Balancing Privacy, Fairness, and Accuracy: A Comparative Study of Standard and Selective DP-SGD on the UCI Adult Income Dataset.

Enrico Madani[1], Han Zhang[1] and Aston Tandiono[1]

[1]University of Chicago

Abstract

Machine learning and AI are often trained on personal data that may reveal private information. Differentially-Private Stochastic Gradient Descent (DP-SGD) is one method to prevent this by clipping gradient norms and adding noise during training. In this project, we wanted to see how DP methods impact utility and fairness in the UCI adult dataset. We initially focused on comparing the baseline model (the one without Differential Privacy) and the DP-SGD model. However, standard DP-SGD applies privacy equally to all entries in the training dataset. To address that, we also explored a method called Selective Differentially-Private Stochastic Gradient Descent (S-DP-SGD). S-DP-SGD adds privacy protection selectively during training based on a policy that targets sensitive attributes [Shi et al., 2021]. This project evaluates if S-DP-SGD can achieve privacy protection comparable to DP-SGD, but with better accuracy and fairness.

Introduction

Motivation

Machine learning and AI in healthcare, government, or other domains are often trained on personal data. The output or process of that training may reveal private information. Differential Privacy (DP) is one method that provides a formal mathematical framework for limiting such information leakage.

However, the application of DP in the training, for example, in the process of stochastic gradient descent (SGD), requires per-sample gradient clipping and the addition of noise, which can worsen model performance. Recent studies have also shown that applying DP can also unintentionally increase algorithmic fairness [1].

One criticism for the implementation of DP with SGD is that it applies the same protection on all features when only a subset of variables are truly sensitive [2]. This leads to the exploration of selective differential privacy (S-DP) that just protects sensitive features more directly and avoids unnecessary noise to the non-sensitive features.

Fairness in ML is very important. Models used without attention to group-level impacts may perpetuate existing biases. Men and certain racial groups are often more represented in higher-income categories in datasets like the Adult Census Income dataset. Models trained without careful design risk perpetuating these disparities. Thus, balancing privacy, utility, and fairness remains an important open challenge in machine learning.

Research Question

Although DP-SGD has been well explored by many machine learning researchers [3, 4, 5], selective differential privacy remains relatively unexplored. The study by Shi et al. [2], which became our main inspiration, showed that Selective DP-SGD can achieve better privacy-utility trade-offs compared to standard DP-SGD in language modeling tasks by applying noise only to sensitive attributes. However, their analysis only focused on utility metrics such as perplexity and accuracy, and did not study how selective privacy protection affected fairness.

Our work extends the past studies by investigating whether the selective privacy mechanism can simultaneously improve both utility and fairness. This is particularly important given recent findings that standard DP-SGD can exacerbate algorithmic bias [1].

More specifically, our research question is: *can S-DP-SGD, which selectively applies differential privacy noise to gradients associated with sensitive features (e.g., race and sex), preserve model utility and fairness more effectively than DP-SGD?*

We compare three training strategies (all with the same logistic regression classifier) on the Adult Census Income dataset:

- No differential privacy (baseline)
- Standard DP-SGD, which applies gradient clipping and noise across the entire parameter space
- Selective DP-SGD, which applies noise only to gradients associated with explicitly sensitive features

Background

Differential Privacy

Basic Definition Differential Privacy (DP) [6]: This provides a mathematical foundation for calculating the privacy guarantees. The deletion of data for any individual will not significantly impact the algorithm's results. This prevents the leakage of information regarding any individual.

Definition 0.1 (Differential Privacy)**.** *Given a domain $\mathcal{D}$, any two datasets $D, D' \subseteq \mathcal{D}$ that differ in exactly one record are called neighboring datasets. A randomized algorithm $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ is (ϵ, δ)-differentially private if for all neighboring datasets D and D' and all subsets $T \subseteq \mathcal{R}$,*

$$\Pr[\mathcal{M}(D) \in T] \leq e^{\epsilon} \Pr[\mathcal{M}(D') \in T] + \delta. \quad (1)$$

Privacy Budget Interpretation The first, ϵ, is the factor that controls privacy loss. The smaller the epsilon, the stronger the privacy guarantee, which becomes perfect if and only if epsilon equals 0. The other parameter, δ, is the probability of failure for the privacy guarantee. We often pick δ to be very small.

In essence, ϵ measures the amount of information that can be gathered regarding an individual from the result of the algorithm. The smaller the value of ϵ, the more similar the distribution of the result of the algorithm whether it includes information regarding an individual or not, meaning the stronger the privacy.

Differentially Private Stochastic Gradient Descent (DP-SGD)

Abadi et al. [3] introduced DP-SGD, which adapts stochastic gradient descent to satisfy differential privacy guarantees. The key modifications are:

1. *Per-sample gradient clipping*, where each individual gradient g_i is clipped to have bounded ℓ_2 norm, $\tilde{g}_i = g_i / \max(1, \|g_i\|_2 / C)$ for a clipping threshold C.
2. *Noise addition*, where calibrated Gaussian noise is added to the aggregated clipped gradients before the parameter, θ, update.

Formally, at each training step $t \in [T]$, DP-SGD performs:

$$\tilde{g}_i = g_i / \max\left(1, \frac{\|g_i\|_2}{C}\right) \quad (2)$$

$$\tilde{g}_t = \frac{1}{|B|}\left(\sum_{i \in B} \tilde{g}_i + \mathcal{N}(0, \sigma^2 C^2 I)\right) \quad (3)$$

$$\theta_{t+1} = \theta_t - \eta \tilde{g}_t \quad (4)$$

Output: θ_T, which is the optimized parameter.

Note: B is the batch, σ is the noise multiplier, C is the clipping threshold, and η is the learning rate.

Impact on Accuracy and Fairness The process of gradient clipping and adding noise impacts the accuracy of the model. The noise added for privacy protection hinders the optimization process.[4].

Recent work has also shown the fact that the standard use of DP on model parameters can have negative effects to the fairness of algorithms [1]. The noise used for privacy protection disproportionately impacts different demographic subgroups when smaller subgroups are represented in the training data. Since DP adds noise according to the size of the overall dataset, rather than the subgroup size, minority groups may face a more significant relative accuracy drop. Another study has also shown that DP-SGD can distort learned relationships in ways that amplify existing biases [7].

Selective Differential Privacy

Core Concept Selective Differential Privacy (S-DP) [2] helps address this downside of DP. In many real-world datasets, sensitive information is sparse. Not every attribute within a data record needs protection. Sometimes, classical DP may cause unnecessary utility loss in protecting a whole record uniformly.

S-DP introduces a *policy function* that specifies which attributes within a record are sensitive:

Definition 0.2 (Policy Function)**.** *A policy function $F : \tau \to \{0,1\}^{n_r}$ indicates which attributes of a record $r \in \tau$ are sensitive ($F(r)_i = 0$) or non-sensitive ($F(r)_i = 1$), where n_r is the number of attributes in r.*

Under S-DP, two datasets are said to be neighbors only if they are different in the sensitive attributes identified by the policy function:

Definition 0.3 (F-Neighbors)**.** *Given datasets D, D' and a policy function F, D' is an F-neighbor of D (denoted $D' \in \mathcal{N}_F(D)$) if and only if $\exists r \in D$ such that $F(r)$ contains at least one private attribute, $\exists r' \in D'$ such that $F(r)$ and $F(r')$ differ by at least one private attribute, and $D' = D \setminus \{r\} \cup \{r'\}$.*

Definition 0.4 (Selective Differential Privacy)**.** *Given a policy function F, a randomized algorithm $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ satisfies (F, ϵ, δ)-selective differential privacy if for all $D, D' \in \mathcal{N}_F(D)$ and all $T \subseteq \mathcal{R}$,*

$$\Pr[\mathcal{M}(D) \in T] \leq e^{\epsilon} \Pr[\mathcal{M}(D') \in T] + \delta. \quad (5)$$

Selective DP-SGD Mechanism Shi et al. [2] then proposed Selective-DP-SGD. This is a modification of DP-SGD where only gradient components of sensitive attributes are clipped and noised. Non-sensitive

components follow standard SGD. This reduces unnecessary protection that might hinder the model's performance.

For our work on tabular data with logistic regression, we apply noise only to the gradient components corresponding to sensitive features (e.g., race and sex), and leave gradients for non-sensitive features (e.g., education, occupation) unperturbed.

Ethical Implications The ethical implication is that we need to correctly identify which features are considered sensitive. In this project, we chose to protect the sex and race features.

Dataset

UCI Adult Income Dataset

We used the UCI Adult Census Income dataset, one of the most studied datasets in fairness and privacy research. The target variable is binary income (whether an individual makes over $50K). There are 48,842 rows and 14 attributes related to socioeconomic factors. The dataset has both numerical and categorical variables. Some examples of numerical variables include age, number of education, and working hours per week. Some examples of the categorical variables are working class, occupation, and native country.

The following two attributes, `race` and `sex`, are treated as sensitive features for fairness evaluation and selective DP-SGD.

Preprocessing Pipeline

All models (No DP, DP-SGD, S-DP-SGD) go through the same preprocessing steps:

Label Cleaning: The target variable (`income`) is binarized into:

- 0: $\leq$ \$50K
- 1: $>$ \$50K

Data Cleaning: Missing values in the `workclass`, `occupation`, and `native-country` columns were indicated by the string "?" in the original dataset. All missing values in these categorical columns were imputed by filling with the mode (most frequent value) of each column.

Feature Transformation: We use a `ColumnTransformer` consisting of:

- `StandardScaler` for numerical features
- `OneHotEncoder` (with `handle_unknown='ignore'`) for categorical features

This results in 105 encoded features, with expanded one-hot representations for sensitive attributes (e.g., `race_White`, `race_Black`, `sex_Male`, etc.).

Train–Test Split: An 80/20 train-test split with `random_state=42` is applied. The same split is used across all models.

Methodology

Models

1. Baseline Model (No Differential Privacy) The baseline is a `sklearn` logistic regression classifier trained on the preprocessed features. The baseline functions as the high-utility reference point against the DP methods that are evaluated.

2. DP-SGD Logistic Regression To use differential privacy, we reimplemented logistic regression in PyTorch that is compatible with Opacus [8], a library for DP-SGD.

3. Selective DP-SGD Logistic Regression Selective DP-SGD modifies the DP mechanism to apply gradient clipping and addition of noise only to the sensitive coordinates of the model parameters corresponding to `race` and `sex`.

Both the DP-SGD and S-DP-SGD implementations use proper per-sample gradient computation, where each sample's gradient is individually clipped before aggregation and noise addition, as done in Abadi et al. [3]. For DP-SGD, we used Opacus library's PrivacyEngine. For Selective DP-SGD, we manually use per-sample gradient computation to apply selective noise application.

Identifying Sensitive Parameter Because one-hot encoding expands the categorical attributes `race` and `sex` into several binary columns (e.g., `race_White`, `race_Black`, `sex_Male`, `sex_Female`), we first locate the indices of these expanded columns in the 105-dimension feature space. In our implementation, we identify these indices after preprocessing and store them in a list called `sensitive_indices`. These indices mark which components of the gradients need to be privatized.

Selective Gradient Protection Our implementation of Selective DP-SGD performs the following operations at every training step:

1. **Standard forward and backward pass:** Compute the usual forward pass with predictions, and then compute gradients normally for all parameters.

2. **Create a gradient mask:** Construct a binary mask that identifies which gradient components correspond to sensitive features (`race` and `sex`). Components corresponding to the sensitive features are set to 1 and the rest is set to 0.
3. **Selective gradient clipping:** Apply the gradient norm bound C to the masked (sensitive) gradient components. If the ℓ_2 norm of sensitive gradients exceeds C, they are scaled down to C. Gradients for non-sensitive features (such as education, occupation, and working hours per week) are not clipped.
4. **Selective noise addition:** Add calibrated Gaussian noise of standard deviation $\sigma \cdot C$ only to sensitive gradient components. Non-sensitive gradients do not get any noise addition.
5. **Parameter update:** Update model weights using the combined gradient.

This creates a model where the influence of the attributes `race` and `sex` on the learned parameters is differentially private. Other non-sensitive features contribute to learning without privacy-induced degradation.

Utility

We measure the models' predictive performance on the test set with these metrics:

- Accuracy (our main focus)
- AUC

Fairness

Fairness is assessed using common group fairness metrics across race and sex:

Disparate Impact (Positive Rate):

$$\text{PR}_g = P(\hat{Y} = 1 \mid A = g) \tag{6}$$

True Positive Rate parity (Equal Opportunity):

$$\text{TPR}_g = P(\hat{Y} = 1 \mid Y = 1, A = g) \tag{7}$$

False Positive Rate parity (a component of Equalized Odds):

$$\text{FPR}_g = P(\hat{Y} = 1 \mid Y = 0, A = g) \tag{8}$$

We compute group gaps for each metric (max $-$ min), which measure disparity. Smaller gaps mean greater fairness.

Experimental Setup

Analysis 1: Fixed Configuration Comparison In this first analysis, we compare all three models (No DP, DP-SGD, S-DP-SGD) with a single fixed hyperparameter setting. All models are also trained with identical settings:

- Train-test split: 80/20 with `random_state=42`
- Preprocessing pipeline: shared across all models
- Logistic regression architecture
- Learning rate: 0.01
- Number of epochs: 10
- Batch size: 64

For DP-SGD and Selective DP-SGD, we additionally fix:

- Noise multiplier σ: 1
- Gradient norm bound C: 1

We chose these parameter values because they provide moderate privacy protection (not too low or high).

Results from this analysis are presented in a single comparative table showing utility, fairness, and privacy metrics across all three models. This controlled comparison isolates the effect of the privacy mechanism, while all other factors are held constant.

Analysis 2: Sensitivity Analysis We also do a sensitivity analysis by systematically changing one parameter and keeping all others constant. This is to understand how the privacy-utility-fairness trade-offs vary with the choice of hyperparameters. Specifically, we investigate:

- **Gradient norm bound C:** We vary $C \in \{0.05, 0.5, 0.75, 1.0, 1.5, 2.0, 3.0, 5.0\}$ to study the impact of gradient clipping on models' utility and fairness. Lower C values lead to smaller ϵ (stronger privacy) but potentially lower accuracy and fairness.
- **Noise multiplier σ:** We vary $\sigma \in \{0.5, 0.75, 1.0, 1.5, 2.0, 3.0, 5.0]\}$ to learn how increasing privacy protection (higher noise) affects model utility and fairness. Higher σ values lead to smaller ϵ (stronger privacy) but potentially lower accuracy and fairness.

For each parameter sweep, we train both DP-SGD and S-DP-SGD models. We then plot the resulting accuracy and fairness metrics. This analysis shows:

1. How sensitive each privacy mechanism is to hyperparameter choices
2. Whether S-DP-SGD still has advantages over different hyperparameter settings

The results from this analysis are presented as line plots displaying the relationship between the varied parameter and outcome metrics. This helps us to visualize the privacy-utility-fairness trade off for each privacy model.

Results

Analysis 1: Fixed Configuration Comparison

We start by comparing the three models, Baseline (No DP), DP-SGD, and Selective DP-SGD (S-DP-SGD), on a fixed set of hyperparameters. We then evaluate the trade-offs between privacy, utility, and fairness. The specifications of the hyperparameters are described in the Experimental Setup section above.

Utility Comparison Table 1 presents the baseline predictive model has the highest overall accuracy (85.08%) and AUC-ROC (90.42%).

Table 1: *Model utility metrics under fixed hyperparameters.*

Model	Accuracy	AUC-ROC
Baseline (No DP)	0.851	0.904
DP-SGD	0.832	0.885
S-DP-SGD	0.845	0.899

DP-SGD results in notable utility cost, with accuracy dropping by 1.89 percentage points (pp) and AUC-ROC dropping by 2.03 pp from the baseline model. The decrease in performance occurs because of the privacy protection to all 105 features during training.

S-DP-SGD helps alleviate the loss in utility. It achieves an accuracy of 84.48% (just 0.6 pp lower than the baseline) and an AUC-ROC score of 89.88% (just 0.54 pp lower than the baseline). In comparison to DP-SGD, S-DP-SGD improves accuracy by 1.29 pp. These slight improvements are because in S-DP-SGD, it only protects the data in the 7 sensitive feature indices that correspond to `race` and `sex`, hence maintaining cleaner gradients in the other 103 non-sensitive features.

These results imply that S-DP-SGD outperforms DP-SGD in terms of the trade-off between utility and privacy. Although both algorithms offer formal guarantees of differential privacy, the selective protection mechanism in Selective DP-SGD ensures that the model's performance stays much closer to the baseline model.

Fairness Comparison We evaluate fairness across two protected attributes: `sex` (Male vs. Female) and `race_grouped` (White vs. Non-White). Tables 2 and 4 present group-specific metrics and fairness disparities.

Fairness by Sex From Table 3, we see that DP-SGD decreases fairness for sex. The Disparate Impact (DI) drops from 0.298 (baseline) to 0.149, suggesting that the number of positive predictions for females is less than 15% of the number of positive predictions for males. The TPR difference (Equal Opportunity gap) more than doubles relative to the baseline, from -0.120 to -0.3047, suggesting that in the population where actual earnings are above $ 50k, females are correctly classified at a rate 30.5 pp lower than males. This worsening of bias is consistent with prior findings that DP can disproportionately harm minority groups [1].

Table 2: *Fairness metrics by sex under fixed hyperparameters.*

Model	Group	Pos. Rate	TPR	FPR
Baseline	Male	0.254	0.612	0.097
	Female	0.076	0.493	0.023
DP-SGD	Male	0.227	0.536	0.093
	Female	0.034	0.231	0.009
S-DP-SGD	Male	0.262	0.616	0.107
	Female	0.079	0.479	0.029

Table 3: *Fairness disparities by sex. DI = Disparate Impact, TPR diff = True Positive Rate difference, FPR diff = False Positive Rate difference.*

Model	DI	TPR diff	FPR diff
Baseline	0.298	-0.120	-0.074
DP-SGD	0.149	-0.305	-0.084
S-DP-SGD	0.304	-0.137	-0.079

S-DP-SGD has better fairness for sex. It achieved DI = 0.304 (even better than the baseline of 0.296) and TPR difference = -0.1371 (compared to DP-SGD of -0.298). The False Positive Rate difference remains close to the baseline(-0.079 v.s. -0.074), whereas DP-SGD gets -0.084. For this specific set of hyperparameters, S-DP-SGD has better fairness for sex compared to DP-SGD.

Table 4: *Fairness metrics by race under fixed hyperparameters.*

Model	Group	Pos. Rate	TPR	FPR
Baseline	White	0.208	0.604	0.074
	Non-White	0.112	0.497	0.042
DP-SGD	White	0.169	0.489	0.061
	Non-White	0.119	0.479	0.054
S-DP-SGD	White	0.192	0.564	0.066
	Non-White	0.086	0.411	0.027

Fairness by Race From Table 5, fairness analysis with respect to races indicates an interesting phenomenon. DP-SGD actually has a better overall fairness compared to baseline and S-DP-SGD. DP-SGD has an improved DI (0.704, very close to 1) compared to the baseline, which has DI = 0.537. DP-SGD also results in the TPR difference that is very close to 0 ($+0.009$,) which is a strong signal for fairness. It also has the lowest FPR difference across all

Table 5: *Fairness disparities by race.*

Model	DI	TPR diff	FPR diff
Baseline	0.537	−0.106	−0.032
DP-SGD	0.704	+0.009	−0.007
S-DP-SGD	0.449	−0.154	−0.038

models(−0.007).

For this specific set of parameters, S-DP-SGD does not seem to work very well with fairness for race and consistently produced the lowest fairness metrics. However, the metrics are not that far off from the baseline. For example, S-DP-SGD achieves a DI of 0.449, which is only 0.08 worse than the baseline. The TPR and FPR differences also increased slight for S-DP-SGD compared to baseline.

Summary of Fairness Findings

- **Sex fairness:** S-DP-SGD has better fairness and even surpasses the baseline model for some the fairness metrics.
- **Race fairness:** Unlike in sex fairness, DP-SGD has the best race fairness metrics (even compared to baseline). S-DP-SGD has slightly worse fairness compared to the baseline.

Analysis 2: Sensitivity Analysis

In order to demonstrate privacy-utility-fairness tradeoffs for our models, we conduct a sensitivity analysis. For each hyperparameter, we varied its value while holding the other hyperparameters constant.

Effect of Gradient Clipping Norm (C) The value of the gradient clipping norm C specifies the bound for the gradients to be clipped before adding the noise. Smaller C values give stronger privacy protection but may also limit the model's ability to learn effectively. We sweep $C \in \{0.05, 0.5, 0.75, 1.0, 1.5, 2.0, 3.0, 5.0\}$ while holding the noise multiplier fixed at $\sigma = 1.0$, learning rate = 0.01, epochs = 10, and batch size = 64.

Impact of C on Accuracy Figure 1 shows the accuracy for both DP-SGD and S-DP-SGD across different clipping norm values. When clipping norms are very small (high privacy), DP-SGD's performance is observed to sharply decline to 76.8% accuracy at $C = 0.05$. This is because aggressive uniform clipping really corrupts the data. When the value of C increases, the accuracy of DP-SGD improves gradually to 85.0% for $C = 5.0$.

On the other hand, the behavior of S-DP-SGD is quite counterintuitive. It starts high at 84% accuracy at $C = 0.05$, remains relatively stable through intermediate values, but then *decreases* slightly at higher clipping norms. This is contrary to the general expectation that a decrease in clipping constraints should lead to improvement or at least result in the same accuracy. Putting aside this unexpected pattern, S-DP-SGD maintains higher accuracy at the restrictive end of the spectrum ($C \leq 1.5$).

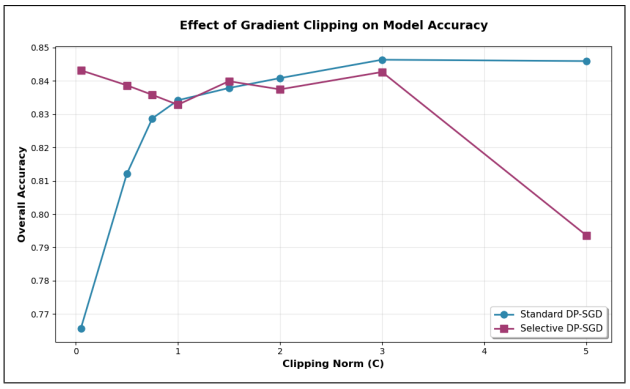

Figure 1: *Effect of gradient clipping norm on model accuracy of DP-SGD (blue circles) and S-DP-SGD (magenta squares)*

Impact on Fairness Metrics Figure 2 presents three key fairness metrics: Disparate Impact, TPR difference (Equal Opportunity Difference), and FPR difference (component of Equalized Odds) as functions of the clipping norm.

Disparate Impact (Top Row): For sex-based fairness (left), neither of the methods performs well. DP-SGD has low but more stable DI values (0.15-0.30). S-DP-SGD shows more variability ($0.10 - 0.6$). S-DP-SGD seems to be more fair across the moderate values of C ($0.5 - 3$), but neither passes the 0.8 fairness threshold. Both models show that females receive fewer positive predictions than males.

For race (right), the trends differ significantly. DP-SGD is quite stable and is within range 0.7 to 1.1. In contrast, S-DP-SGD shows more sensitivity (especially with larger clipping norms). The DI initially fluctuates, but S-DP-SGD is generally closer to the ideal green dashed line at 1.0 for moderate and lower values of C, before increasing sharply at $C = 5$. We find that S-DP-SGD may offer better fairness under stricter privacy (lower clipping norms).

TPR Difference (Middle Row): For sex (left), the TPR difference between the two genders for both models is mostly in the negative range. This means both models still have the bias of males having a higher TPR rate compared to females. For DP-SGD, the pattern is steadier. TPR parity is declining from 0 to −0.30, before rising steadily for larger clipping norms. The TPR parity for S-DP-SGD is unstable but still better than DP-SGD for moderate values of C ($0.5 - 3$), indicated by most of the magenta squares being in the green zone.

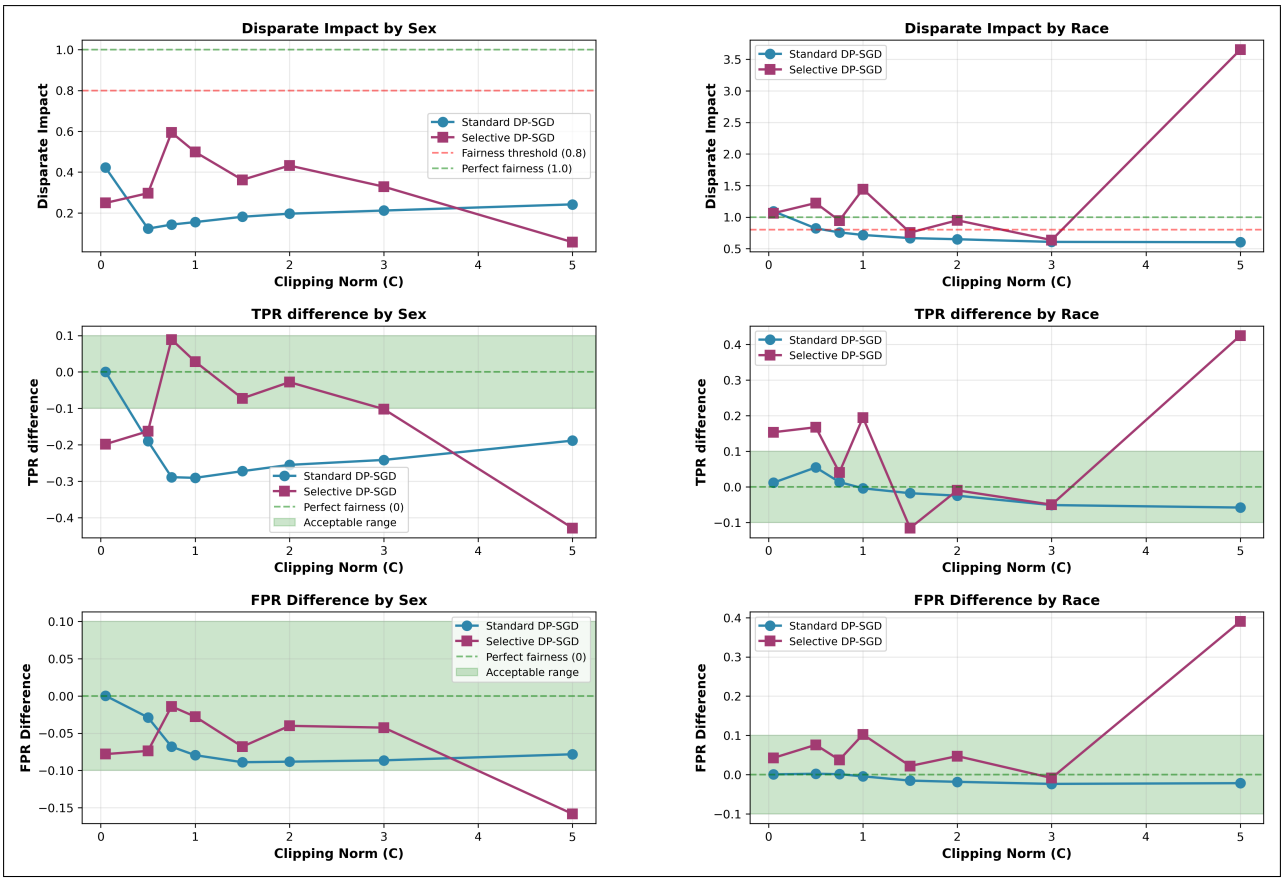

Figure 2: *Effect of clipping norms on fairness metrics by sex (left column) and race (right column). Top row: Disparate Impact. Middle row: TPR Difference. Bottom row: FPR Difference. Green shaded regions indicate acceptable fairness ranges. S-DP-SGD is the magenta squares and DP-SGD is the blue circles.*

For race (right), DP-SGD maintains near-perfect TPR parity (± 0.05). S-DP-SGD shows a concerning fluctuation and upward trend, even reaching $+0.4$ at $C = 5.0$. This mirrors the Analysis 1 where we find that S-DP-SGD is not very good for race fairness.

FPR Difference (Bottom Row): The sex-based FPR differences (left) remain fairly small (-0.15 to 0.00) and within the acceptable ranges for both models. For stricter clipping norms (low values of C), we see that DP-SGD performs better, but for medium values of C (0.75 to 3), S-DP-SGD performs better.

For race (right), DP-SGD keeps FPR parity very close to perfection (± 0.02). S-DP-SGD shows occasional fluctuations but is mostly within the acceptable range. Most of the FPR parity for S-DP-SGD is positive, implying that selectively perturbing race-related gradients may not mitigate bias, but instead could reverse its direction, with non-White groups now exhibiting higher false positive prediction rates than White groups.

Summary of Fairness-Clipping Interactions

- **Divergent fairness across demographics:** S-DP-SGD performs comparably to DP-SGD for sex-based metrics but shows severe instability for racial fairness, particularly at $C \geq 3.0$.
- **Catastrophic failure at high clipping:** S-DP-SGD's racial fairness deteriorates dramatically when clipping is relaxed, with DI exceeding 3.5 and both TPR/FPR differences violating acceptable ranges.
- **DP-SGD stability advantage:** Even with poor sex-based fairness, DP-SGD demonstrates consistent, predictable behavior across clipping values, particularly for racial metrics.
- **Optimal clipping depends on priorities:** For S-DP-SGD, $C \in [0.5, 2.0]$ offers the best compromise. For DP-SGD, higher values ($C \geq 2.0$) improve utility without substantially worsening fairness.

These show mechanisms that improve one fairness dimension may harm another. Therefore, we are required to do a multi-metric evaluation before using the model.

Effect of Noise Multiplier (σ) The noise multiplier σ directly controls the magnitude of Gaussian noise added to gradients, with higher values providing stronger privacy guarantees (lower ϵ) but po-

tentially degrading utility and fairness. We sweep $\sigma \in \{0.5, 0.75, 1.0, 1.5, 2.0, 3.0, 5.0\}$ while keeping the clipping norm fixed at $C = 1.0$, learning rate = 0.01, epochs = 10, and batch size = 64.

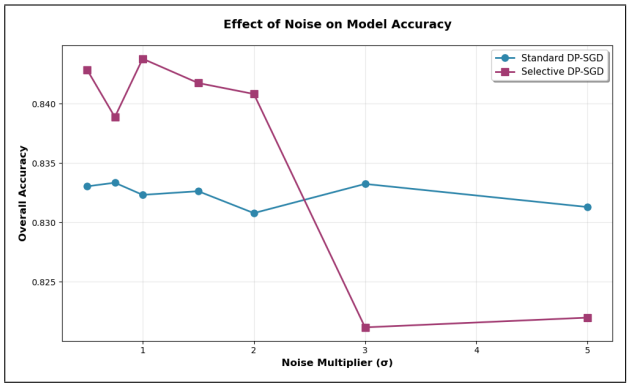

Figure 3: *Effect of noise multiplier on accuracy of DP-SGD (blue) and S-DP-SGD (magenta) models*

Impact of σ on Accuracy From Figure 3, we observe that DP-SGD's accuracy is very stable across the entire noise range (between 83.0% and 83.3%).

S-DP-SGD, on the other hand, is very sensitive to the level of noise. At a low noise level ($\sigma = 0.5$), it achieves the highest accuracy at around 84.3%. Performance remains strong through $\sigma = 2$ (around 84.3%), but then the accuracy drops sharply to approximately 82.1% at $\sigma = 3.0$ and stay around 82.2% at $\sigma = 5.0$. That is a loss of over 2 percentage points compared to the low-noise model. The sharp decline after $\sigma = 2.0$ clearly tells us a threshold beyond which sensitive features become too corrupted to contribute meaningfully.

Impact on Fairness Metrics Figure 4 presents the three fairness metrics: Disparate Impact, TPR Difference, and FPR Difference as functions of the noise multiplier.

Disparate Impact (Top Row): For sex (left), DP-SGD keeps its steady values of DI at around 0.15-0.20 for any level of σ, showing poor yet predictable performance. S-DP-SGD, on the other hand, shows large variation values. Starting at 0.27 ($\sigma = 0.5$), it gradually drops to 0.21 ($\sigma = 0.75$). Then it gradually rises to 0.87 at $\sigma = 3.0$ (approaching the fairness threshold) before collapsing back to 0.16 at $\sigma = 5.0$. The brief improvement at $\sigma = 3.0$ coincides with the accuracy collapse in Figure 3. The model may have shifted toward more conservative (but less accurate) predictions that happen to be more balanced across sex.

For race (right), the pattern is even more extreme. DP-SGD keeps a constant DI of around 0.70–0.75. S-DP-SGD starts at 0.95 (for σ = 0.5), jumps to 1.40 at σ = 0.75 (reflecting extreme over-prediction for Non-White persons), and stays between 0.43 and 1.50 at intermediate points and peaking at 1.50 for σ = 5.0. This volatility of DI for S-DP-SGD shows how unpredictable its racial fairness is and how it depends sensitively on the exact noise level.

TPR Difference (Middle Row): For sex (left), DP-SGD shows consistent negative disparities (-0.28 to -0.30), below the acceptable green zone. S-DP-SGD starts with similar disparities (-0.15 to -0.27 for $\sigma \leq 1.5$), but briefly enters the acceptable range at $\sigma = 2.0$ (-0.05), then returns to -0.24 at $\sigma = 5.0$. Overall, S-DP-SGD is often closer to the acceptable range than DP-SGD.

In terms of race (right), DP-SGD excellently supports TPR fairness (values are within ± 0.02 pp from zero for all noise range). In contrast, S-DP-SGD again shows a fluctuating pattern. Almost no magenta squares are outside the acceptable range, demonstrating how unpredictable and incompatible S-DP-SGD is for racial fairness.

FPR Difference (Bottom Row): For sex (left), both methods remain within or near the acceptable range. DP-SGD shows small negative values (-0.08 to -0.09). S-DP-SGD fluctuates between -0.06 and $+0.03$, with being closest to parity at $\sigma = 3.0$. Across all noise inputs, S-DP-SGD again shows a better sex fairness than DP-SGD, but slightly more volatile.

For race (right), DP-SGD maintains near-perfect FPR parity (± 0.01). S-DP-SGD shows positive FPR differences that generally increase with noise, ranging from $+0.03$ at $\sigma = 0.5$ to $+0.09$ at $\sigma = 0.75$, with notable variation across the range. We can also spot some positive FPR parity for S-DP-SGD at certain values of C, indicating Non-White individuals experience more false positives than White individuals.

Summary of Noise Multiplier Effects

- **Stability** DP-SGD remains stable in both utility and fairness across noise levels. This signifies that applying uniform noise results in a stable learning environment. In contrast, S-DP-SGD shows considerable sensitivity to σ values and depends on σ significantly.
- **S-DP-SGD Sharp Fall:** S-DP-SGD suffers a sharp utility fall at $\sigma \geq 2.0$, losing over 2 pp in accuracy. This suggests that selective noise creates an imbalance at high noise levels, where sensitive features become too corrupted to be useful.
- **Unpredictable fairness:** S-DP-SGD's fairness metrics oscillate unpredictably, making it difficult to tune σ for specific fairness objectives.
- **Optimal operating range:** For S-DP-SGD, $\sigma \in [0.5, 1.5]$ is the best range that gives reasonable

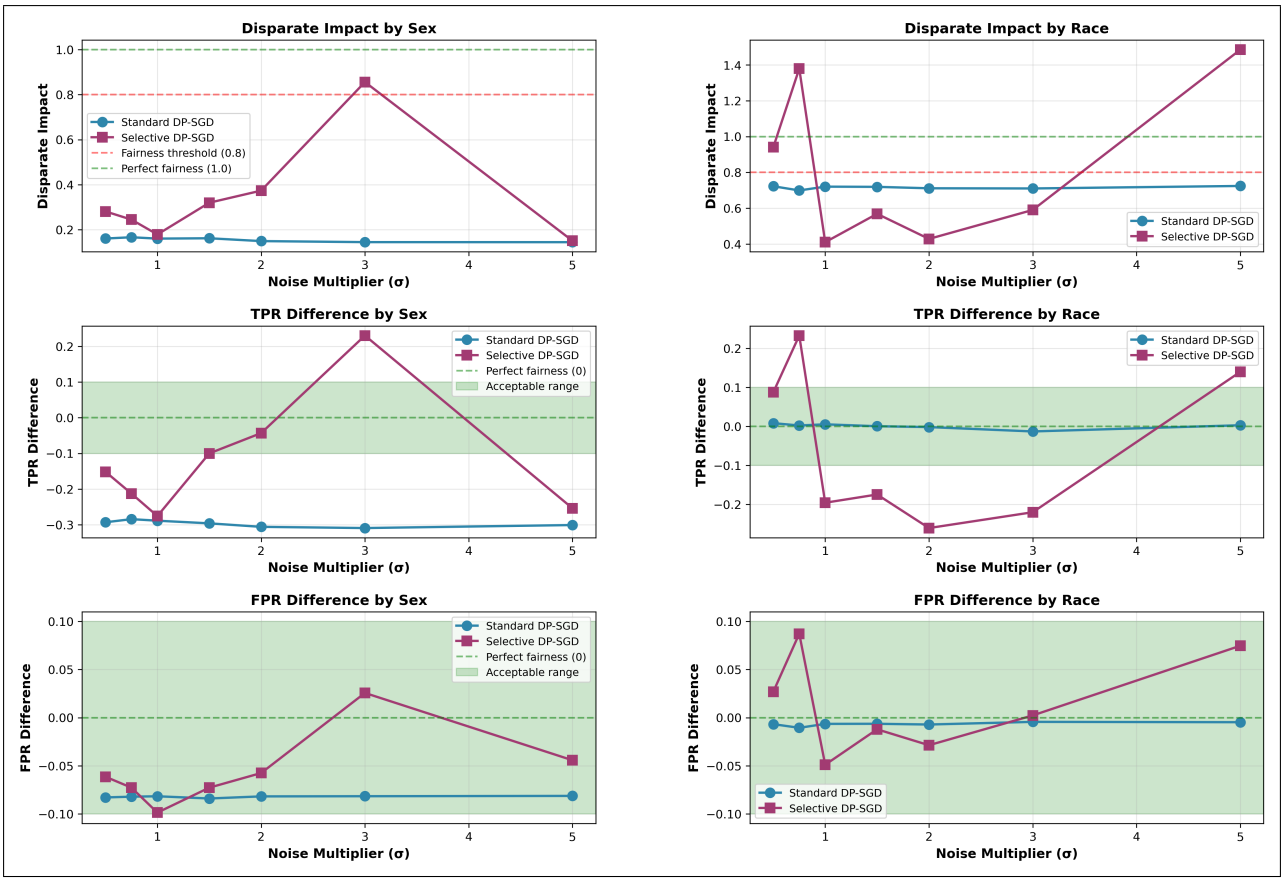

Figure 4: *Effect of noise multiplier on fairness metrics by sex (left column) and race (right column). Top row: Disparate Impact. Middle row: TPR Difference. Bottom row: FPR Difference. Green shaded regions indicate acceptable fairness ranges. S-DP-SGD is the magenta squares, and DP-SGD is the blue circles.*

utility (84.1-84.3% accuracy) with relatively controlled fairness metrics. Beyond this range, both utility and fairness become worse.

- **DP-SGD for high privacy:** When strong privacy guarantees are required (high σ), DP-SGD is the more reliable choice. It has more stable utility and excellent racial fairness across all noise levels, even with slightly low sex fairness.

Discussion

Summary of Key Findings

Analysis 1: Fixed Configuration Results Under identical hyperparameters ($\sigma = 1.0$, $C = 1.0$), S-DP-SGD has utility advantages over DP-SGD, achieving 84.48% accuracy compared to 83.19%—a 1.29 percentage point improvement that narrows the gap with the non-private baseline (85.08%).

However, fairness outcomes are more complex. For sex-based metrics, S-DP-SGD substantially outperformed DP-SGD, having DI of 0.304 (vs. 0.149) and TPR difference of -0.137 (vs. -0.305). Yet for racial fairness, the pattern reversed: DP-SGD achieved better metrics (DI = 0.704, TPR diff = $+0.009$, FPR diff = -0.007) compared to S-DP-SGD (DI = 0.449, TPR diff = -0.154, FPR diff = -0.038). This divergence suggests that privacy mechanisms interact differently with various demographic dimensions (sex vs race). This might also be due to the differences in how these sensitive dimensions correlate with other variables in the dataset.

Analysis 2: Hyperparameter Sensitivity Sensitivity analysis shows significant stability differences between the two privacy methods that we used. DP-SGD is very stable in terms of fairness. The trends are mostly monotonic and predictable. This stability makes DP-SGD a more reliable choice when strong privacy guarantees are required (high noise), in spite of its relatively worse sex-based fairness compared to S-DP-SGD.

S-DP-SGD shows extreme parameter sensitivity. For noise multipliers, S-DP-SGD had a sharp fall beyond $\sigma \geq 2$, with a precision that fell over 2 percentage points. More concerning, fairness metrics exhibited wild, non-monotonic oscillations. For example, racial DI ranged from 0.43 to 2.3 across clipping values, and sex-based Equal Opportunity (TPR

difference) briefly approached parity at $\sigma = 3$ only to collapse at $\sigma = 5$.

This instability is likely from the fundamental asymmetry S-DP-SGD introduces, when only 7 out of 105 features receive noise and clipping. The optimization becomes imbalanced. At high noise levels, sensitive features become so corrupted that the model may either ignore them entirely or fail to properly integrate protected and unprotected information. We found that S-DP-SGD has a narrow good operating range ($C \in [0.5, 2]$, $\sigma \in [0.5, 1.5]$) beyond which both utility and fairness become very unstable and poor.

Answering the Research Question

Back to our central question—*can S-DP-SGD preserve model utility and fairness more effectively than standard DP-SGD?*—the answer is: **it depends on the operating conditions and fairness priorities**.

With moderate privacy settings ($\sigma \leq 1.5$, $C \in [0.5, 2.0]$), S-DP-SGD has several advantages, such as superior accuracy (1 - 1.5 pp improvement) and better sex-based fairness.

However, S-DP-SGD fails when: (1) strong privacy guarantees are needed (high σ or low C), where catastrophic utility collapse occurs; (2) racial fairness is prioritized, because DP-SGD consistently outperforms; (3) operating in environments where hyperparameter tuning is infeasible; and (4) fairness must be guaranteed across multiple demographic dimensions simultaneously.

DP-SGD is very stable across parameter ranges and produces reliable racial fairness metrics. It makes DP-SGD the safer choice for high-stakes applications where robustness matters more than peak performance. The sex-based fairness degradation it causes is worse than S-DP-SGD, but at least it is consistent and quantifiable.

Pitfalls and How We Avoided Them

There are multiple pitfalls we faced along the way with our specified methodology. Here are some examples and our way to avoid them:

- **Data leakage**: We used the same train-test split (80/20 with seed=42) for all three models.
- **Unfair comparison**: We used identical preprocessing, Logistic Regression model, and training parameters. The specific details about the model and parameters we held constant are described in the Methodology section.
- **Cherry-picking results**: We acknowledged the limitation of only doing one set of parameters (Analysis 1) and extended our study with sensitivity analysis (Analysis 2). We tested across 8 clipping norms × 7 noise levels (56 configurations).
- **Single fairness metric**: We evaluated 3 metrics (DI, TPR, FPR) × 2 groups (sex, race) = 6 dimensions of analysis.
- **Overstating findings**: We acknowledged S-DP-SGD's instability and race fairness failures.

Limitations

Several limitations constrain the generalizability of our findings:

Single Dataset: Our analysis focuses only on the UCI Adult dataset. The specific correlation structure between sensitive and non-sensitive features in this dataset may not reflect other domains/datasets.

Model simplicity: We only use logistic regression for the analysis. The privacy-fairness interactions found here may differ in more complex models (like deep learning or neural networks).

Binary protected attributes: We evaluate sex and race separately, ignoring intersectional groups (such as Black women) whose fairness outcomes may differ significantly.

Fixed policy function: We assumed perfect knowledge of which features are sensitive, hand-coding the indices for race and sex. In reality, choosing what is sensitive is subjective and very context-dependent. Incomplete or incorrect policy functions could violate privacy or waste privacy budget.

Single task and metric set: Income prediction may not generalize to other domains (healthcare, criminal justice) where sensitive attributes interact differently with outcomes.

Privacy accounting assumptions: The ϵ values for both models are not directly comparable. Standard DP-SGD uses Opacus' built-in accountant with uniform noise across all parameters. S-DP-SGD uses manual calculation with noise applied only to sensitive features. Empirical testing (such as via membership inference) may be useful for future projects.

Future Research Directions

With more resources and time, we would like to extend our analysis to several directions, such as:

Complex model extensions: Extend the comparison study to multi-layer and deep neural models beyond logistic regression.

Expanded fairness analysis: Study the interaction between selective privacy and a broader set of fairness metrics (beyond the three that we already did).

Adaptive policy functions: Investigate adaptive or learned policy functions for more robust identification of sensitive features (instead of hand-coding like we did).

Empirical privacy testing: Evaluate and compare the privacy budget of different models empirically using inference attack methods.

Selective privacy accounting: Develop a formal privacy accountant tailored to selective DP mechanisms to enable principled and comparable ϵ guarantees.

Conclusion

This paper provides an analysis of Selective DP-SGD and its effect on both utility and fairness, unlike earlier works, which only focused on utility. The findings contradict the claim that more specific privacy tools are necessarily more effective. Although S-DP-SGD is proven to perform well under specific conditions, it is also found to be very unstable and has worse race fairness.

The privacy-fairness analysis is more complex than a simple trade-off between two competing objectives. We observed fairness-fairness trade-offs (sex fairness vs race fairness). There were also stability-efficiency trade-offs (predictable DP-SGD vs. higher-performing but volatile S-DP-SGD). Therefore, it is hard to find a universal privacy method that works well in every aspect.

Our key contribution is to show that while selective privacy tools are appealing theoretically and, at times, empirically better, they also imply certain optimization and fairness problems. The choice between DP-SGD or S-DP-SGD is then based on questions like which features need protection, which groups' fairness matters most, and how much instability we tolerate. Understanding these privacy-utility-fairness trade-offs is important especially with differential privacy becoming more common in machine learning practices.

Links

All the data and the notebook used for the analysis can be accessed through the following link: `https://github.com/Data259-Autumn25/group-project-han-enrico-aston/tree/main`.

References

[1] Ferdinando Fioretto et al. "Differential privacy and fairness in decisions and learning tasks: A survey". In: *arXiv preprint arXiv:2202.08187* (2022).

[2] Weiyan Shi et al. "Selective differential privacy for language modeling". In: *arXiv preprint arXiv:2108.12944* (2021).

[3] Martin Abadi et al. "Deep learning with differential privacy". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 2016, pp. 308–318.

[4] Nicolas Papernot et al. "Tempered sigmoid activations for deep learning with differential privacy". In: *arXiv preprint arXiv:2007.14191* (2020).

[5] Om Thakkar, Galen Andrew, and H Brendan McMahan. "Differentially private learning with adaptive clipping". In: *arXiv preprint arXiv:1905.03871* (2019).

[6] Cynthia Dwork, Aaron Roth, et al. "The algorithmic foundations of differential privacy". In: *Foundations and Trends in Theoretical Computer Science* 9.3-4 (2014), pp. 211–407.

[7] Anderson Santana de Oliveira et al. "An empirical analysis of fairness notions under differential privacy". In: *arXiv preprint arXiv:2302.02910* (2023).

[8] Ashkan Yousefpour et al. "Opacus: User-friendly differential privacy library in PyTorch". In: *arXiv preprint arXiv:2109.12298* (2021).