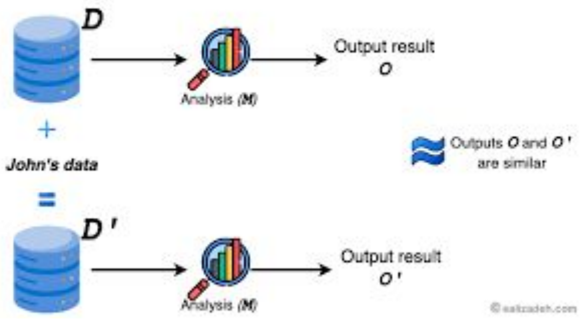
Privacy and Fairness

Group 3

Differential Privacy and Fairness in Decisions and Learning Tasks

Motivation & Background

- **AI/ML** increasingly used in decisions: legal, hiring, healthcare, policy
- DP protects sensitive data – adopted by Census, Google, Apple
- But: DP noise may harm underrepresented groups
- Goal: Understand when privacy and fairness align or conflict



Foundations – DP & Fairness

- **DP:**
  - A randomized mechanism M: X -> Y is is (ε, δ)-DP if:

$$P[M(x) = y] \leq e^{\epsilon} \cdot P[M(x') = y] + \delta$$

- **Fairness Concepts:**
  - Individual Fairness: Similar individuals -> similar outcomes
  - Group Fairness: Equal outcomes across groups

Decision Tasks vs. Learning Tasks

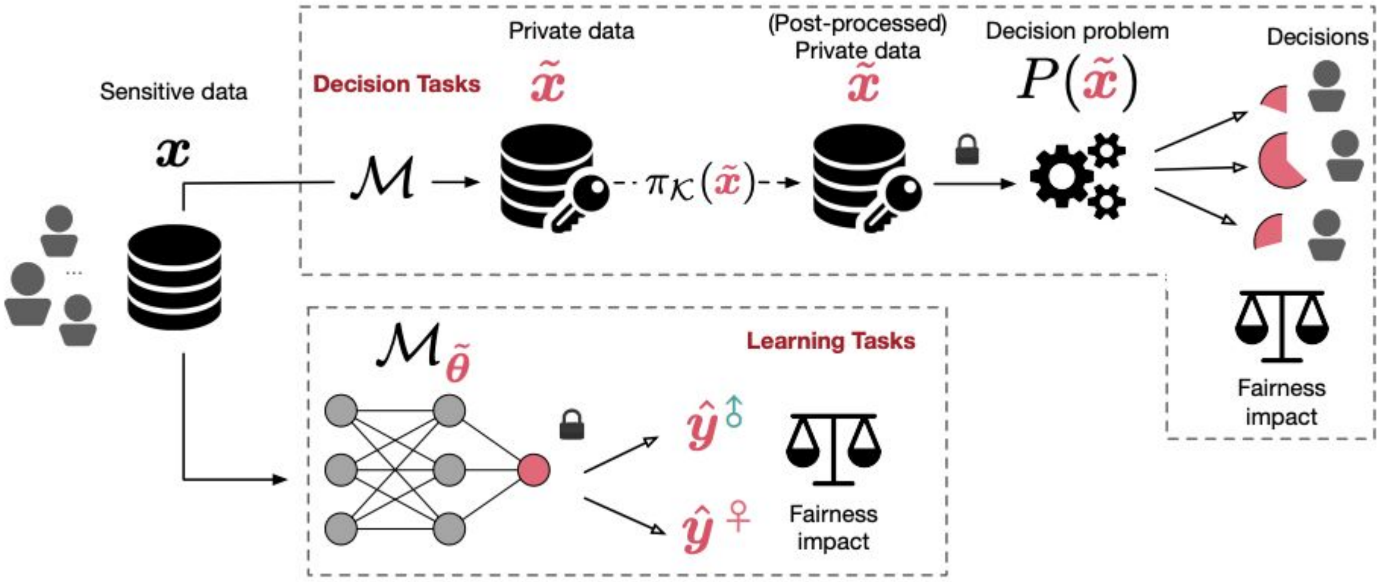
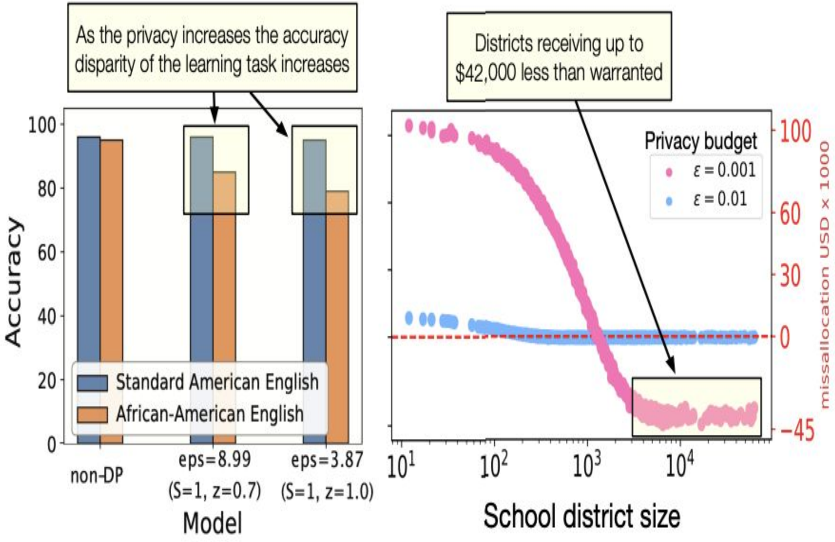


Figure 2: Setting analyzed in this survey.

How DP Affects Fairness

- **In Decision Tasks:**
  - DP adds noise to data (e.g., census counts) -> can distort decisions
  - **Bias arises** when:
    - The decision function is non-linear
    - Post processing shifts values unevenly
- **In Learning Tasks:**
  - DP methods like **Dp-SGD** add noise during training
  - Disparate impact occurs due to:
    - **Gradient clipping** penalizing high-norm groups
    - **Noise** disproportionately affecting underrepresented samples



Why Does Privacy Hurt Fairness?

- **In Decision Tasks:**
  - Bias from non-linearity (Taylor expansion):
    - Post-processing introduces asymmetric errors
- **Learning Tasks:**
  - Minority groups often have:
    - Higher gradient norms
    - Data far from decision boundary -> affected by noise + clipping

$$B_i^P = \frac{1}{2} H_{P_i}(x) \cdot \mathrm{Var}[\eta]$$

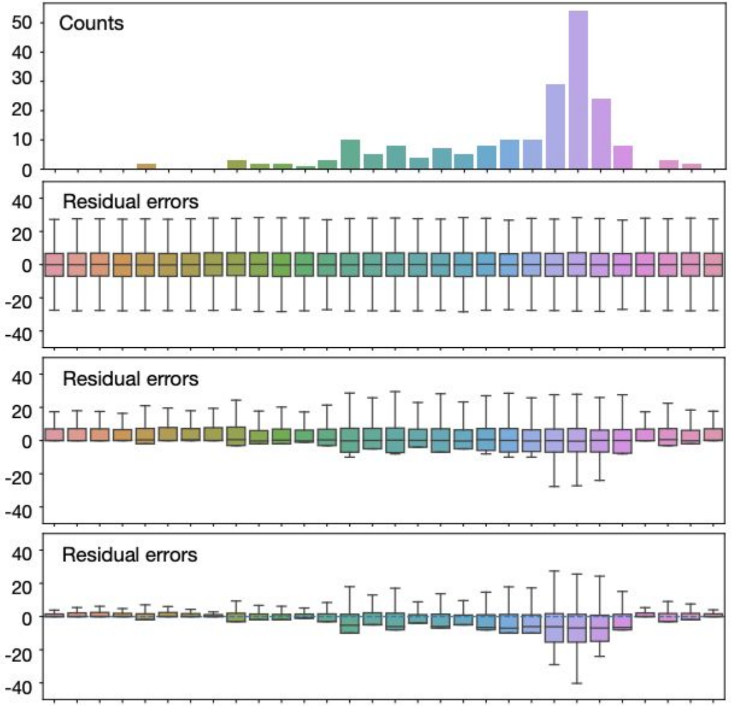


Figure 3: Bias and variance in DP post-processing.

Mitigating Fairness Under DP

- **In Decision Tasks:**
  - Linear proxy problems
  - Fair projections to reduce group disparity

- **In Learning Tasks:**
  - Group aware gradient clipping
  - Excessive risk constraints
  - Early stopping

Challenges and Takeaway

- **Unresolved Issues:**
  - No unified theory linking ε, accuracy, and fairness
  - Hyperparameters affect fairness under DP
  - Robustness, privacy, and fairness are entangled
  - Limited tools for DP + fairness auditing
- **Takeaway:**
  - Achieving fairness under DP is hard – but not impossible

The Impact of Differential Privacy on Model Accuracy

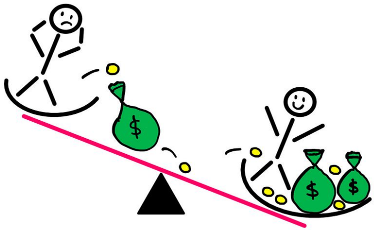
Differential Privacy (DP)

- **Differential Privacy**
  - Bounds the influence of any single input on the output of a computation

- **Differentially Private Stochastic Gradient Descent (DP-SGD)**
  - A training algorithm that achieves DP by computing gradients on mini-batches, clipping individual gradients, and adding noise

- **ε**
  - Parameter that controls privacy loss - the tradeoff between privacy and model accuracy

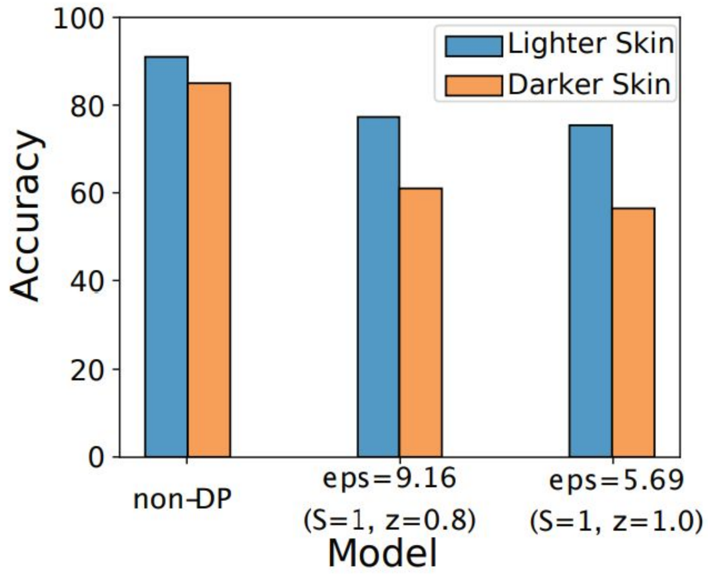
Limitations of Differential Privacy



- Reductions in training accuracy incurred by DP disproportionately impacts underrepresented and complex subgroups
    - Smaller subgroups experience a greater reduction in accuracy compared to larger groups
- DP is biased towards popular elements of the distribution learned
- “The Poor Get Poorer” Effect
    - Classes with lower accuracy in the non-DP model experience the largest accuracy drops when DP is applied

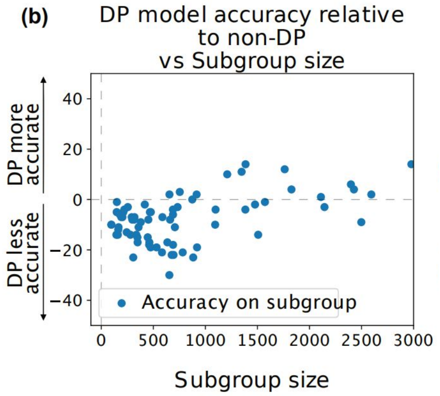
Gender Classification

- Model performs gender classification based on facial imagery
- 29,500 images of individuals with lighter skin color
- 500 images of individuals with darker skin color
- DP-SGD leads to greater accuracy degradation for darker-skinned faces compared to lighter-skinned ones

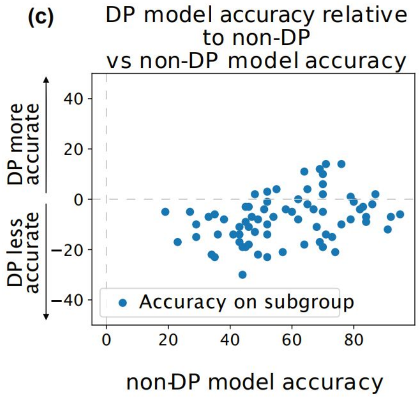


Age Classification

- Model estimates an individual's age based on their facial image
- Accuracy of model is measured across subgroups defined by the intersection of age, gender, and skin color attributes
- 60,000 images randomly sampled from Diversity in Faces (DiF) dataset
- Model evaluated on 72 intersections (subgroups)



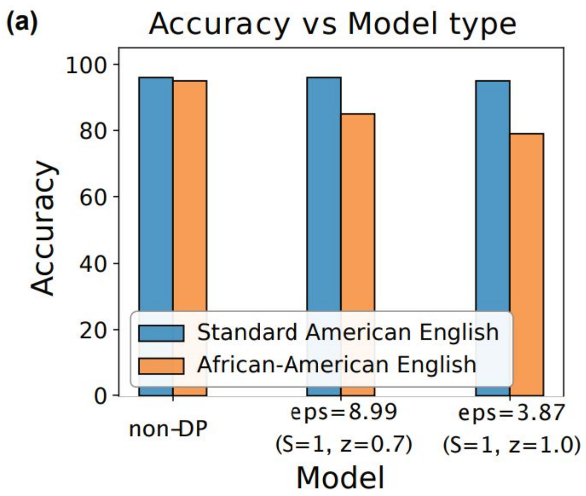
DP Model less accurate on smaller subgroups



"The Poor Get Poorer" Effect

Sentiment Analysis of Tweets

- Model classifies Twitter posts as positive or negative
- Trained on 60,000 STA tweets and 1,000 AAE tweets
- The accuracy of the DP model drops more than the non-DP model
    - Disproportionately degrades accuracy for users writing in African-American English



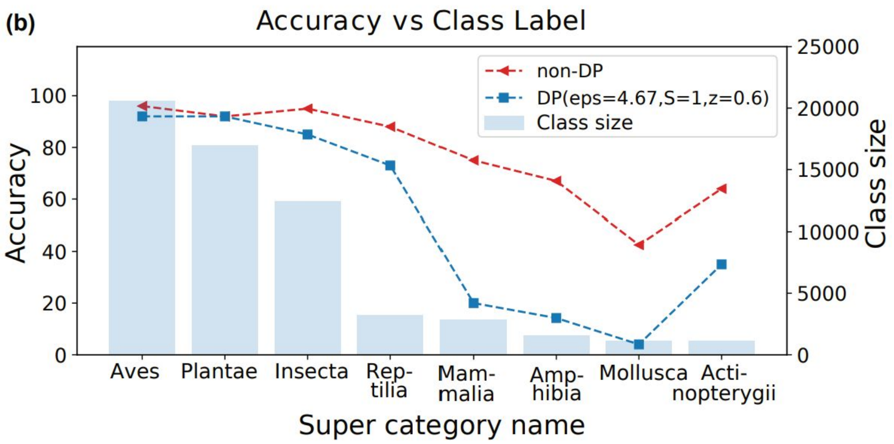
Species Classification

- Trained on 60,000 images from iNaturalist dataset, which contains hierarchically labeled images of plants and animals
  - Largest: Aves, 20,574 images
  - Smallest: Actinopterygii, 1,119 images

- Model classifies images into 8 classes

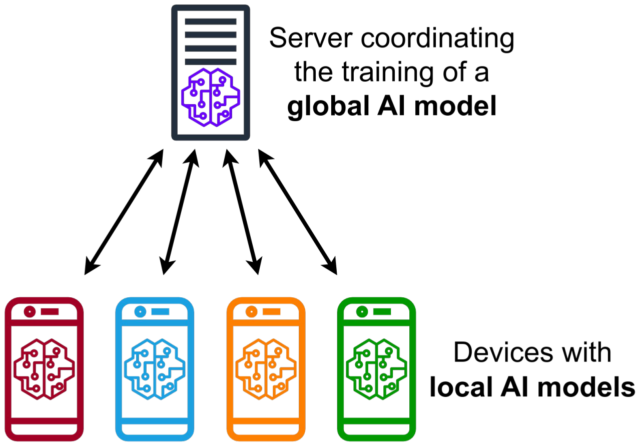
- Accuracy is lower for underrepresented/smaller classes

- Accuracy of DP model almost matches the accuracy of the non-DP model in well-represented classes



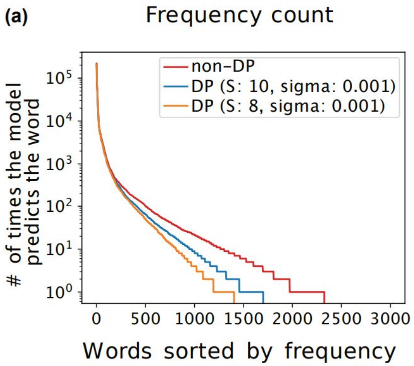
Federated Learning

- Participants jointly train a model
- In each round, a global server distributes the current global model to a subgroup
- Each participant in the subgroup trains the the global model on their private data, producing their own local model
- The global server aggregates the local models and uses them to update the global model
- The process repeats

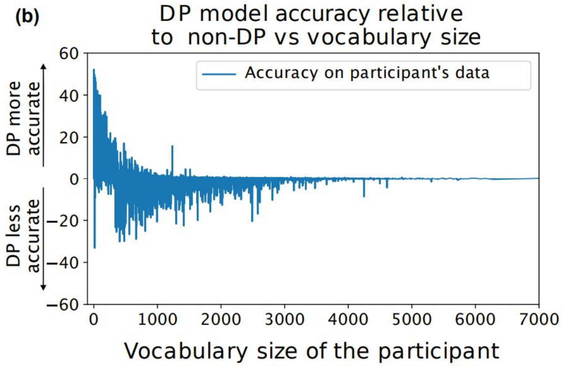


Federated Learning of Language Models

- Trained on public Reddit posts made in November 2017 by users who have made 150-500 posts
- Model is trained to predict the next word given a partial word sequence
- Vocabulary is restrict to 50K most frequently used words, and unpopular words, emojis, and special symbols are replaced with <unk>
- Accuracy is lower for users with larger vocabularies, and higher for those with smaller vocabularies
    - DP model predicts most popular words



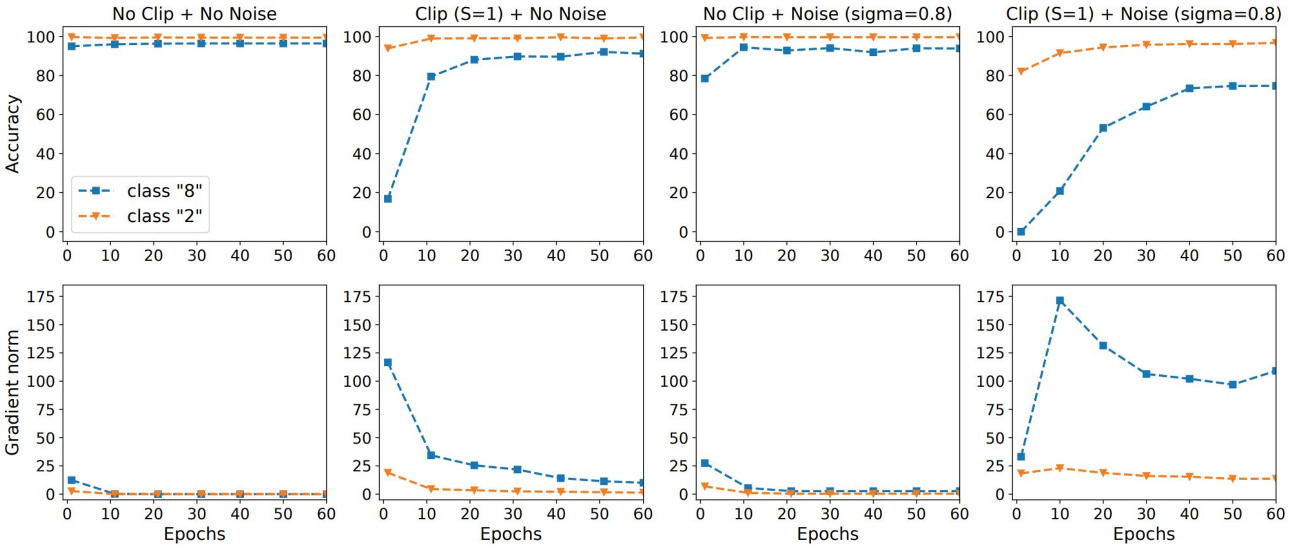
Non-DP model is more accurate than DP model



Accuracy decreases vocabulary size increases

Effect of Clipping and Noise on MNIST Training

- MNIST is a numbers classification dataset
- Higher accuracy with no clipping and no noise
    - Trade-off between accuracy and privacy



Discussion

- How can the trade-off between accuracy and privacy be mitigated in models with differential privacy, particularly in respect to its disparate impact on underrepresented/small subgroups?

- In what scenarios would having underrepresented/small subgroups benefit a model?

Differentially Empirical Risk Minimization Under the Fairness Lens

Excessive Risk & Fairness

- **Excessive Risk measures how much performance is lost due to privacy:**

$$R(\theta; D) = \mathbb{E}_{\mathcal{M}}[\mathcal{L}(\hat{\theta}; D)] - \mathcal{L}(\theta^*; D)$$

- **Fairness Definition: Excessive Risk Gap**
  - Fairness gap for group α:

$$\xi_a = |R_a(\theta) - R(\theta)|$$

  - If gap for a is large -> group a suffers more
  - Goal: minimize $\max_a \xi_a$ to achieve fairness

Two DP Strategies in ERM - Output Perturbation

- **Output Perturbation**
  - Train standard ERM model, then add noise to final model parameters
  - Advantage: easy to implement, post-processing

$$\hat{\theta} = \theta^* + \mathcal{N}(0, \sigma^2 I)$$

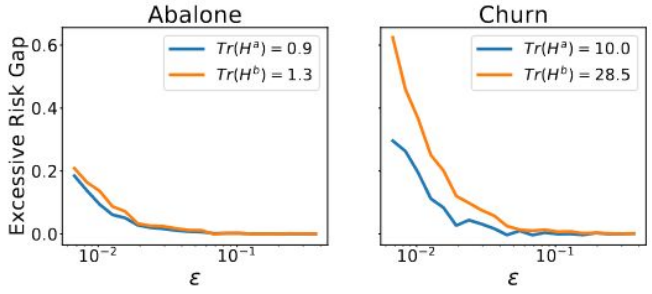


Figure 1: Correlation between excessive risk gap and Hessian Traces at varying of the privacy loss ϵ.

- **Where Does Unfairness Come From?**
  - Excessive risk gap caused by curvature difference:

$$\xi_a \approx \frac{1}{2}\Delta^2\sigma^2 \left|\mathrm{Tr}(H_\ell^a) - \mathrm{Tr}(H_\ell)\right|$$

  - Measures how sensitive the model is to parameter changes for that group
  - Groups with larger Hessian trace are more affected by the same noise

Two DP Strategies in ERM - DP-SGD

$$\mathbb{E}\left[\mathcal{L}(\theta_{t+1}; D_a)\right] \approx \underbrace{\mathcal{L}(\theta_t; D_a) - \eta \langle g_{D_a}, g_D \rangle + \frac{\eta^2}{2}\mathbb{E}\left[g_B^T \boldsymbol{H}_\ell^a g_B\right]}_{\textit{non-private term}} \tag{4}$$

$$+ \underbrace{\eta\left(\langle g_{D_a}, g_D \rangle - \langle g_{D_a}, \bar{g}_D \rangle\right) + \frac{\eta^2}{2}\left(\mathbb{E}\left[\bar{g}_B^T \boldsymbol{H}_\ell^a \bar{g}_B\right] - \mathbb{E}\left[g_B^T \boldsymbol{H}_\ell^a g_B\right]\right)}_{\textit{private term due to clipping}} \qquad (R_a^{\text{clip}})$$

$$+ \underbrace{\frac{\eta^2}{2}\mathrm{Tr}(\boldsymbol{H}_\ell^a) C^2 \sigma^2}_{\textit{private term due to noise}} \qquad (R_a^{\text{noise}})$$

- **DP-SGD**
  - DP-SGD adds noise to each gradient update
- **Where Does Unfairness Come From**

$$R_a = R_a^{\text{clip}} + R_a^{\text{noise}}$$

  - **Clipping Risk**:
    - Gradient vectors are clipped to a fixed norm bound C
    - Groups with larger gradient norms lose more directional info → These groups may learn less → performance degrades
  - **Noise Risk:**
    - After clipping, Gaussian noise is added. Groups with higher Hessian curvature are more sensitive to perturbation → Noise causes more error for these groups

Clipping Risk

Theorem 3. *Let $p_z = |D_z|/|D|$ be the fraction of training samples in group $z \in \mathcal{A}$. For groups $a, b \in \mathcal{A}$, $R_a^{\text{clip}} > R_b^{\text{clip}}$ whenever:*

$$\|g_{D_a}\|\left(p_a - \frac{p_a^2}{2}\right) \geq \frac{5}{2}C + \|g_{D_b}\|\left(1 + p_b + \frac{p_b^2}{2}\right). \qquad (5)$$

- Theorem 3 provides a sufficient condition for which a group may have larger excessive risk than another solely based on the clipping term analysis.
- It relates unfairness with the average (non-private) gradient norms between groups and the clipping value C.
- The gradient norm of a group is strongly correlated with Input norm // X // . → groups with larger input features are more likely to be affected

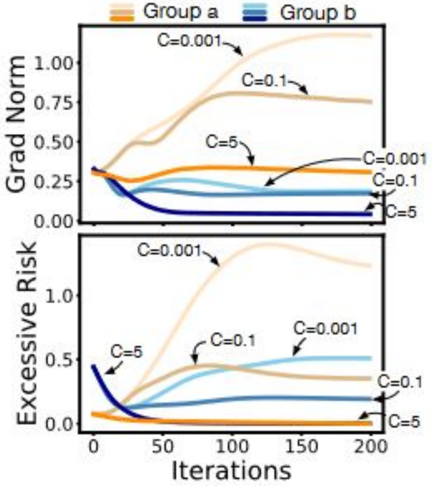


Figure 3: Impact of gradient clipping on gradient norms for different clipping bounds. Bank dataset.

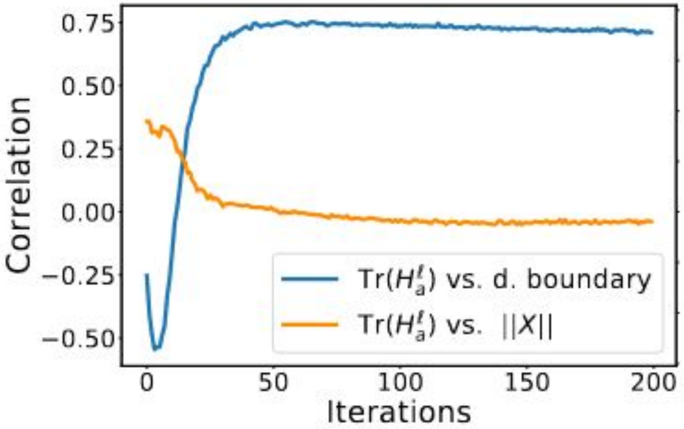
Noise risk

- After clipping, Gaussian noise is added to ensure differential privacy. Noise affects all groups but its impact is not equal

- Theorem 4: If one group has a higher loss curvature (i.e., larger Hessian trace), then it will suffer more from the same amount of noise:

$$R_a^{\text{noise}} > R_b^{\text{noise}} \quad \text{if} \quad \text{Tr}(H_\ell^a) > \text{Tr}(H_\ell^b)$$

- What Drives High Curvature?
  - **Proximity to decision boundary**: Samples near the boundary → prediction uncertainty → higher curvature
  - **Input norm** // **X** // : Larger input vectors → higher second-order sensitivity



Mitigation Strategy

- Objective is to minimize both:
  - Differences in gradient norms (clipping risk)
  - Differences in curvature/Hessian (noise risk)

- Modified optimization objective: $$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; D) + \sum_{a \in \mathcal{A}} \left(\gamma_1 \left| \langle \boldsymbol{g}_{D_a} - \boldsymbol{g}_D, \boldsymbol{g}_D - \bar{\boldsymbol{g}}_D \rangle \right| + \gamma_2 \left| \mathrm{Tr}(\boldsymbol{H}_\ell^a) - \mathrm{Tr}(\boldsymbol{H}_\ell) \right| \right)$$

  Where:

  - $\mathcal{L}$: standard ERM loss
  - g_{D_a}: group-level gradients
  - H_ℓ^a: group-level Hessian
  - γ_1: Controls gradient alignment (for clipping fairness)
  - γ_2: Controls curvature similarity (for noise fairness)

- Hessians are expensive to compute, so the paper replaces them with: $$\mathrm{Tr}(H_\ell^a) \approx \mathbb{E}_{X \sim D_a} \left[1 - \sum_k f_{\theta,k}^2(X) \right]$$

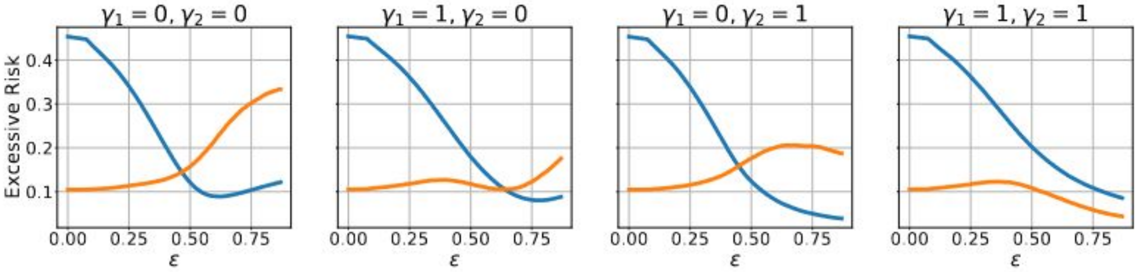
Mitigation results

Baseline: $\gamma_1 = \gamma_2 = 0$

Only $\gamma_1 > 0$: targets clipping fairness

Only $\gamma_2 > 0$: targets noise fairness

Both $\gamma_1, \gamma_2 > 0$: full mitigation



The paper's core contributions:

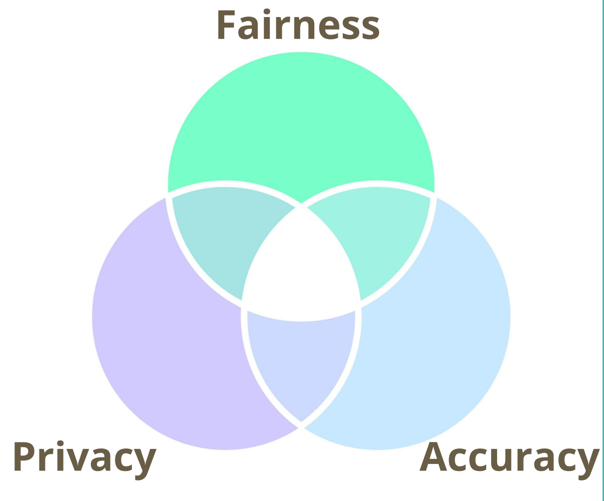
- Introduces **Excessive Risk Gap** as a fairness-aware metric under DP
- Shows how **gradient clipping** and **noise addition** can cause disparate impacts
- Identifies **input norm and boundary proximity** as hidden drivers of unfairness
- Proposes a effective **mitigation method** that improves fairness **without sacrificing utility**

Discussion

- If a system treats all data the same (like adding equal noise), but harms some groups more than others, is that acceptable?

- Would making group membership (e.g., gender, race) explicit during training help or harm fairness?

Are Fairness and Privacy Compatible?



The Data Universe

- Let χ be a data universe consisting of elements of the form $z = (x,a,y)$ where x are the element's features, a is a protected (binary) binary attribute, and y is a binary label

Ex: Loan Applications

- x: applicant's income and credit score, a: whether the applicant is a racial minority, and y: whether the applicant intends to repay her loan

Database: A collection of these individuals ($Z = (z_1,z_2,\ldots,z_n)$) with entries drawn i.i.d from a distribution D over χ

- From this data set train a classifier, h

Prelim: Defining Differential Privacy

- **Differential privacy**: a strong guarantee for individuals whose data is used for training
- Model learns aggregate information without encoding information about specific individuals
- Neighboring databases:
  - Neighboring samples: two finite samples differing in at most one entry
  - **ζ**-closeness: two distributions where the statistical distance between them is at most **ζ**
- A randomized algorithm A is (ϵ,δ)-differentially private if for all pairs of neighboring databases D,D' and for all sets S ∈ Range(A) of outputs:

$$\Pr[\mathcal{A}(D) \in \mathcal{S}] \leq \exp(\epsilon) \Pr[\mathcal{A}(D') \in \mathcal{S}] + \delta.$$

- If δ = 0, A is ϵ-differentially private

Prelim: Exact Fairness - Equal Opportunity

- For the analysis of exact fairness, a database is considered as a distribution over the data universe
- **Equal Opportunity**: equality of group-conditional true positive classification rates for different values of the protected attribute (a = 0 and a = 1) given the positive label (Y=1)
- Used as the notion of exact fairness

$$\gamma_{ya}(h) := \Pr[h = 1 | Y = y, A = a]$$

- Fairness definition requires equality of group conditional true positive classification rates and assumes that $P_{ya} > 0$ for $a,y \in \{0,1\}$

The Impossibility of Exact Fairness with DP

- A hypothesis fair under one distribution may be unfair under a neighboring one
- DP prevents output change based on small distribution changes
- For two neighboring distributions D and D’
  - Any hypothesis, h, fair on D will not be fair on D’
  - DP constraint implies we cannot change the output significantly between D and D’
- Thus, no algorithm can achieve DP and exact fairness simultaneously with better than trivial accuracy

Approximate Fairness

Why?

- Achieving exact fairness is impossible when learning from a finite sample
- Exact fairness is incompatible with differential privacy

Use $\boldsymbol{\alpha}$-discrimination:

- More robust to sampling noise and compatible with DP
- A binary predictor, h, is $\boldsymbol{\alpha}$-discriminatory if the absolute difference between group-conditional true positive rates on the sample Z is no more than $\boldsymbol{\alpha}$

Approximate Fairness Definitions

- Define subgroup conditional true positive classification rates

$$\gamma_{ya}(h) := \Pr[h = 1 | Y = y, A = a]$$

- A classifier is α-discriminatory if:

$$\max_{y \in \{0,1\}} |\gamma^{Z}_{y0}(h) - \gamma^{Z}_{y1}(h)| \leq \alpha.$$

- $\alpha = 0$: exact fairness
- $\alpha > 0$: approximate fairness

Achieving Approximate Fairness with DP

- Goal: Learn a classifier h such that with high probability:
  - h has low error
  - h is $\boldsymbol{\alpha}$-discriminatory
  - The algorithm is (ϵ,δ)-DP
- Use concepts from:
  - Agnostic PAC learning
  - Differential Privacy (Laplace + Exponential Mechanisms)
- Laplace: Privately estimate subgroup sizes (the number of positively labeled individuals with A=1)
- Helps ensure no single datapoint affects the subgroup size

Agnostic PAC Learning

- Probably Approximately Correct learning without assuming that a perfect hypothesis (h) exists in the class $\mathcal{H}$
- Find a hypothesis h $\in \mathcal{H}$ such that:

$$\Pr[err(h) \leq OPT + \alpha] \geq 1 - \beta,$$

Why?

- Framework handles imperfect data by minimizing error as best as possible
- Labels might be noisy and not match any hypothesis in $\mathcal{H}$

Learning Algorithm: Exponential Mechanism

- Used to select a fair and accurate classifier from $\mathcal{H}$
- Each hypothesis receives a utility score: $u(Z,h) = \text{error}_Z(h) + \boldsymbol{\Gamma}_Z(h)$

Main ideas:

- Adds enough noise to maintain DP
- Ensures that a hypothesis with small loss is sampled with high probability
- Encourages low error and fairness while preserving privacy

Algorithm 1 Approximately Fair Private Learner $\mathcal{A}(\mathcal{H}, Z, n, \epsilon)$

Input: hypothesis class $\mathcal{H}$, sample Z of size n, privacy parameter ϵ
Set $u(Z, h) = \Gamma^Z(h) + err^Z(h)$ and $\delta = \exp(-\sqrt{n})$
Sample $Y \sim \mathrm{Lap}(1/\epsilon)$
Set $M = \min_a |Z_{1a}| + Y - \ln(1/\delta)(1/\epsilon)$
Set $\Delta = \frac{2}{M-1} + \frac{1}{n}$
Sample hypothesis $h \in \mathcal{H}$ with probability proportional to

$$\exp\left(-\frac{\epsilon \cdot u(Z, h)}{2\Delta}\right)$$

Output: sampled hypothesis h

Putting it All Together (Algorithm 1)

Efficient Algorithm for Approximately Fair Classification

Problem: Exponential Mechanism requires evaluating all hypotheses in $\mathcal{H}$ (all linear classifiers)

Private-FairNR Algorithm (Private Fair No-Regret)

- Use no-regret dynamics in a game between:
  - Learner (cast as two-player zero-sum game): minimizes error + fairness loss
  - Auditor: identifies fairness violations
- Relies on oracles for cost-sensitive classification
- Privacy is preserved using Prave Follow-The-Perturbed-Leader
  - Private, online learning algorithm used in each round
  - Adds noise using Laplace to ensure privacy in repeated interactions

Discussion

- Can fairness and privacy ever be fully aligned in practice, or must we always accept trade-offs?
    - Are there real-world contexts where you'd prioritize one over the other?
    - What would be the risks of prioritizing privacy over fairness (or vice versa)?

- What kinds of real-world systems would benefit from using a fair & private learner like the one proposed in this paper?

- How should practitioners balance fairness and privacy in deployment scenarios?