

Differential Privacy and Fairness in Automated Decision Making

Arijit Das, PhD

IFoA Research Section Knowledge Sharing Session

February 1, 2022

About Me

- *2022-Present, Senior Data Scientist, ERGO Group AG, Duesseldorf*
- 2017-2021, Independent Research Group Leader, University Hospital, Cologne
Selfsupervised anomaly detection in medical images
- 2012-2017, PhD Machine Learning and Computational Biology, Max Plank Institute, Cologne.
Design and analysis of learning algorithms which control false discoveries

About Me

- *2022-Present, Senior Data Scientist, ERGO Group AG, Duesseldorf*
- 2017-2021, Independent Research Group Leader, University Hospital, Cologne
Selfsupervised anomaly detection in medical images
- 2012-2017, PhD Machine Learning and Computational Biology, Max Plank Institute, Cologne.
Design and analysis of learning algorithms which control false discoveries

About Me

- *2022-Present, Senior Data Scientist, ERGO Group AG, Duesseldorf*
- 2017-2021, Independent Research Group Leader, University Hospital, Cologne
Selfsupervised anomaly detection in medical images
- 2012-2017, PhD Machine Learning and Computational Biology, Max Plank Institute, Cologne.
Design and analysis of learning algorithms which control false discoveries

Outline

- 1 Automated Decision Making
 - Introduction
 - Bias in Machine Learning
- 2 Measuring Privacy and Fairness
 - Differential Privacy
 - Notions of Fairness
 - Impossibility Results
- 3 Learning Differentially Private and Fair Models
 - Fair Empirical Risk Minimization
 - UCI Credit-Card Default Dataset

Outline

- 1 Automated Decision Making
 - Introduction
 - Bias in Machine Learning
- 2 Measuring Privacy and Fairness
 - Differential Privacy
 - Notions of Fairness
 - Impossibility Results
- 3 Learning Differentially Private and Fair Models
 - Fair Empirical Risk Minimization
 - UCI Credit-Card Default Dataset

Differential Privacy

- Consider the credit risk scoring problem, where given a new applicant, the financial institution has to decide if they should get the loan.
- Increasingly, machine learning frameworks are being used to make these decisions, which are trained on a database of past clients.
- Assume, that a past client wants their data removed from future use (GDPR 2016).
- Question: Would this removal lead to a significant change in the learned model parameters?

Differential Privacy

- Consider the credit risk scoring problem, where given a new applicant, the financial institution has to decide if they should get the loan.
- Increasingly, machine learning frameworks are being used to make these decisions, which are trained on a database of past clients.
- Assume, that a past client wants their data removed from future use (GDPR 2016).
- Question: Would this removal lead to a significant change in the learned model parameters?

Differential Privacy

- Consider the credit risk scoring problem, where given a new applicant, the financial institution has to decide if they should get the loan.
- Increasingly, machine learning frameworks are being used to make these decisions, which are trained on a database of past clients.
- Assume, that a past client wants their data removed from future use (GDPR 2016).
- Question: Would this removal lead to a significant change in the learned model parameters?

Differential Privacy

- Consider the credit risk scoring problem, where given a new applicant, the financial institution has to decide if they should get the loan.
- Increasingly, machine learning frameworks are being used to make these decisions, which are trained on a database of past clients.
- Assume, that a past client wants their data removed from future use (GDPR 2016).
- Question: Would this removal lead to a significant change in the learned model parameters?

Fairness

- Automated decision making also leads to decreasing human involvement, which raises concerns of bias and discrimination based on race, gender, sexual orientation, etc.
- Perceived lack of transparency, leads to less satisfaction among clients.
- Question: Can we provably ensure “fairness” among the protected classes as required by GDPR 2016?

Fairness

- Automated decision making also leads to decreasing human involvement, which raises concerns of bias and discrimination based on race, gender, sexual orientation, etc.
- Perceived lack of transparency, leads to less satisfaction among clients.
- Question: Can we provably ensure “fairness” among the protected classes as required by GDPR 2016?

Fairness

- Automated decision making also leads to decreasing human involvement, which raises concerns of bias and discrimination based on race, gender, sexual orientation, etc.
- Perceived lack of transparency, leads to less satisfaction among clients.
- Question: Can we provably ensure “fairness” among the protected classes as required by GDPR 2016?

Achieving Balance

- The financial institutions are concerned with achieving high accuracy, which directly translates into profits.
- Unfortunately, the three requirements of privacy, fairness and accuracy cannot be satisfied at the same time.
- Question: How does one balance privacy and fairness in automated decision making while maintaining high accuracy?

Achieving Balance

- The financial institutions are concerned with achieving high accuracy, which directly translates into profits.
- Unfortunately, the three requirements of privacy, fairness and accuracy cannot be satisfied at the same time.
- Question: How does one balance privacy and fairness in automated decision making while maintaining high accuracy?

Achieving Balance

- The financial institutions are concerned with achieving high accuracy, which directly translates into profits.
- Unfortunately, the three requirements of privacy, fairness and accuracy cannot be satisfied at the same time.
- Question: How does one balance privacy and fairness in automated decision making while maintaining high accuracy?

Outline

- 1 Automated Decision Making
 - Introduction
 - Bias in Machine Learning
- 2 Measuring Privacy and Fairness
 - Differential Privacy
 - Notions of Fairness
 - Impossibility Results
- 3 Learning Differentially Private and Fair Models
 - Fair Empirical Risk Minimization
 - UCI Credit-Card Default Dataset

Machine Learning models are objective

- Machine Learning models can be assumed to be objective.
- When well specified, they learn to make decisions, purely based on data
- Removes any subjectivity or bias which humans may have

Machine Learning models are objective

- Machine Learning models can be assumed to be objective.
- When well specified, they learn to make decisions, purely based on data
- Removes any subjectivity or bias which humans may have

Machine Learning models are objective

- Machine Learning models can be assumed to be objective.
- When well specified, they learn to make decisions, purely based on data
- Removes any subjectivity or bias which humans may have

Models maybe objective but ...

- Labeling Bias:

- They imbibe the inherent social/ideological biases in historical data, which is usually against marginalized groups.

- Algorithmic Bias:

- Unbalanced sampling and noisy data from certain sub-groups, nudges the algorithm to favor dominant sub-groups with better sampling.

- Intervention Bias:

- Past decisions, algorithmic or otherwise, lead to bias in newly generated data.

- Model Interpretability:

- The GDPR 2016 EU regulation requires the right to explanation for any decisions made. However, opaque/black box models make it hard to detect and thereby correct biases using traditional approaches.

Models maybe objective but ...

- Labeling Bias:
 - They imbibe the inherent social/ideological biases in historical data, which is usually against marginalized groups.
- Algorithmic Bias:
 - Unbalanced sampling and noisy data from certain sub-groups, nudges the algorithm to favor dominant sub-groups with better sampling.
- Intervention Bias:
 - Past decisions, algorithmic or otherwise, lead to bias in newly generated data.
- Model Interpretability:
 - The GDPR 2016 EU regulation requires the right to explanation for any decisions made. However, opaque/black box models make it hard to detect and thereby correct biases using traditional approaches.

Models maybe objective but ...

- Labeling Bias:
 - They imbibe the inherent social/ideological biases in historical data, which is usually against marginalized groups.
- Algorithmic Bias:
 - Unbalanced sampling and noisy data from certain sub-groups, nudges the algorithm to favor dominant sub-groups with better sampling.
- Intervention Bias:
 - Past decisions, algorithmic or otherwise, lead to bias in newly generated data.
- Model Interpretability:
 - The GDPR 2016 EU regulation requires the right to explanation for any decisions made. However, opaque/black box models make it hard to detect and thereby correct biases using traditional approaches.

Models maybe objective but ...

- Labeling Bias:
 - They imbibe the inherent social/ideological biases in historical data, which is usually against marginalized groups.
- Algorithmic Bias:
 - Unbalanced sampling and noisy data from certain sub-groups, nudges the algorithm to favor dominant sub-groups with better sampling.
- Intervention Bias:
 - Past decisions, algorithmic or otherwise, lead to bias in newly generated data.
- Model Interpretability:
 - The GDPR 2016 EU regulation requires the right to explanation for any decisions made. However, opaque/black box models make it hard to detect and thereby correct biases using traditional approaches.

Models maybe objective but ...

- Labeling Bias:
 - They imbibe the inherent social/ideological biases in historical data, which is usually against marginalized groups.
- Algorithmic Bias:
 - Unbalanced sampling and noisy data from certain sub-groups, nudges the algorithm to favor dominant sub-groups with better sampling.
- Intervention Bias:
 - Past decisions, algorithmic or otherwise, lead to bias in newly generated data.
- Model Interpretability:
 - The GDPR 2016 EU regulation requires the right to explanation for any decisions made. However, opaque/black box models make it hard to detect and thereby correct biases using traditional approaches.

Models maybe objective but ...

- Labeling Bias:
 - They imbibe the inherent social/ideological biases in historical data, which is usually against marginalized groups.
- Algorithmic Bias:
 - Unbalanced sampling and noisy data from certain sub-groups, nudges the algorithm to favor dominant sub-groups with better sampling.
- Intervention Bias:
 - Past decisions, algorithmic or otherwise, lead to bias in newly generated data.
- Model Interpretability:
 - The GDPR 2016 EU regulation requires the right to explanation for any decisions made. However, opaque/black box models make it hard to detect and thereby correct biases using traditional approaches.

Models maybe objective but ...

- Labeling Bias:
 - They imbibe the inherent social/ideological biases in historical data, which is usually against marginalized groups.
- Algorithmic Bias:
 - Unbalanced sampling and noisy data from certain sub-groups, nudges the algorithm to favor dominant sub-groups with better sampling.
- Intervention Bias:
 - Past decisions, algorithmic or otherwise, lead to bias in newly generated data.
- Model Interpretability:
 - The GDPR 2016 EU regulation requires the right to explanation for any decisions made. However, opaque/black box models make it hard to detect and thereby correct biases using traditional approaches.

Models maybe objective but ...

- Labeling Bias:
 - They imbibe the inherent social/ideological biases in historical data, which is usually against marginalized groups.
- Algorithmic Bias:
 - Unbalanced sampling and noisy data from certain sub-groups, nudges the algorithm to favor dominant sub-groups with better sampling.
- Intervention Bias:
 - Past decisions, algorithmic or otherwise, lead to bias in newly generated data.
- Model Interpretability:
 - The GDPR 2016 EU regulation requires the right to explanation for any decisions made. However, opaque/black box models make it hard to detect and thereby correct biases using traditional approaches.

Removing bias is not easy

- Removing sensitive variables from dataset: Leads to redlining and reverse tokenism.
- Incomplete data can be recovered with high probability. Example: Matrix Completion Algorithms. The phenomena is also known as “Birds of a feather flock together”.
- Demographic parity is not enough. Example: Gerrymandering in US elections
- Important: There is no ground truth, hence it is empirically impossible to be unbiased without making explicit assumptions.

Removing bias is not easy

- Removing sensitive variables from dataset: Leads to redlining and reverse tokenism.
- Incomplete data can be recovered with high probability. Example: Matrix Completion Algorithms. The phenomena is also known as “Birds of a feather flock together”.
- Demographic parity is not enough. Example: Gerrymandering in US elections
- Important: There is no ground truth, hence it is empirically impossible to be unbiased without making explicit assumptions.

Removing bias is not easy

- Removing sensitive variables from dataset: Leads to redlining and reverse tokenism.
- Incomplete data can be recovered with high probability. Example: Matrix Completion Algorithms. The phenomena is also known as “Birds of a feather flock together”.
- Demographic parity is not enough. Example: Gerrymandering in US elections
- Important: There is no ground truth, hence it is empirically impossible to be unbiased without making explicit assumptions.

Removing bias is not easy

- Removing sensitive variables from dataset: Leads to redlining and reverse tokenism.
- Incomplete data can be recovered with high probability. Example: Matrix Completion Algorithms. The phenomena is also known as “Birds of a feather flock together”.
- Demographic parity is not enough. Example: Gerrymandering in US elections
- Important: There is no ground truth, hence it is empirically impossible to be unbiased without making explicit assumptions.

Outline

- 1 Automated Decision Making
 - Introduction
 - Bias in Machine Learning
- 2 Measuring Privacy and Fairness
 - Differential Privacy
 - Notions of Fairness
 - Impossibility Results
- 3 Learning Differentially Private and Fair Models
 - Fair Empirical Risk Minimization
 - UCI Credit-Card Default Dataset

Differential Privacy

- Clearly, a person's privacy cannot be compromised by a statistical model based on a dataset if their data was never in it.
- Consequently, the goal in differential privacy (Dwork 2006, Dwork et al. 2006b) is to ensure that each individual in the dataset roughly has the same privacy, as if their data was removed.
- Let $\mathcal{M}(X)$ and $\mathcal{M}(X')$ be a randomized algorithm learned based on two data sets X, X' that differ only at one point. If for all possible outputs y

$$\Pr(\mathcal{M}(X) = y) \leq \exp(\epsilon) \Pr(\mathcal{M}(X') = y) + \delta$$

Then \mathcal{M} is defined to be (ϵ, δ) -differentially private.

Differential Privacy

- Clearly, a person's privacy cannot be compromised by a statistical model based on a dataset if their data was never in it.
- Consequently, the goal in differential privacy (Dwork 2006, Dwork et al. 2006b) is to ensure that each individual in the dataset roughly has the same privacy, as if their data was removed.
- Let $\mathcal{M}(X)$ and $\mathcal{M}(X')$ be a randomized algorithm learned based on two data sets X, X' that differ only at one point. If for all possible outputs y

$$\Pr(\mathcal{M}(X) = y) \leq \exp(\epsilon) \Pr(\mathcal{M}(X') = y) + \delta$$

Then \mathcal{M} is defined to be (ϵ, δ) -differentially private.

Differential Privacy

- Clearly, a person's privacy cannot be compromised by a statistical model based on a dataset if their data was never in it.
- Consequently, the goal in differential privacy (Dwork 2006, Dwork et al. 2006b) is to ensure that each individual in the dataset roughly has the same privacy, as if their data was removed.
- Let $\mathcal{M}(X)$ and $\mathcal{M}(X')$ be a randomized algorithm learned based on two data sets X, X' that differ only at one point. If for all possible outputs y

$$\Pr(\mathcal{M}(X) = y) \leq \exp(\epsilon) \Pr(\mathcal{M}(X') = y) + \delta$$

Then \mathcal{M} is defined to be (ϵ, δ) -differentially private.

Differential Privacy

- Thus we want to limit the statistical influence of each individual data point on the learning algorithm.
- Consider the modified empirical risk minimization problem $\hat{\theta}_{\epsilon, z} := \arg \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n L(z_i, \theta) + \epsilon L(z, \theta)$. A classic result (Cook & Weisberg, 1982) tells us that the influence of upweighting z on the parameters θ is given by

$$\mathcal{I}(z) := \frac{d\hat{\theta}_{\epsilon, z}}{d\epsilon} \Big|_{\epsilon=0} = -H_{\hat{\theta}}^{-1} \nabla_{\theta} L(z, \hat{\theta})$$

where $H_{\hat{\theta}} := \frac{1}{n} \sum_{i=1}^n \nabla_{\theta}^2 L(z_i, \hat{\theta})$ is the Hessian and is positive definite by assumption.

- A differentially private learning algorithm would simply limit the influence of each data point by adding the above term as a constraint in the empirical risk minimization problem.

Differential Privacy

- Thus we want to limit the statistical influence of each individual data point on the learning algorithm.
- Consider the modified empirical risk minimization problem $\hat{\theta}_{\epsilon, z} := \arg \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n L(z_i, \theta) + \epsilon L(z, \theta)$. A classic result (Cook & Weisberg, 1982) tells us that the influence of upweighting z on the parameters θ is given by

$$\mathcal{I}(z) := \frac{d\hat{\theta}_{\epsilon, z}}{d\epsilon} \Big|_{\epsilon=0} = -H_{\hat{\theta}}^{-1} \nabla_{\theta} L(z, \hat{\theta})$$

where $H_{\hat{\theta}} := \frac{1}{n} \sum_{i=1}^n \nabla_{\theta}^2 L(z_i, \hat{\theta})$ is the Hessian and is positive definite by assumption.

- A differentially private learning algorithm would simply limit the influence of each data point by adding the above term as a constraint in the empirical risk minimization problem.

Differential Privacy

- Thus we want to limit the statistical influence of each individual data point on the learning algorithm.
- Consider the modified empirical risk minimization problem $\hat{\theta}_{\epsilon, z} := \arg \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n L(z_i, \theta) + \epsilon L(z, \theta)$. A classic result (Cook & Weisberg, 1982) tells us that the influence of upweighting z on the parameters θ is given by

$$\mathcal{I}(z) := \frac{d\hat{\theta}_{\epsilon, z}}{d\epsilon} \Big|_{\epsilon=0} = -H_{\hat{\theta}}^{-1} \nabla_{\theta} L(z, \hat{\theta})$$

where $H_{\hat{\theta}} := \frac{1}{n} \sum_{i=1}^n \nabla_{\theta}^2 L(z_i, \hat{\theta})$ is the Hessian and is positive definite by assumption.

- A differentially private learning algorithm would simply limit the influence of each data point by adding the above term as a constraint in the empirical risk minimization problem.

Outline

- 1 Automated Decision Making
 - Introduction
 - Bias in Machine Learning
- 2 Measuring Privacy and Fairness
 - Differential Privacy
 - **Notions of Fairness**
 - Impossibility Results
- 3 Learning Differentially Private and Fair Models
 - Fair Empirical Risk Minimization
 - UCI Credit-Card Default Dataset

Notions of Fairness

- Well calibrated prediction: Predictions on random subgroups match the predictions over the whole population.
 - Example: Medical testing and diagnosis - Is decision-making applied uniformly across different groups of patients?
- Balance for the positive class i.e. FPR: the average score received by people constituting positive instances should be the same in each subgroup.
 - Example: Different advertising and commercial content on social media, maybe shown based on gender or racial groups.
 - Females maybe shown lower-paying job adverts compared to equally qualified males.

Notions of Fairness

- Well calibrated prediction: Predictions on random subgroups match the predictions over the whole population.
 - Example: Medical testing and diagnosis - Is decision-making applied uniformly across different groups of patients?
- Balance for the positive class i.e. FPR: the average score received by people constituting positive instances should be the same in each subgroup.
 - Example: Different advertising and commercial content on social media, maybe shown based on gender or racial groups.
 - Females maybe shown lower-paying job adverts compared to equally qualified males.

Notions of Fairness

- Well calibrated prediction: Predictions on random subgroups match the predictions over the whole population.
 - Example: Medical testing and diagnosis - Is decision-making applied uniformly across different groups of patients?
- Balance for the positive class i.e. FPR: the average score received by people constituting positive instances should be the same in each subgroup.
 - Example: Different advertising and commercial content on social media, maybe shown based on gender or racial groups.
 - Females maybe shown lower-paying job adverts compared to equally qualified males.

Notions of Fairness

- Well calibrated prediction: Predictions on random subgroups match the predictions over the whole population.
 - Example: Medical testing and diagnosis - Is decision-making applied uniformly across different groups of patients?
- Balance for the positive class i.e. FPR: the average score received by people constituting positive instances should be the same in each subgroup.
 - Example: Different advertising and commercial content on social media, maybe shown based on gender or racial groups.
 - Females maybe shown lower-paying job adverts compared to equally qualified males.

Notions of Fairness

- Well calibrated prediction: Predictions on random subgroups match the predictions over the whole population.
 - Example: Medical testing and diagnosis - Is decision-making applied uniformly across different groups of patients?
- Balance for the positive class i.e. FPR: the average score received by people constituting positive instances should be the same in each subgroup.
 - Example: Different advertising and commercial content on social media, maybe shown based on gender or racial groups.
 - Females maybe shown lower-paying job adverts compared to equally qualified males.

Notions of Fairness

- Balance for the negative class i.e. FNR: the average score received by people constituting negative instances should be the same in each subgroup.
 - Example: Criminal Justice System - Decisions about bail, sentencing, or parole, are increasingly based on models which predict the probability of recidivism, based on past history e.g. COMPAS Risk Tool.
- Individual Fairness: Models are Lipschitz continuous with respect to a certain notion of individual similarity.
 - Example: Credit Scoring - Are two "similar" people scored similarly? That is, does slightly changing attributes of a person, change their score dramatically? Example: Strategic manipulation of FICO scores.

Notions of Fairness

- Balance for the negative class i.e. FNR: the average score received by people constituting negative instances should be the same in each subgroup.
 - Example: Criminal Justice System - Decisions about bail, sentencing, or parole, are increasingly based on models which predict the probability of recidivism, based on past history e.g. COMPAS Risk Tool.
- Individual Fairness: Models are Lipschitz continuous with respect to a certain notion of individual similarity.
 - Example: Credit Scoring - Are two "similar" people scored similarly? That is, does slightly changing attributes of a person, change their score dramatically? Example: Strategic manipulation of FICO scores.

Notions of Fairness

- Balance for the negative class i.e. FNR: the average score received by people constituting negative instances should be the same in each subgroup.
 - Example: Criminal Justice System - Decisions about bail, sentencing, or parole, are increasingly based on models which predict the probability of recidivism, based on past history e.g. COMPAS Risk Tool.
- Individual Fairness: Models are Lipschitz continuous with respect to a certain notion of individual similarity.
 - Example: Credit Scoring - Are two “similar” people scored similarly? That is, does slightly changing attributes of a person, change their score dramatically? Example: Strategic manipulation of FICO scores.

Notions of Fairness

- Balance for the negative class i.e. FNR: the average score received by people constituting negative instances should be the same in each subgroup.
 - Example: Criminal Justice System - Decisions about bail, sentencing, or parole, are increasingly based on models which predict the probability of recidivism, based on past history e.g. COMPAS Risk Tool.
- Individual Fairness: Models are Lipschitz continuous with respect to a certain notion of individual similarity.
 - Example: Credit Scoring - Are two “similar” people scored similarly? That is, does slightly changing attributes of a person, change their score dramatically? Example: Strategic manipulation of FICO scores.

Outline

- 1 Automated Decision Making
 - Introduction
 - Bias in Machine Learning
- 2 Measuring Privacy and Fairness
 - Differential Privacy
 - Notions of Fairness
 - Impossibility Results
- 3 Learning Differentially Private and Fair Models
 - Fair Empirical Risk Minimization
 - UCI Credit-Card Default Dataset

Impossibility: Statistical Parity

- Notions of group fairness intuitively ask that our model should have the same effectiveness regardless of group membership.
- One might therefore hope that it would be feasible to achieve all of them simultaneously i.e. achieve statistical parity.
- However, unfortunately it is impossible to achieve statistical parity, even approximately, unless [Kleinberg et al 2017]
 - the classification algorithm is perfect and
 - all underlying subgroups are statistically equivalent.

Impossibility: Statistical Parity

- Notions of group fairness intuitively ask that our model should have the same effectiveness regardless of group membership.
- One might therefore hope that it would be feasible to achieve all of them simultaneously i.e. achieve statistical parity.
- However, unfortunately it is impossible to achieve statistical parity, even approximately, unless [Kleinberg et al 2017]
 - the classification algorithm is perfect and
 - all underlying subgroups are statistically equivalent.

Impossibility: Statistical Parity

- Notions of group fairness intuitively ask that our model should have the same effectiveness regardless of group membership.
- One might therefore hope that it would be feasible to achieve all of them simultaneously i.e. achieve statistical parity.
- However, unfortunately it is impossible to achieve statistical parity, even approximately, unless [Kleinberg et al 2017]
 - the classification algorithm is perfect and
 - all underlying subgroups are statistically equivalent

Impossibility: Statistical Parity

- Notions of group fairness intuitively ask that our model should have the same effectiveness regardless of group membership.
- One might therefore hope that it would be feasible to achieve all of them simultaneously i.e. achieve statistical parity.
- However, unfortunately it is impossible to achieve statistical parity, even approximately, unless [Kleinberg et al 2017]
 - the classification algorithm is perfect and
 - all underlying subgroups are statistically equivalent

Impossibility: Statistical Parity

- Notions of group fairness intuitively ask that our model should have the same effectiveness regardless of group membership.
- One might therefore hope that it would be feasible to achieve all of them simultaneously i.e. achieve statistical parity.
- However, unfortunately it is impossible to achieve statistical parity, even approximately, unless [Kleinberg et al 2017]
 - the classification algorithm is perfect and
 - all underlying subgroups are statistically equivalent

Non-Trivial Problem: Composition

- Most decision making processes, automated or otherwise, are based on multiple decisions composed together.
 - Example: Mixture of experts model, sequential filtering
- Unfortunately, compositions of individually fair decisions may not be fair [Zemel et al 2013].
 - Example: Ranked choice voting systems with more than 2 candidates can be unfair (Arrow's Impossibility Theorem)
- Therefore, it is important to look at the problem of fairness holistically while training the model.

Non-Trivial Problem: Composition

- Most decision making processes, automated or otherwise, are based on multiple decisions composed together.
 - Example: Mixture of experts model, sequential filtering
- Unfortunately, compositions of individually fair decisions may not be fair [Zemel et al 2013].
 - Example: Ranked choice voting systems with more than 2 candidates can be unfair (Arrow's Impossibility Theorem)
- Therefore, it is important to look at the problem of fairness holistically while training the model.

Non-Trivial Problem: Composition

- Most decision making processes, automated or otherwise, are based on multiple decisions composed together.
 - Example: Mixture of experts model, sequential filtering
- Unfortunately, compositions of individually fair decisions may not be fair [Zemel et al 2013].
 - Example: Ranked choice voting systems with more than 2 candidates can be unfair (Arrow's Impossibility Theorem)
- Therefore, it is important to look at the problem of fairness holistically while training the model.

Non-Trivial Problem: Composition

- Most decision making processes, automated or otherwise, are based on multiple decisions composed together.
 - Example: Mixture of experts model, sequential filtering
- Unfortunately, compositions of individually fair decisions may not be fair [Zemel et al 2013].
 - Example: Ranked choice voting systems with more than 2 candidates can be unfair (Arrow's Impossibility Theorem)
- Therefore, it is important to look at the problem of fairness holistically while training the model.

Non-Trivial Problem: Composition

- Most decision making processes, automated or otherwise, are based on multiple decisions composed together.
 - Example: Mixture of experts model, sequential filtering
- Unfortunately, compositions of individually fair decisions may not be fair [Zemel et al 2013].
 - Example: Ranked choice voting systems with more than 2 candidates can be unfair (Arrow's Impossibility Theorem)
- Therefore, it is important to look at the problem of fairness holistically while training the model.

Differential Privacy and Fairness

- Since privacy is achieved by limiting the influence of the whole data point, there is a trade-off with controlling fairness, which requires more fine-grained analysis.
- It further implies that, unlike privacy, ensuring fairness does not imply increasing generalizability of the model.
- Therefore, the models of interest in practice ought to be both differentially private and fair.

Differential Privacy and Fairness

- Since privacy is achieved by limiting the influence of the whole data point, there is a trade-off with controlling fairness, which requires more fine-grained analysis.
- It further implies that, unlike privacy, ensuring fairness does not imply increasing generalizability of the model.
- Therefore, the models of interest in practice ought to be both differentially private and fair.

Differential Privacy and Fairness

- Since privacy is achieved by limiting the influence of the whole data point, there is a trade-off with controlling fairness, which requires more fine-grained analysis.
- It further implies that, unlike privacy, ensuring fairness does not imply increasing generalizability of the model.
- Therefore, the models of interest in practice ought to be both differentially private and fair.

Outline

- 1 Automated Decision Making
 - Introduction
 - Bias in Machine Learning
- 2 Measuring Privacy and Fairness
 - Differential Privacy
 - Notions of Fairness
 - Impossibility Results
- 3 Learning Differentially Private and Fair Models
 - Fair Empirical Risk Minimization
 - UCI Credit-Card Default Dataset

Empirical Risk Minimization

- The problem of learning any model can be reduced to the problem of empirical risk minimization

$$f^* = \arg \min_{f \in \mathcal{H}} \frac{1}{n} \sum_i [L(f(x_i), y_i)]$$

where L is the loss function, \mathcal{H} is the model space and $\{(x_i, y_i)\}_{i=1}^n$ is the dataset.

- In Fair Empirical Risk Minimization [Donini et al 2018] the idea is to regularize the optimization problem using Karush-Kuhn-Tucker theory

$$f^* = \arg \min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_i [L(f(x_i), y_i)] + \lambda \mathcal{F}(f) \right\}$$

where $\mathcal{F}(f)$ is the fairness constraint on the learned model.

Empirical Risk Minimization

- The problem of learning any model can be reduced to the problem of empirical risk minimization

$$f^* = \arg \min_{f \in \mathcal{H}} \frac{1}{n} \sum_i [L(f(x_i), y_i)]$$

where L is the loss function, \mathcal{H} is the model space and $\{(x_i, y_i)\}_{i=1}^n$ is the dataset.

- In Fair Empirical Risk Minimization [Donini et al 2018] the idea is to regularize the optimization problem using Karush-Kuhn-Tucker theory

$$f^* = \arg \min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_i [L(f(x_i), y_i)] + \lambda \mathcal{F}(f) \right\}$$

where $\mathcal{F}(f)$ is the fairness constraint on the learned model.

Differentially Private SGD

Differentially-Private Stochastic Gradient Descent

- Compute per sample gradient $\nabla_{\theta} L(z, \hat{\theta})$
- Clip the gradients $\nabla_{\theta} L(z, \hat{\theta})$ to a fixed maximum norm so that the corresponding influence can be effectively bounded

$$\|\mathcal{I}(z)\| = \|\hat{H}_{\hat{\theta}}^{-1} \nabla_{\theta} L(z, \hat{\theta})\| \leq \epsilon$$

- Aggregate them back into a single parameter gradient
- Add Gaussian noise to the clipped gradients and perform standard SGD iteration.

Differentially Private SGD

Differentially-Private Stochastic Gradient Descent

- Compute per sample gradient $\nabla_{\theta} L(z, \hat{\theta})$
- Clip the gradients $\nabla_{\theta} L(z, \hat{\theta})$ to a fixed maximum norm so that the corresponding influence can be effectively bounded

$$\| \mathcal{I}(z) \| = \| H_{\hat{\theta}}^{-1} \nabla_{\theta} L(z, \hat{\theta}) \| \leq \epsilon$$

- Aggregate them back into a single parameter gradient
- Add Gaussian noise to the clipped gradients and perform standard SGD iteration.

Differentially Private SGD

Differentially-Private Stochastic Gradient Descent

- Compute per sample gradient $\nabla_{\theta} L(z, \hat{\theta})$
- Clip the gradients $\nabla_{\theta} L(z, \hat{\theta})$ to a fixed maximum norm so that the corresponding influence can be effectively bounded

$$\| \mathcal{I}(z) \| = \| H_{\hat{\theta}}^{-1} \nabla_{\theta} L(z, \hat{\theta}) \| \leq \epsilon$$

- Aggregate them back into a single parameter gradient
- Add Gaussian noise to the clipped gradients and perform standard SGD iteration.

Differentially Private SGD

Differentially-Private Stochastic Gradient Descent

- Compute per sample gradient $\nabla_{\theta} L(z, \hat{\theta})$
- Clip the gradients $\nabla_{\theta} L(z, \hat{\theta})$ to a fixed maximum norm so that the corresponding influence can be effectively bounded

$$\| \mathcal{I}(z) \| = \| H_{\hat{\theta}}^{-1} \nabla_{\theta} L(z, \hat{\theta}) \| \leq \epsilon$$

- Aggregate them back into a single parameter gradient
- Add Gaussian noise to the clipped gradients and perform standard SGD iteration.

Outline

- 1 Automated Decision Making
 - Introduction
 - Bias in Machine Learning
- 2 Measuring Privacy and Fairness
 - Differential Privacy
 - Notions of Fairness
 - Impossibility Results
- 3 Learning Differentially Private and Fair Models
 - Fair Empirical Risk Minimization
 - UCI Credit-Card Default Dataset

UCI Credit-Card Default Dataset

Google Colaboratory Notebook:

<https://colab.research.google.com/drive/1yrL7FqP6I14w-Kaly6T3nHvF40Vipk6I?usp=sharing>

Summary

- Ensuring Fairness **is hard** even if measuring it is relatively easy.
- **Complete Fairness** is impossible among all groups, there are tradeoffs.
- Ensuring Fairness **is expensive**, and there has to be a political/managerial discussion, to what is acceptable.
- Ensuring Differential privacy is **relatively easy**, and improves **generalization** performance of the learned model.

Summary

- Ensuring Fairness is **hard** even if measuring it is relatively easy.
- **Complete Fairness** is impossible among all groups, there are tradeoffs.
- Ensuring Fairness is **expensive**, and there has to be a political/managerial discussion, to what is acceptable.
- Ensuring Differential privacy is **relatively easy**, and improves **generalization** performance of the learned model.

Summary

- Ensuring Fairness **is hard** even if measuring it is relatively easy.
- **Complete Fairness** is impossible among all groups, there are tradeoffs.
- Ensuring Fairness **is expensive**, and there has to be a political/managerial discussion, to what is acceptable.
- Ensuring Differential privacy is **relatively easy**, and improves **generalization** performance of the learned model.





Summary

- Ensuring Fairness **is hard** even if measuring it is relatively easy.
- **Complete Fairness** is impossible among all groups, there are tradeoffs.
- Ensuring Fairness **is expensive**, and there has to be a political/managerial discussion, to what is acceptable.
- Ensuring Differential privacy is **relatively easy**, and improves **generalization** performance of the learned model.





Thank You!

- Thank you for your attention! Any questions?





References

-  Donini et al. risk minimization under fairness constraints. Advances in Neural Information Processing Systems, 2018.
-  Lowy et al. FERMI: Fair Empirical risk minimization via Exponential Renyi Mutual Information://arxiv.org/pdf/2102.12586.pdf, 2021
-  Zemel et al. Learning Fair Representations of the 30th International Conference on Machine Learning, 2013
-  Kleinberg et al. Inherent Trade-Offs in the Fair Determination of Risk Scores of Innovations in Theoretical Computer Science (ITCS), 2017





References

-  Donini et al. risk minimization under fairness constraints. Advances in Neural Information Processing Systems, 2018.
-  Lowy et al. FERMI: Fair Empirical risk minimization via Exponential Renyi Mutual Information://arxiv.org/pdf/2102.12586.pdf, 2021
-  Zemel et al. Learning Fair Representations of the 30th International Conference on Machine Learning, 2013
-  Kleinberg et al. Inherent Trade-Offs in the Fair Determination of Risk Scores of Innovations in Theoretical Computer Science (ITCS), 2017

References

-  [Donini et al. risk minimization under fairness constraints.](#)
Advances in Neural Information Processing Systems, 2018.
-  [Lowy et al.](#)
FERMI: Fair Empirical risk minimization via Exponential Renyi Mutual Information://arxiv.org/pdf/2102.12586.pdf, 2021
-  [Zemel et al.](#)
Learning Fair Representations of the 30th International Conference on Machine Learning, 2013
-  [Kleinberg et al.](#)
Inherent Trade-Offs in the Fair Determination of Risk Scores of Innovations in Theoretical Computer Science (ITCS), 2017

References

-  [Donini et al. risk minimization under fairness constraints.](#)
Advances in Neural Information Processing Systems, 2018.
-  [Lowy et al.](#)
FERMI: Fair Empirical risk minimization via Exponential Renyi Mutual Information://arxiv.org/pdf/2102.12586.pdf, 2021
-  [Zemel et al.](#)
Learning Fair Representations of the 30th International Conference on Machine Learning, 2013
-  [Kleinberg et al.](#)
Inherent Trade-Offs in the Fair Determination of Risk Scores of Innovations in Theoretical Computer Science (ITCS), 2017