Fairness in Generative Modelling

Alexandre Vérine

Machine Intelligence and Learning Systems, LAMSADE, Université Paris-Dauphine-PSL

June 3rd 2025

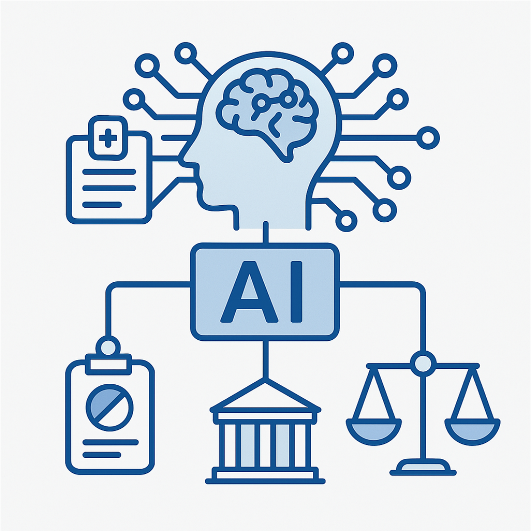
1794 ENS | PSL

Dauphine UNIVERSITÉ PARIS | PSL LAMSADE UMR 7243 MILES Machine Intelligence and Learning Systems

What is Fairness?

Generic Definition and Origins

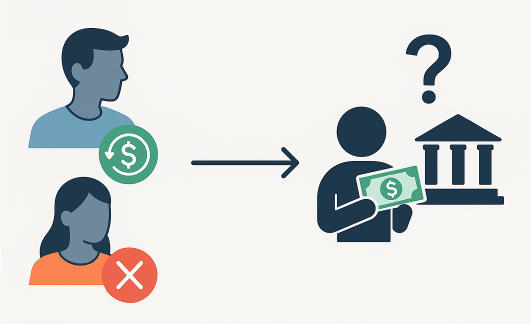
- Fairness: the quality of being free from bias, favoritism, or injustice.
- Earliest mentions date to the late 1960s in the context of test scoring and employment.
- Renewed interest in early 2010s with high-profile algorithmic decision issues.



Mathematical Context

Formalizing the Problem

- X: feature vector (e.g., income, age, credit history)
- Y: true binary label (e.g., loan repayment: 1 for yes, 0 for no)
- $\hat{Y}$: predicted label output by model
- A: sensitive attribute (e.g., gender, race)

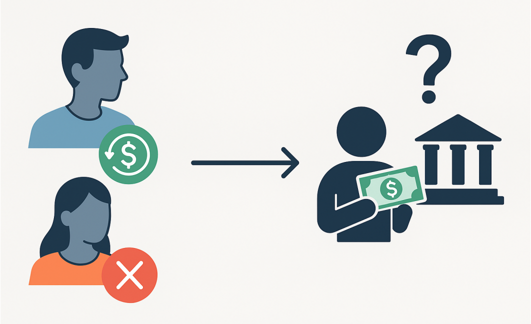


Mathematical Context

Formalizing the Problem

Example: Loan Approval

- X: age, income, employment history, etc.
- $Y = 1$: applicant repaid the loan
- $\hat{Y} = 1$: model predicts repayment $\rightarrow$ approve the loan
- A: gender

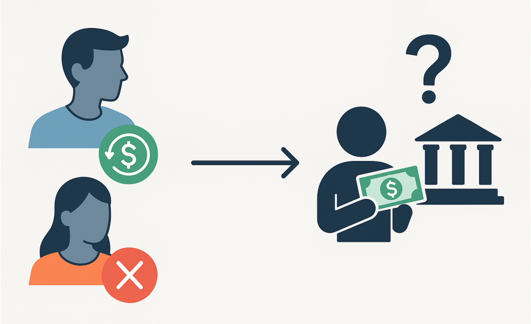


Mathematical Context

Formalizing the Problem

Objective and Fairness Concern

- Goal: Learn $f : X \to \hat{Y}$ that is **accurate** and **fair**.
- Fairness = model should not produce systematically different outcomes for groups defined by A.
- Complication: A may not be used explicitly but can be encoded in X.



Confusion Matrix

Basics and Example

	True label = 1	**True label = 0**	**Total**
Predicted = 1	TP	FP	PP
Predicted = 0	FN	TN	PN
Total	P	N	

- TP: predicted 1, true 1 → correctly predicted good payer
- FP: predicted 1, true 0 → incorrectly predicted good payer
- FN: predicted 0, true 1 → missed a good payer
- TN: predicted 0, true 0 → correctly rejected a bad payer

Confusion Matrix

Basics and Example

	True label = 1	**True label = 0**	**Total**
Predicted = 1	56	14	70
Predicted = 0	18	12	30
Total	74	26	100

Confusion Matrix

Basics and Example

	True label = 1	**True label = 0**	**Total**
Predicted = 1	20 + 36	4 + 10	24 + 46
Predicted = 0	12 + 6	5 + 7	15 + 15
Total	32 + 42	9 + 17	41 + 59

Accuracy

Fairness via Overall Performance

Definition

The accuracy of a binary classifier is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

- **Interpretation:** Proportion of correct predictions over all predictions.
- **Pros:** Simple, intuitive, and widely used.
- **Cons:** Does not reflect disparities between subgroups. A model can be accurate overall while being unfair to some groups.

Accuracy

Fairness via Overall Performance

Definition

The accuracy of a binary classifier is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

- **Example (global):**

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} = \frac{56 + 12}{56 + 14 + 12 + 18} = \frac{68}{100} = 0.68$$

- **By group:**
  - Male: $\frac{20+5}{41} \approx 0.610$
  - Female: $\frac{36+7}{59} \approx 0.728$

Accuracy

Fairness via Overall Performance

Definition

The accuracy of a binary classifier is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

- **Example (global):**

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} = \frac{56 + 12}{56 + 14 + 12 + 18} = \frac{68}{100} = 0.68$$

- **By group:**
  - Male: $\frac{20+5}{41} \approx 0.610$
  - Female: $\frac{36+7}{59} \approx 0.728$
  - The model appear to be more accurate for men than for female.

Precision and Recall

Fairness via Performance Decomposition

Definitions

$$\text{Recall} = \frac{TP}{TP + FN}, \quad \text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall}^- = \frac{TN}{TN + FP}, \quad \text{Precision}^- = \frac{TN}{TN + FN}$$

- **Recall**: Among all actual positives, how many did we correctly predict?
- **Precision**: Among all positive predictions, how many were correct?
- **Recall$^-$**: Among actual negatives, how many did we correctly reject?
- **Precision$^-$**: Among all negative predictions, how many were correct?

Precision and Recall

Fairness via Performance Decomposition

Definitions

$$\text{Recall} = \frac{TP}{TP + FN}, \quad \text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall}^- = \frac{TN}{TN + FP}, \quad \text{Precision}^- = \frac{TN}{TN + FN}$$

- Recall$^-$ is also known as **True Negative Rate (TNR)** or **Specificity**. 1$-$Recall$^-$ is the Type I error rate.
- Recall is also known as **True Positive Rate (TPR)** or **Sensitivity**. 1$-$Recall is the Type II error rate.

Precision and Recall

Example Calculations by Group

From the Confusion Matrix:

- Male:

$$\text{Recall}_M = \frac{20}{32} \approx 0.625, \quad \text{Precision}_M = \frac{20}{24} \approx 0.833$$

$$\text{Recall}^-_M = \frac{5}{9} \approx 0.556, \quad \text{Precision}^-_M = \frac{5}{17} \approx 0.294$$

- Female:

$$\text{Recall}_F = \frac{36}{42} \approx 0.857, \quad \text{Precision}_F = \frac{36}{46} \approx 0.783$$

$$\text{Recall}^-_F = \frac{7}{17} \approx 0.412, \quad \text{Precision}^-_F = \frac{7}{13} \approx 0.538$$

Equal Opportunity

Fairness via True Positive Parity

Definition

A predictor satisfies **Equal Opportunity** if:

$$\text{Recall}_a = \text{Recall}_b$$

where Recall (True Positive Rate) = $\frac{TP}{TP+FN}$

- **Interpretation:** Equal chance of correctly identifying positives across groups.
- **Pros:** Focuses on fairness in access to opportunities (beneficial outcomes).
- **Cons:** Ignores potential disparities in false positive rates.

Equal Opportunity

Fairness via True Positive Parity

Definition

A predictor satisfies **Equal Opportunity** if:

$$\text{Recall}_a = \text{Recall}_b$$

where Recall (True Positive Rate) = $\frac{TP}{TP+FN}$

Recall Values in Our Example:

- Male: $\text{Recall}_\text{M} = \frac{20}{20+12} = \frac{20}{32} \approx 0.625$
- Female: $\text{Recall}_\text{F} = \frac{36}{36+6} = \frac{36}{42} \approx 0.857$

Equal Opportunity

Fairness via True Positive Parity

Relaxed Definition (Additive)

A predictor satisfies ε**-Equal Opportunity (additive)** if:

$$|\text{Recall}_a - \text{Recall}_b| \leq \varepsilon$$

where Recall (True Positive Rate) = $\frac{TP}{TP+FN}$

Recall Values in Our Example:

- Male: $\text{Recall}_{\text{M}} = \frac{20}{20+12} = \frac{20}{32} \approx 0.625$
- Female: $\text{Recall}_{\text{F}} = \frac{36}{36+6} = \frac{36}{42} \approx 0.857$
- Difference: 0.232

Equal Opportunity

Fairness via True Positive Parity

Relaxed Definition (Multiplicative)

A predictor satisfies ε**-Equal Opportunity (multiplicative)** if:

$$\left| \frac{\text{Recall}_a}{\text{Recall}_b} - 1 \right| \leq \varepsilon$$

where Recall (True Positive Rate) = $\frac{TP}{TP+FN}$

Recall Values in Our Example:

- Male: $\text{Recall}_\text{M} = \frac{20}{20+12} = \frac{20}{32} \approx 0.625$
- Female: $\text{Recall}_\text{F} = \frac{36}{36+6} = \frac{36}{42} \approx 0.857$
- Ratio: $\left|\frac{0.625}{0.857} - 1\right| \approx 0.271$

Equalized Odds

Fairness via Equal True and False Positive Rates

Definition

A predictor satisfies **Equalized Odds** if:

$$\text{Recall}_a = \text{Recall}_b \quad \text{and} \quad \text{Recall}_a^- = \text{Recall}_b^-$$

where Recall = $\frac{TP}{TP+FN}$ and Recall$^-$ = $\frac{TN}{TN+FP}$

- **Interpretation:** Requires equal true positive rates (Recall) and equal true negative rates (Recall$^-$) across groups.
- **Pros:** Stronger fairness criterion than Equal Opportunity; ensures both equal access and equal errors.
- **Cons:** May be incompatible with maximizing overall accuracy.
  For example, to equalize error rates between groups, the model might reject more applicants from a high-performing group, lowering overall accuracy.

Equalized Odds

Fairness via Equal True and False Positive Rates

Definition

A predictor satisfies **Equalized Odds** if:

$$\text{Recall}_a = \text{Recall}_b \quad \text{and} \quad \text{Recall}_a^- = \text{Recall}_b^-$$

where Recall = $\frac{TP}{TP+FN}$ and Recall$^-$ = $\frac{TN}{TN+FP}$

Recall and Recall$^-$ Values in Our Example:

- Male: $\text{Recall}_\text{M} = \frac{20}{20+12} = \frac{20}{32} \approx 0.625$, $\text{Recall}_\text{M}^- = \frac{5}{5+4} = \frac{5}{9} \approx 0.556$
- Female: $\text{Recall}_\text{F} = \frac{36}{36+6} = \frac{36}{42} \approx 0.857$, $\text{Recall}_\text{F}^- = \frac{7}{7+10} = \frac{7}{17} \approx 0.412$

Equalized Odds

Fairness via Equal True and False Positive Rates

Relaxed Definition (Additive)

A predictor satisfies ε-**Equalized Odds (additive)** if:

$$|\mathrm{Recall}_a - \mathrm{Recall}_b| \leq \varepsilon \quad \text{and} \quad \left|\mathrm{Recall}^-_a - \mathrm{Recall}^-_b\right| \leq \varepsilon$$

Recall and Recall$^-$ Values in Our Example:

- Male: $\mathrm{Recall}_\mathrm{M} = \frac{20}{20+12} = \frac{20}{32} \approx 0.625$, $\mathrm{Recall}^-_\mathrm{M} = \frac{5}{5+4} = \frac{5}{9} \approx 0.556$
- Female: $\mathrm{Recall}_\mathrm{F} = \frac{36}{36+6} = \frac{36}{42} \approx 0.857$, $\mathrm{Recall}^-_\mathrm{F} = \frac{7}{7+10} = \frac{7}{17} \approx 0.412$
- Recall difference: 0.232; Recall$^-$ difference: 0.144

Equalized Odds

Fairness via Equal True and False Positive Rates

Relaxed Definition (Multiplicative)

A predictor satisfies ε**-Equalized Odds (multiplicative)** if:

$$\left|\frac{\text{Recall}_a}{\text{Recall}_b} - 1\right| \leq \varepsilon \quad \text{and} \quad \left|\frac{\text{Recall}_a^-}{\text{Recall}_b^-} - 1\right| \leq \varepsilon$$

Recall and Recall$^-$ Values in Our Example:

- Male: $\text{Recall}_\text{M} = \frac{20}{20+12} = \frac{20}{32} \approx 0.625$, $\text{Recall}_\text{M}^- = \frac{5}{5+4} = \frac{5}{9} \approx 0.556$
- Female: $\text{Recall}_\text{F} = \frac{36}{36+6} = \frac{36}{42} \approx 0.857$, $\text{Recall}_\text{F}^- = \frac{7}{7+10} = \frac{7}{17} \approx 0.412$
- Recall ratio: $\left|\frac{0.625}{0.857} - 1\right| \approx 0.271$; Recall$^-$ ratio: $\left|\frac{0.556}{0.412} - 1\right| \approx 0.350$

Disparate Impact

Fairness via Outcome Rate Parity

Definition

A predictor satisfies **Disparate Impact** fairness if:

$$\frac{\mathrm{PP}_a}{\mathrm{PP}_b} = 1$$

where $\mathrm{PP} = \frac{\mathrm{TP}+\mathrm{FP}}{\mathrm{Total}}$ is the **Positive Prediction Rate** for a group a or b.

- **Interpretation:** Groups should have similar rates of receiving the positive prediction.
- **Pros:** Easy to check from outcomes, no need for true labels.
- **Cons:** Does not account for underlying qualification differences across groups.
- Disparate impact is often used in legal contexts, e.g., in hiring or lending practices.

Disparate Impact

Fairness via Outcome Rate Parity

Definition

A predictor satisfies **Disparate Impact** fairness if:

$$\frac{\text{PP}_a}{\text{PP}_b} = 1$$

where $\text{PP} = \frac{\text{TP}+\text{FP}}{\text{Total}}$ is the **Positive Prediction Rate** for a group a or b.

Positive Prediction Rates in Our Example:

- Male: $\frac{20+4}{41} = \frac{24}{41} \approx 0.585$
- Female: $\frac{36+10}{59} = \frac{46}{59} \approx 0.780$

Disparate Impact

Fairness via Outcome Rate Parity

Relaxed Definition (Additive)

Disparate impact may be relaxed by requiring:

$$|\mathrm{PP}_a - \mathrm{PP}_b| \leq \varepsilon$$

Positive Prediction Rates in Our Example:

- Male: $\frac{20+4}{41} = \frac{24}{41} \approx 0.585$
- Female: $\frac{36+10}{59} = \frac{46}{59} \approx 0.780$
- Difference: $0.780 - 0.585 = 0.195$

Disparate Impact

Fairness via Outcome Rate Parity

Relaxed Definition (Multiplicative)

Disparate impact may be relaxed by requiring:

$$\left|\frac{\mathrm{PP}_a}{\mathrm{PP}_b} - 1\right| \leq \varepsilon$$

Positive Prediction Rates in Our Example:

- Male: $\frac{20+4}{41} = \frac{24}{41} \approx 0.585$
- Female: $\frac{36+10}{59} = \frac{46}{59} \approx 0.780$
- Ratio: $\left|\frac{0.585}{0.780} - 1\right| \approx 0.25$
- Legally acceptable ratio is often set at 0.8, meaning the positive prediction rate for one group should not be less than 80% of the other group.

Why Are These Notions Important?

Motivating the Limits of Fairness

- Equal Opportunity ensures fairness in beneficial outcomes. In the case of loans, it ensures that individuals who would repay the loan (true positives) are equally likely to be approved, regardless of whether they are male or female.
- Equalized Odds ensures equal error rates across groups. In the loan context, both the true positive rate (recall) and the true negative rate (specificity) are equal for men and women—meaning good payers and bad payers are equally well treated regardless of gender.
- Disparate Impact ensures fairness without needing ground-truth labels. For loans, this means the proportion of applicants who are approved (positive predictions) is similar across gender groups, even without knowing who would actually repay the loan.
- These fairness notions often conflict with each other and with overall performance. For example, a model might increase fairness for one group at the cost of decreased accuracy overall or fairness for another group.

Why Are These Notions Important?

Motivating the Limits of Fairness

- **But are they enough? Are the equivalent?**
- Shouldn't fairness also consider **whether predictions reflect actual probabilities**?
- This leads us to consider three perspectives on fairness: **Independence**, **Separation**, and **Sufficiency**.

Two Perspectives on Fairness

Separation, Sufficiency, and Independence

- **Separation:** People who have the same true outcome should be treated the same by the algorithm, regardless of their group.
- **Sufficiency:** The algorithm's prediction should carry the same meaning for everyone—if two people get the same score, their actual outcomes should be equally likely, regardless of group.
- **Independence:** The probability of receiving a certain prediction (e.g., being approved for a loan) should be the same for all groups, no matter their actual likelihood of repaying.
- **Illustration (Loan Example):**
  - **Separation:** Among those who actually repay, approval rates should be the same for men and women.
  - **Sufficiency:** Among those predicted to repay, the actual repayment rate should be the same for men and women.
  - **Independence:** Approval rates should be the same for men and women, regardless of their true repayment behavior.

Independence

Fairness via Demographic Parity

Definition

A predictor satisfies **Independence** if:

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

Interpretation

The model's predictions should be equally distributed across sensitive groups (e.g., men and women have the same rate of positive predictions, regardless of actual repayment).

- This notion corresponds to **Disparate Impact** when using binary predictions.

Pros:

- Does not require ground-truth labels (can be checked from outcomes alone).
- Simple to compute and legally relevant in some contexts.

Cons:

- Ignores differences in qualification or base rates between groups.
- May require accepting less qualified applicants from one group to equalize rates.

Separation

Fairness via Error Rate Parity

Definition (in terms of performance)

A predictor satisfies **Separation** if:

$$\text{Recall}_a = \text{Recall}_b \quad \text{and} \quad \text{Recall}_a^- = \text{Recall}_b^-$$

Definition (in terms of independence)

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

The prediction is independent of the sensitive attribute given the true label.

- **Note:** This is equivalent to equalized odds.

Sufficiency

Fairness via Calibration Parity

Definition (in terms of performance)

A predictor satisfies **Sufficiency** if:

$$\text{Precision}_a = \text{Precision}_b \quad \text{and} \quad \text{Precision}_a^- = \text{Precision}_b^-$$

Definition (in terms of independence)

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

The true label is independent of the sensitive attribute given the prediction.

Many Definitions of Fairness

Each with a Specific Intuition

- **Statistical parity**: Ensure equal predicted positive rates across groups.
- **Group fairness**: Generic term for parity in outcomes across groups.
- **Demographic parity**: Same as statistical parity, often used in hiring.
- **Conditional statistical parity**: Allow fairness conditioned on certain features.
- **Equal opportunity**: Equal true positive rate across groups.
- **Equalized odds**: Equal true and false positive rates across groups.
- **Conditional procedure accuracy equality**: Accuracy conditioned on true label should be equal.
- **Disparate mistreatment**: Equal error rates across groups.
- **Balance for positive class**: Predictions for positive labels should be similar across groups.
- **Balance for negative class**: Predictions for negative labels should be similar across groups.
- **Predictive equality**: Equal false positive rate across groups.
- **Conditional use accuracy equality**: Accuracy conditioned on prediction should be equal.
- **Predictive parity**: Equal precision across groups.
- **Calibration**: Predicted probabilities should match actual outcomes across groups.

Taxonomy of Fairness Definitions

Equivalence and Relaxations

Name	Closest relative	Note
Statistical parity	Independence	Equivalent
Group fairness	Independence	Equivalent
Demographic parity	Independence	Equivalent
Conditional statistical parity	Independence	Relaxation
Equal opportunity	Separation	Relaxation
Equalized odds	Separation	Equivalent
Conditional procedure accuracy equality	Separation	Equivalent
Disparate mistreatment	Separation	Equivalent
Balance for positive class	Separation	Relaxation
Balance for negative class	Separation	Relaxation
Predictive equality	Separation	Relaxation
Conditional use accuracy equality	Sufficiency	Equivalence
Predictive parity	Sufficiency	Relaxation
Calibration	Sufficiency	Equivalence

Relations shown in Barocas et al. [2023]; Verma and Rubin [2018].

Impossibility of Fairness

The COMPAS Case Study

Background:

- **COMPAS** is a commercial algorithm used in U.S. courts to predict a defendant's risk of reoffending.
- An investigation by ProPublica (2016) revealed that the COMPAS algorithm disproportionately labeled Black defendants as high risk for recidivism, even when they did not reoffend, and white defendants as low risk, even when they did reoffend.
- Specifically, Black defendants were nearly twice as likely as white defendants to be falsely labeled as future criminals (false positives), while white defendants were more often mislabeled as low risk (false negatives).
- These findings highlighted significant racial disparities in algorithmic decision-making and sparked widespread debate about how to define and enforce fairness in automated systems, particularly in high-stakes domains such as criminal justice.

Interpretation:

- The COMPAS score was *calibrated*—i.e., among individuals with the same predicted score, the probability of recidivism was approximately the same across races.
- This means COMPAS satisfied **sufficiency** (score is conditionally independent of sensitive attribute given the outcome).
- However, the false positive and false negative rates differed significantly by race, violating **separation**.
- This case illustrates the incompatibility between fairness criteria: satisfying sufficiency may require violating separation, and vice versa.

Impossibility of Fairness

Theoretical Limits of Fairness

Key Concepts:

- **Base Rates**: The prevalence of a condition in a population (e.g., recidivism rates).
- **Fairness Paradigms**: Different definitions of fairness, such as independence, separation, and sufficiency.

Impossibility Theorem:

- When base rates differ across groups (e.g., differing prevalence of recidivism), it is mathematically impossible to simultaneously satisfy the three fundamental fairness paradigms:
  1. **Independence (Demographic Parity)**: $\hat{Y} \perp A$
  2. **Separation (Equalized Odds)**: $\hat{Y} \perp A \mid Y$
  3. **Sufficiency (Calibration)**: $Y \perp A \mid \hat{Y}$
- In other words, you cannot ensure all three types of fairness at once when groups have different base rates — trade-offs are unavoidable.

Fairness in Generative Models

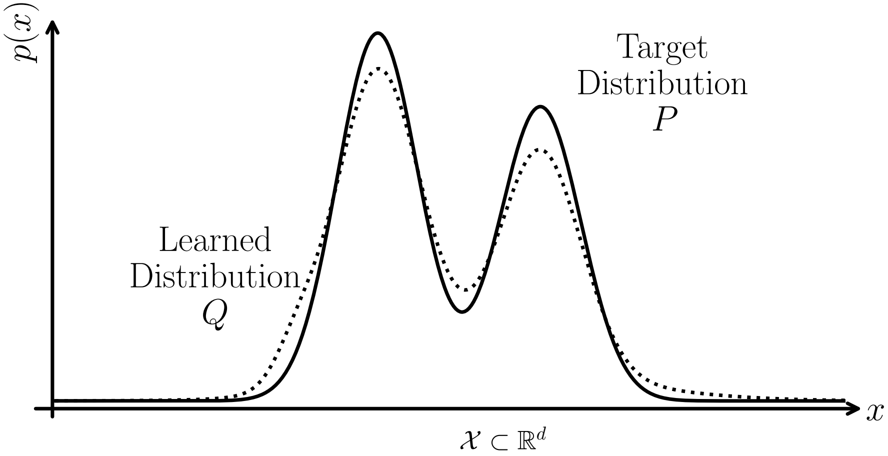
A New Frontier

- Much of the research on algorithmic fairness has focused on:
    - Classification tasks (e.g., loan approval, hiring)
    - Recommendation systems (e.g., job ads, content ranking)
- Comparatively little work has addressed fairness in **generative models** (e.g., image, text, audio synthesis).
- Yet generative models are increasingly deployed in high-impact domains (e.g., education, media, design).
- We highlight two recent efforts:
    - **"Fair Generative Modeling via Weak Supervision"** by Choi et al. [2020]
    - **"On Measuring Fairness in Generative Models"** by Teo et al. [2023]
- We will explore these two approaches and conclude with broader reflections and open questions.

Mathematical Setup

Target vs Learned Distributions

- Let $x \in \mathcal{X} \subset \mathbb{R}^d$ denote the data (e.g., images, text, audio).
- Let P be the **target distribution** (e.g., real-world data).
- Let Q be the **learned distribution** produced by the generative model.
- The model is trained to generate samples $x \sim Q$ such that $Q \approx P$.



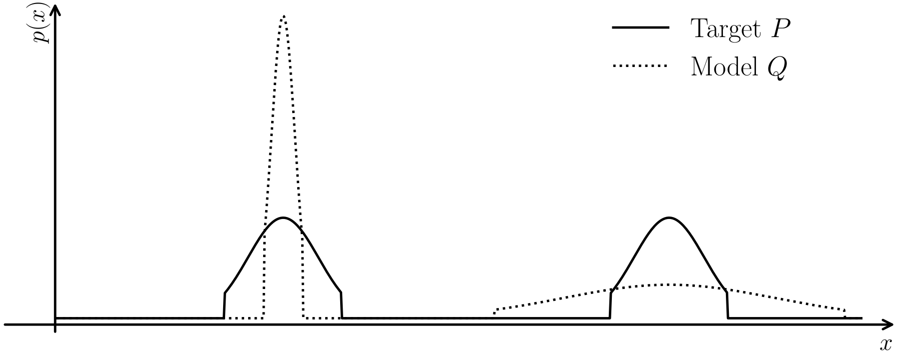
Setup with Attributes

Oracle-based Evaluation

- Suppose an oracle attribute function $\phi : \mathcal{X} \to \mathcal{A}$ maps samples to sensitive attributes (e.g., gender, race).
- Then we can partition the space $\mathcal{X}$ into disjoint subsets based on the attribute values.
- This leads to a decomposition of the distributions:

$$P = \sum_a \pi_a^P P_a \quad \text{and} \quad Q = \sum_a \pi_a^Q Q_a$$

where P_a and Q_a are distributions with disjoint support corresponding to attribute a, and π_a^P, π_a^Q are their mixing proportions.



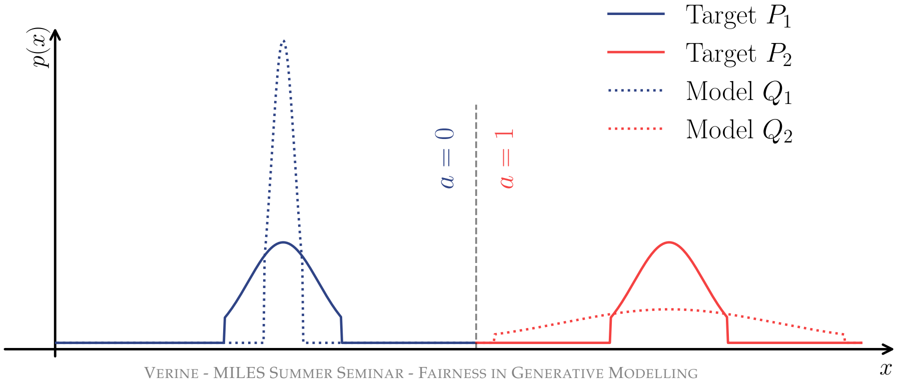
Setup with Attributes

Oracle-based Evaluation

- Suppose an oracle attribute function $\phi : \mathcal{X} \to \mathcal{A}$ maps samples to sensitive attributes (e.g., gender, race).
- Then we can partition the space $\mathcal{X}$ into disjoint subsets based on the attribute values.
- This leads to a decomposition of the distributions:

$$P = \sum_a \pi_a^P P_a \quad \text{and} \quad Q = \sum_a \pi_a^Q Q_a$$

where P_a and Q_a are distributions with disjoint support corresponding to attribute a, and π_a^P, π_a^Q are their mixing proportions.



Measuring Fairness

Disparity in Attribute Distributions

Fairness

What does it mean for Q to be **fair** with respect to target distribution P?

Measuring Fairness

Disparity in Attribute Distributions

Fairness

What does it mean for Q to be **fair** with respect to target distribution P?

- **“Fair Generative Modeling via Weak Supervision”** by Choi et al. [2020] and **“On Measuring Fairness in Generative Models”** by Teo et al. [2023] : Criteria of fairness of Q with respect to P and a:

$$\|\pi_a^P - \pi_a^Q\|_2$$

- **"Improving the Fairness of Deep Generative Models without Retraining"** by Tan et al. [2021]: Criterion of fairness of Q:

$$\mathrm{KL}(\pi^Q \| \mathcal{U}(\mathcal{A}))$$

  where $\mathcal{U}(\mathcal{A})$ is the uniform distribution over attributes.

- **"Fairness in Generative Modeling: do it Unsupervised!"** by Zameshina et al. [2022]: Criterion of fairness of Q with respect to P:

$$1 - \inf_{a \text{ s.t.} \pi_a^Q > 0} \frac{\pi_a^P}{\pi_a^Q}$$

  The criterion is 0 if the target proportions are reached and 1 is one attribute is not present in the generated samples.

Two Notions of Generative Fairness

Attribute Proportions in Generated Data

Equalized Generative Odds (EGO)

A generative model distribution Q satisfies **equalized generative odds** if the attribute proportions are uniform:

$$\forall a, a' \in \mathcal{A}, \quad \pi_a^Q = \pi_{a'}^Q$$

That is, all attribute groups appear equally in generated samples.

Matching Generative Odds (MGO)

Two distributions P and Q are said to have **matching generative odds** if:

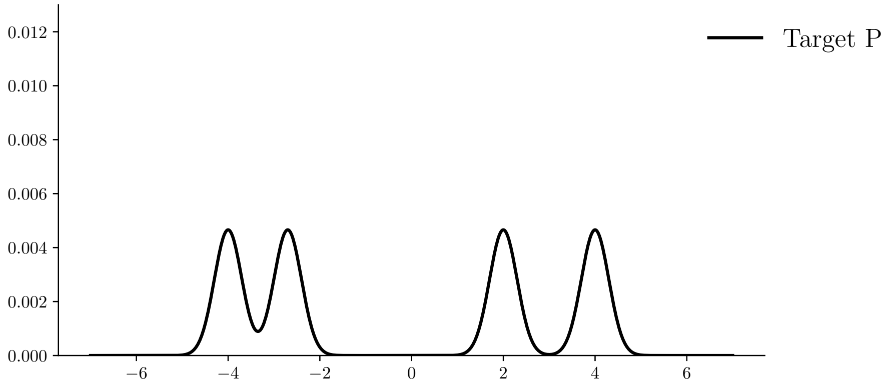
$$\forall a \in \mathcal{A}, \quad \pi_a^Q = \pi_a^P$$

The generated samples reflect the attribute distribution in the training data.

Limitations of Proportion-Based Fairness

EGO and MGO May Still Be Unfair

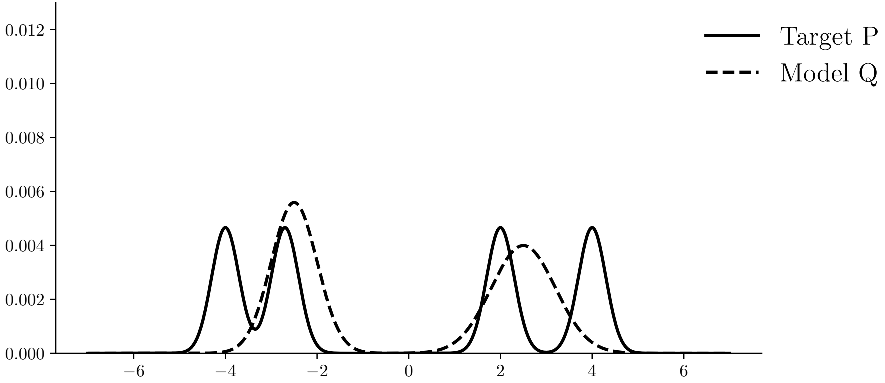
- Both EGO and MGO concern the **global proportions** of attributes in the generated data.
- But they do not account for **how** the samples within each attribute group are distributed.
- It is possible for a model to perfectly match proportions while failing to capture the true diversity or semantics of each group.



Limitations of Proportion-Based Fairness

EGO and MGO May Still Be Unfair

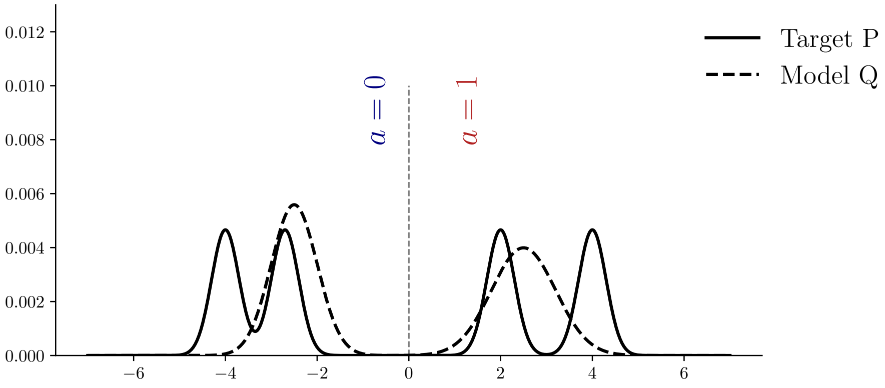
- Both EGO and MGO concern the **global proportions** of attributes in the generated data.
- But they do not account for **how** the samples within each attribute group are distributed.
- It is possible for a model to perfectly match proportions while failing to capture the true diversity or semantics of each group.



Limitations of Proportion-Based Fairness

EGO and MGO May Still Be Unfair

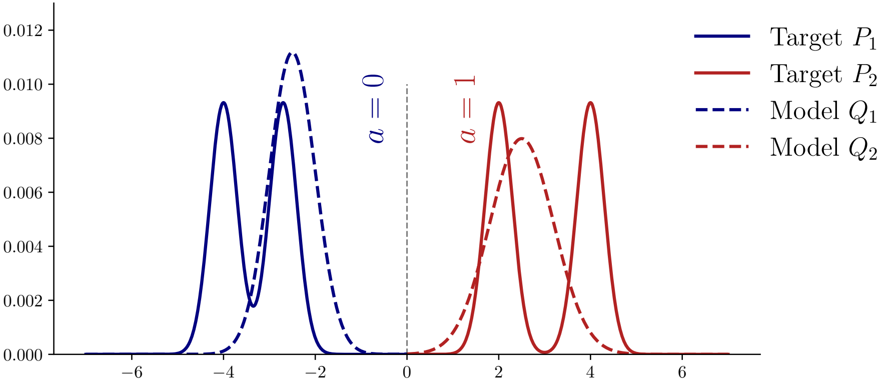
- Both EGO and MGO concern the **global proportions** of attributes in the generated data.
- But they do not account for **how** the samples within each attribute group are distributed.
- It is possible for a model to perfectly match proportions while failing to capture the true diversity or semantics of each group.



Limitations of Proportion-Based Fairness

EGO and MGO May Still Be Unfair

- Both EGO and MGO concern the **global proportions** of attributes in the generated data.
- But they do not account for **how** the samples within each attribute group are distributed.
- It is possible for a model to perfectly match proportions while failing to capture the true diversity or semantics of each group.



- The example below satisfies both EGO and MGO, yet is clearly unfair upon inspection.

Limits of Global Approximation

Local Disparities Remain Possible

Theorem: Local Unfairness Despite Global Fit

Let $P \in \mathcal{P}(\mathcal{X})$ be a target distribution and f be a continuous function such that $\mathcal{D}_f$ defines an f-divergence. Let $\mathcal{S}_{D_f}(P, \epsilon) := \{Q \in \mathcal{P}(\mathcal{X}) \mid D_f(P \| Q) = \epsilon\}$.
For any $\epsilon \in (0, f(0) + \bar{f}(\infty))$ and $\gamma \in (0, \epsilon)$, there exists a distribution $Q^\gamma \in \mathcal{S}_{\mathcal{D}_f}(P, \epsilon)$ such that:

- Q^γ satisfies both Equalized Generative Odds and Matching Generative Odds w.r.t. P.
- Yet, there exists $\bar{a} \in \mathcal{A}$ for which:

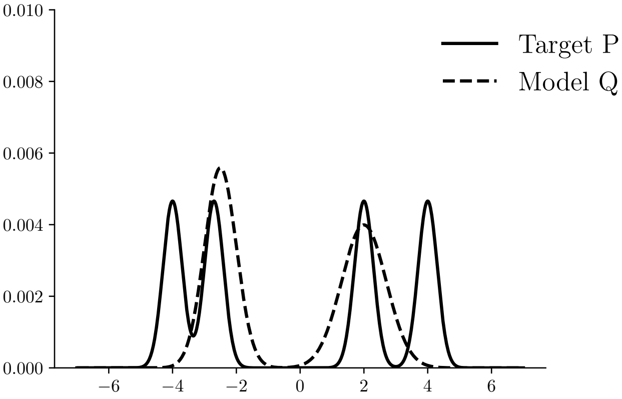
$$\mathcal{D}_f(P_{\bar{a}} \| Q^\gamma_{\bar{a}}) \geq \mathcal{D}_f(P_a \| Q^\gamma_a) + \gamma, \quad \forall a \in \mathcal{A} \setminus \{\bar{a}\}.$$

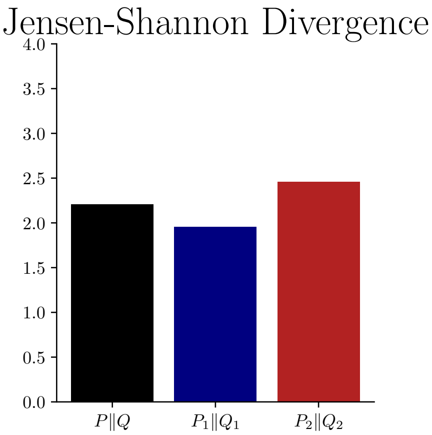
- A small global divergence does not guarantee local small divergences.

Limits of Global Approximation

Example

$$\min_{Q \in \mathcal{Q}} \mathcal{D}_{\mathrm{JS}}(P \| Q)$$





Limits of Global Approximation

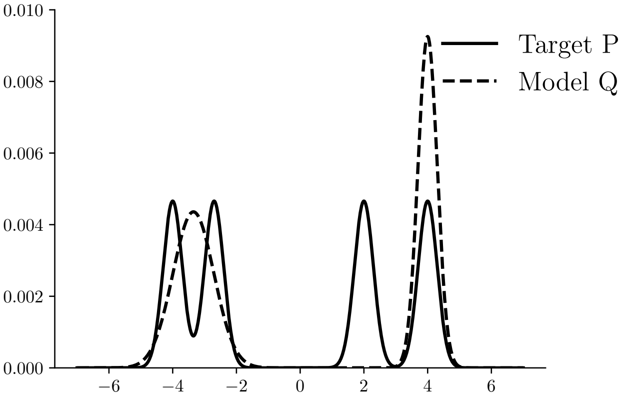
Example

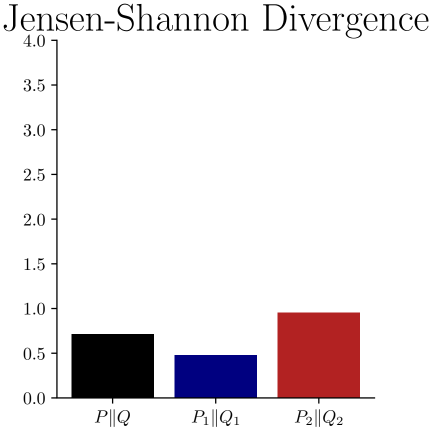
$$\min_{Q \in \mathcal{Q}} \mathcal{D}_{\mathrm{JS}}(P \| Q)$$

Limits of Global Approximation

Example

$$\min_{Q \in \mathcal{Q}} \mathcal{D}_{\text{JS}}(P \| Q)$$





Towards Stronger Generative Fairness

Equalized Generative Treatment

Definition: Equalized Generative Treatment (EGT)

Let $P, Q \in \mathcal{P}(\mathcal{X})$ be two probability measures with $P \ll Q$ and let f be a function such that $\mathcal{D}_f$ defines an f-divergence. For any $\delta > 0$, we say that Q satisfies δ-*equalized generative treatment* of P with respect to $\mathcal{D}_f$ if:

$$\left|\mathcal{D}_f(P_a \| Q_a) - \mathcal{D}_f(P_{a'} \| Q_{a'})\right| \leq \delta, \quad \forall a, a' \in \mathcal{A}.$$

When $\delta = 0$, we say that Q satisfies *equalized generative treatment* of P with respect to $\mathcal{D}_f$.

- Controls **treatment disparity** between groups in the **distributional approximation** step.
- Can be instantiated with common f-divergences: KL, Total Variation, JS, etc.
- Stronger than EGO and MGO: accounts for within-group quality, not just proportions.

Achievability of Fair Generative Treatment

Existence Guarantee

Proposition: Existence of Fair Models

Let $P \in \mathcal{P}(\mathcal{X})$ be a target probability measure and f be a continuous function such that $\mathcal{D}_f$ defines an f-divergence. For any $\epsilon \in (0, f(0) + \bar{f}(\infty))$, there exists a distribution $Q \in \mathcal{S}_{\mathcal{D}_f}(P, \epsilon)$ such that:

- Q satisfies **Matching Generative Odds** w.r.t. P
- Q satisfies **Equalized Generative Treatment** w.r.t. P and $\mathcal{D}_f$

- Such a model achieves both global fidelity and group-level fairness in treatment.
- Demonstrates that the fairness goal is not inherently infeasible — it is attainable under divergence constraints.

Optimization Limits in Practice

Decomposition of Search Space

- In practice, the generator Q is optimized via a neural network.
- $\mathcal{Q}_a$ denotes the set of **reachable conditional distributions** for attribute a:

$$\mathcal{Q}_a := \{R \in \mathcal{P}(\mathcal{X}) \mid \exists Q \in \mathcal{Q} \text{ s.t. } Q_a = R\}$$

- $\Delta_{\mathcal{Q}}$ captures all **reachable mixing proportions**:

$$\Delta_{\mathcal{Q}} := \left\{(\pi_a)_{a \in \mathcal{A}} \in \Delta(\mathcal{A}) \mid \exists Q \in \mathcal{Q} \text{ s.t. } \left(\pi_a^Q\right)_{a \in \mathcal{A}} = \pi_a\right\}$$

- $\bar{\mathcal{Q}}^{\mathcal{A}}$ represents the **idealized search space** if one could freely choose each Q_a and mixing weights:

$$\bar{\mathcal{Q}}^{\mathcal{A}} := \left\{Q \in \mathcal{P}(\mathcal{X}) \mid Q = \sum_{a \in \mathcal{A}} \pi_a^Q Q_a,\ Q_a \in \mathcal{Q}_a,\ \left(\pi_a^Q\right) \in \Delta_{\mathcal{Q}}\right\}$$

Lower Bound on Divergence

Limit of Fairness in Restricted Models

Theorem

Let $P \in \mathcal{P}(\mathcal{X})$ be a target probability measure and f be a function such that $\mathcal{D}_f$ defines an f-divergence.
Let $\mathcal{Q}$ be a set of candidate generators satisfying matching odds with P. Assume:

$$Q^* \in \operatorname*{argmin}_{Q \in \bar{\mathcal{Q}}^{\mathcal{A}}} \mathcal{D}_f(P \| Q)$$

Then, for any $\delta > 0$, if $Q \in \mathcal{Q}$ satisfies δ-equalized generative treatment of P w.r.t. $\mathcal{D}_f$, then:

$$\mathcal{D}_f(P \| Q) \geq \max_{a \in \mathcal{A}} \mathcal{D}_f(P_a \| Q_a^*) - \delta$$

- Even the best fair approximation is lower-bounded by the hardest subgroup to approximate in the idealized space.
- Emphasizes the difficulty of achieving both fairness and global fidelity with realistic generators.

Training Fair Generative Models

Optimization Strategies

- Achieving fairness in generative models requires **explicit integration of fairness constraints** during training.
- Two main paradigms can be considered:
    - **Min-Max Formulation:**

$$\min_{Q\in\mathcal{Q}}\max_{a\in\mathcal{A}} D_f(P_a\|Q_a)$$

    - **Regularized Minimization:**

$$\min_{Q}\mathcal{D}_f(P\|Q)+\lambda\sum_{a,a'\in\mathcal{A}}|\mathcal{D}_f(P_a\|Q_a)-\mathcal{D}_f(P_{a'}\|Q_{a'})|$$

for a fixed trade-off parameter $\lambda > 0$.

Training Fair Generative Models

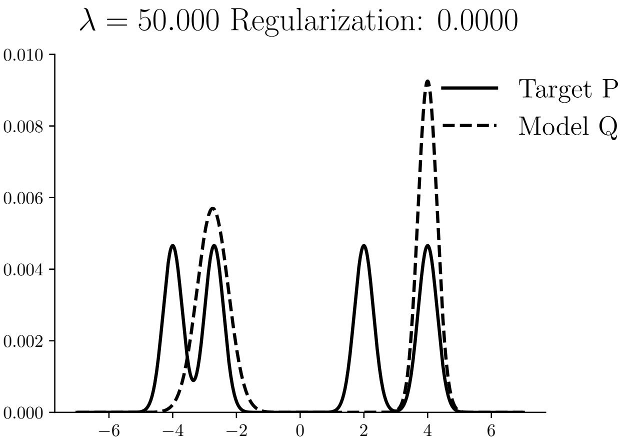
Example

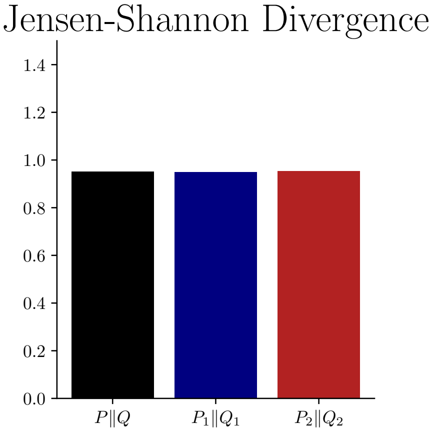
$$\min_{Q} \mathcal{D}_{\mathrm{JS}}(P \| Q) + \lambda \left(\mathcal{D}_{\mathrm{JS}}(P_1 \| Q_1) - \mathcal{D}_{\mathrm{JS}}(P_2 \| Q_2) \right)^2$$

Training Fair Generative Models

Example

$$\min_{Q} \mathcal{D}_{\mathrm{JS}}(P\|Q) + \lambda \left(\mathcal{D}_{\mathrm{JS}}(P_1\|Q_1) - \mathcal{D}_{\mathrm{JS}}(P_2\|Q_2)\right)^2$$

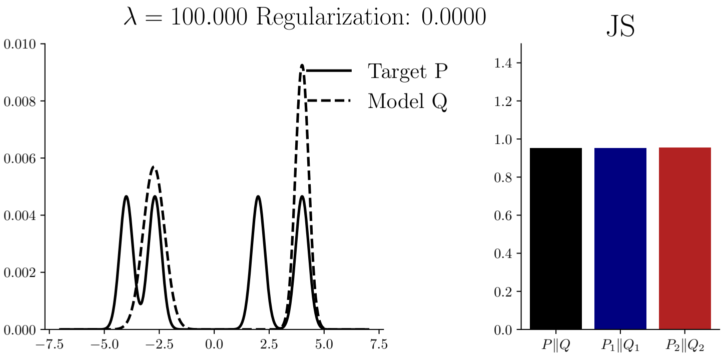




f-Divergence Sensitivity

Diversity–Quality Trade-offs

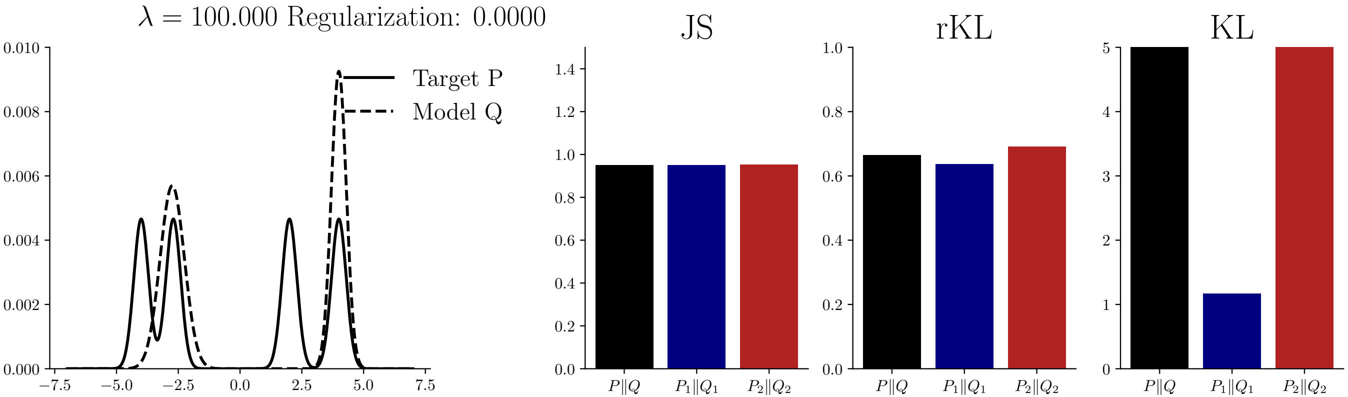
- Different f-divergences correspond to different trade-offs between fidelity (sample quality) and coverage (diversity).
- For instance:
  - KL divergence emphasizes recall: better coverage of P but may include low-quality samples.
  - Reverse KL emphasizes precision: avoids generating from low-density areas in P.
  - JS divergence offers a balance between the two.
- These trade-offs suggest that fairness should be assessed across multiple divergences.



f-Divergence Sensitivity

Diversity–Quality Trade-offs

- Different *f*-divergences correspond to different trade-offs between fidelity (sample quality) and coverage (diversity).
- For instance:
  - KL divergence emphasizes recall: better coverage of P but may include low-quality samples.
  - Reverse KL emphasizes precision: avoids generating from low-density areas in P.
  - JS divergence offers a balance between the two.
- These trade-offs suggest that fairness should be assessed across multiple divergences.



Extending EGT Across Divergences

Diversity–Fairness Trade-offs

Definition: Extended Equalized Generative Treatment (Multi-divergence EGT)

Let $P, Q \in \mathcal{P}(\mathcal{X})$ be two probability measures with $P \ll Q$, and let $\mathcal{F}$ be a family of continuous functions such that each $\mathcal{D}_f$ defines an f-divergence.

For a family of thresholds $(\delta_f)_{f\in\mathcal{F}}$ with each $\delta_f > 0$, we say that Q satisfies *extended equalized generative treatment* of P with respect to $\mathcal{F}$ if:

$$\left|\mathcal{D}_f(P_a\|Q_a) - \mathcal{D}_f(P_{a'}\|Q_{a'})\right| \leq \delta_f, \quad \forall a, a' \in \mathcal{A}, \ \forall f \in \mathcal{F}$$

When $\delta_f = 0$ for all $f \in \mathcal{F}$, we say that Q satisfies *exact extended equalized generative treatment* of P with respect to $\mathcal{F}$.

- This encourages balanced treatment across groups for a range of divergence sensitivities.

Extending EGT Across Divergence

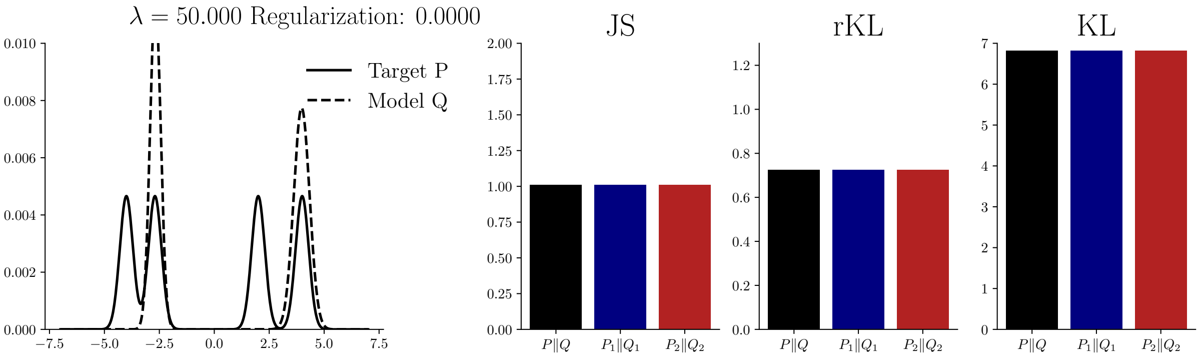
Example

$$\min_{Q} \mathcal{D}_{\text{JS}}(P\|Q) + \lambda \sum_{f\in\mathcal{F}} \left(\mathcal{D}_f(P_1\|Q_1) - \mathcal{D}_f(P_2\|Q_2)\right)^2 \quad \text{with } \mathcal{F} = \{f_{\text{JS}}, f_{\text{KL}}, f_{\text{rKL}}\}$$

Extending EGT Across Divergence

Example

$$\min_{Q} \mathcal{D}_{\mathrm{JS}}(P\|Q) + \lambda \sum_{f\in\mathcal{F}} \left(\mathcal{D}_f(P_1\|Q_1) - \mathcal{D}_f(P_2\|Q_2)\right)^2 \quad \text{with } \mathcal{F} = \{f_{\mathrm{JS}}, f_{\mathrm{KL}}, f_{\mathrm{rKL}}\}$$



Extending EGT Across Divergence

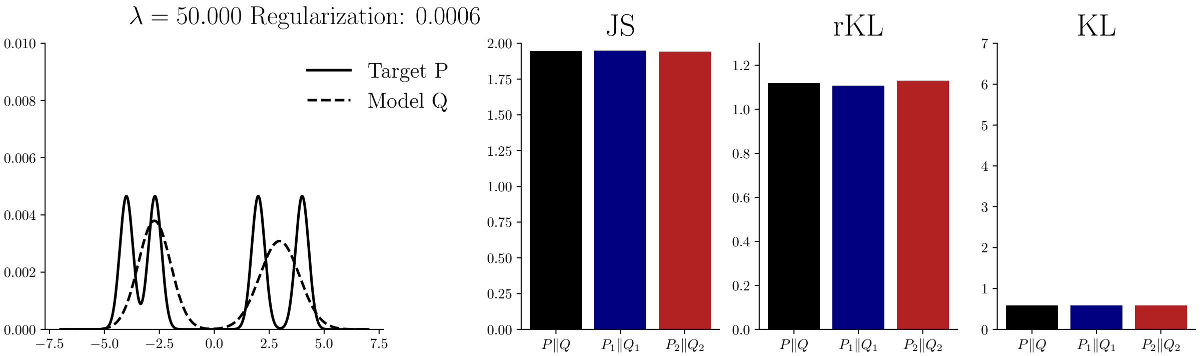
Example

$$\min_{Q} \sum_{f\in\mathcal{F}} \mathcal{D}_f(P\|Q) + \lambda \sum_{f\in\mathcal{F}} \big(\mathcal{D}_f(P_1\|Q_1) - \mathcal{D}_f(P_2\|Q_2)\big)^2 \quad \text{with } \mathcal{F} = \{f_{\text{JS}}, f_{\text{KL}}, f_{\text{rKL}}\}$$

Extending EGT Across Divergence

Example

$$\min_{Q} \sum_{f \in \mathcal{F}} \mathcal{D}_f(P \| Q) + \lambda \sum_{f \in \mathcal{F}} \left(\mathcal{D}_f(P_1 \| Q_1) - \mathcal{D}_f(P_2 \| Q_2)\right)^2 \quad \text{with } \mathcal{F} = \{f_{\text{JS}}, f_{\text{KL}}, f_{\text{rKL}}\}$$



Conclusion and Open Questions

Where Do We Go From Here?

- Generative fairness is a young and evolving field.
- We introduced several fairness notions:
  - Equalized Generative Odds (EGO)
  - Matching Generative Odds (MGO)
  - Equalized Generative Treatment (EGT)
- We saw that global similarity to the target data does not imply group fairness.
- Fairness-aware training objectives can guide model development, but trade-offs remain.

Conclusion and Open Questions

Where Do We Go From Here?

Open Questions:

- Can we classify fairness definitions for generative models similarly to classification (separation, sufficiency, independence)?
- Are there impossibility results in the generative case?
- What are the most effective algorithms for training fair generative models in practice?

Next Goal:

- Observing EGO and MGO and but EGT in LLMS and Diffusion models.
- Training/Finetuning large scale models with fairness criterion.

References I

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and Machine Learning*. Mit press edition.

Choi, K., Grover, A., Singh, T., Shu, R., and Ermon, S. (2020). Fair Generative Modeling via Weak Supervision. arXiv:1910.12008 [cs, stat].

Tan, S., Shen, Y., and Zhou, B. (2021). Improving the Fairness of Deep Generative Models without Retraining. arXiv:2012.04842 [cs].

Teo, C. T. H., Abdollahzadeh, M., and Cheung, N.-M. (2023). On Measuring Fairness in Generative Models.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, pages 1–7, New York, NY, USA. Association for Computing Machinery.

Zameshina, M., Teytaud, O., Teytaud, F., Hosu, V., Carraz, N., Najman, L., and Wagner, M. (2022). Fairness in generative modeling. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 320–323. arXiv:2210.03517 [cs].