)
results.plot()
```

Risk Stratified Analysis

Conditional Parity within Risk Strata

```
# Analyze fairness within specific risk factor slices
# This implements true conditional parity

# Single risk factor analysis
results = ts.diagnose_slicing_fairness(
    features="PAY_1",  # Payment history feature
    group_config=group_config,
    dataset="train",
    metric="AIR"
)
results.plot()
```

Risk Stratified Analysis

Conditional Parity within Risk Strata

```
# Multiple risk factors (implements P(approval | credit_score, income,
    gender))
results = ts.diagnose_slicing_fairness(
    features=("PAY_1", "BILL_AMT1"),  # Payment + balance features
    group_config=group_config, dataset="train",
    metric="AIR", threshold=0.9
)
results.plot("Marriage")

# Comprehensive analysis across all risk factors
feature_names = tuple((x,) for x in ds.feature_names)
results = ts.diagnose_slicing_fairness(
    features=feature_names, group_config=group_config,
    dataset="train", metric="AIR",
    method="auto-xgb1", bins=5
)
```

Retrospective Analysis with Equal Opportunity

Auditing Model Performance

```python
# Note: This would be used AFTER loan outcomes are known
# For demonstration, we use the available labels as proxy
# Equal opportunity analysis (retrospective)
# This conditions on Y=1 (actual positive outcomes)
results = ts.diagnose_fairness(
    group_config=group_config, favorable_label=1,
    metric="Recall",  # This measures P(pred=1|true=1, group)
    threshold=0.8
)

# Compare multiple models for fairness
tsc = TestSuite(ds, models=[model_lgbm, model_xgb])
results = tsc.compare_fairness(
    group_config=group_config, metric="AIR",
    threshold=0.8
)
results.plot()
```

Retrospective Analysis with Equal Opportunity

Auditing Model Performance

```
# Detailed comparison across risk slices
result = tsc.compare_slicing_fairness(
    features="BILL_AMT1",
    group_config=group_config,
    favorable_label=1,
    dataset="train",
    metric="AIR"
)
print("XGB Model Results:", result.table["XGB_model"]["Marriage"])
print("LGBM Model Results:", result.table["LGBM_model"]["Marriage"])
```

Post-Processing Mitigation Strategies

Threshold Adjustment:

- Modify decision boundaries per group
- Lower threshold for unprivileged groups
- Direct impact on AIR
- Preserves model structure

Feature Binning:

- Reduce feature precision
- Group similar values together
- Limits discriminatory patterns
- May reduce accuracy

Threshold Adjustment Implementation

```python
# Threshold adjustment for improved fairness
result = ts.diagnose_mitigate_unfair_thresholding(
    group_config=group_config, favorable_label=1,
    dataset="train", metric="AIR",
    performance_metric="ACC",  # Track accuracy trade-off
    proba_cutoff=(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)
)
# Analyze results
print("Threshold Analysis Results:")
print(result.table)
result.plot()
```

How it Works:

- Test range of probability thresholds
- Measure AIR and accuracy for each
- Find optimal balance point
- Apply group-specific thresholds

Typical Results:

- Lower threshold for unprivileged groups
- Higher threshold for privileged groups
- Improved AIR (closer to 1.0)
- Some accuracy reduction

Feature Binning Implementation

```
# Feature binning for fairness improvement
result = ts.diagnose_mitigate_unfair_binning(
    group_config=group_config, favorable_label=1,
    dataset="train", metric="AIR",
    performance_metric="AUC",  # Monitor predictive performance
    binning_method="quantile",  # Equal frequency bins
    bins=5  # Number of bins to create
)
result.plot()
```

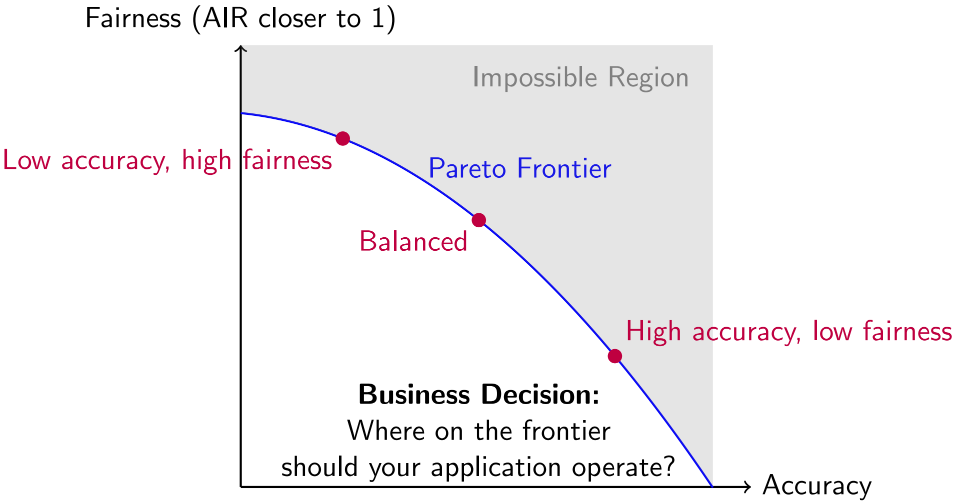
Binning Strategy:

- Start with 3-5 bins
- Test different binning methods
- Monitor both fairness and performance
- Apply to most discriminatory features

Expected Impact:

- Smoother decision boundaries
- Reduced group-specific overfitting
- Improved statistical parity
- Potential accuracy loss

The Fairness-Accuracy Tradeoff



The Fairness-Accuracy Tradeoff

Key Considerations:

- Perfect fairness and accuracy are usually impossible simultaneously
- How large the tradeoff is depends on the populations in question
- The optimal point depends on business context, legal requirements, and ethical considerations
- Different stakeholders may prefer different points on the frontier
- Transparent documentation of tradeoff decisions is essential

The Unified Fairness Framework: Summary

Risk-Based Parity

$Z = \{\text{risk factors}\}$

Forward-looking decisions

Equal Opportunity

$Z = \{Y = 1\}$

Retrospective auditing

$$\hat{Y} \perp A \mid Z$$

Universal Framework: Choice of conditioning variables Z determines fairness type and applicability

Demographic Parity

$Z = \emptyset$

Compliance checking

Equalized Odds

$Z = \{Y\}$

Comprehensive auditing