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

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Enrico Madani   Han Zhang   Aston Tandiono

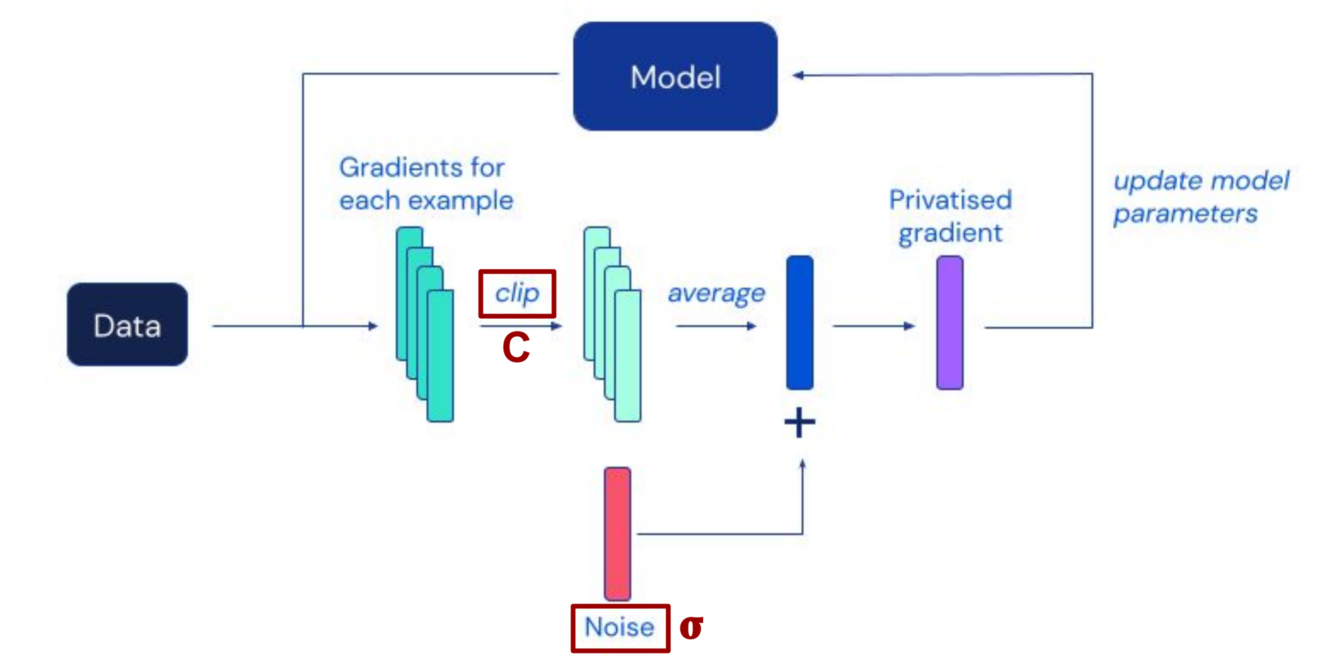
Introduction

Motivation:

- ML models in healthcare, finance, and government use sensitive personal data
- Standard DP-SGD protects ALL features equally → unnecessary utility loss
- Recent work shows DP can **worsen algorithmic bias** for minority groups
- Need methods that balance privacy, accuracy, AND fairness

Research Question:
Can Selective DP-SGD (protecting only race/sex features) achieve better utility and fairness than Standard DP-SGD (protecting all features)?

Differentially Private Stochastic Gradient Descent (DP-SGD)



Selective DP-SGD (S-DP-SGD)
It applies privacy protection ONLY to gradients of sensitive features (race and sex related) and leaving non-sensitive features unperturbed.

Connection to DATA 259 → privacy (DP); fairness (race and sex)

Data

We used the UCI Adult Census Income dataset, which is one of the most studied benchmarks in fairness and privacy research.

This dataset included some missing values in categorical columns. To address this, we imputing using the mode of each column to fill missing values.

Methodology

Baseline	DP-SGD	S-DP-SGD
- No clipping - No noise - No privacy	- Clip ALL features - Noise ALL features - Uniform ε protection	- Clip sensitive features only - Noise sensitive features only - Uniform ε protection

Analysis 1 (Fixed Configuration Comparison)
compared all three models with **one set of hyperparameters** (C=1.0, σ=1.0)

Analysis 2 (Sensitivity Analysis)
varied privacy parameters **one at a time**
C ∈ {0.05, 0.5, 0.75, 1, 1.5, 2, 3, 5},
σ ∈ {0.5, 0.75, 1, 1.5, 2, 3, 5}

Note: C is clipping norm and σ is noise multiplies

Metrics:
Utility: Accuracy and AUC
Fairness: Disparate Impact (DI), TPR parity, and FPR parity

Results

Analysis 1 (Fixed Configuration Comparison) (C=1.0, σ=1.0)

Model	Accuracy	AUC
Baseline	85.08%	90.42%
DP-SGD	83.19%	88.45%
S-DP-SGD	84.48%	89.88%

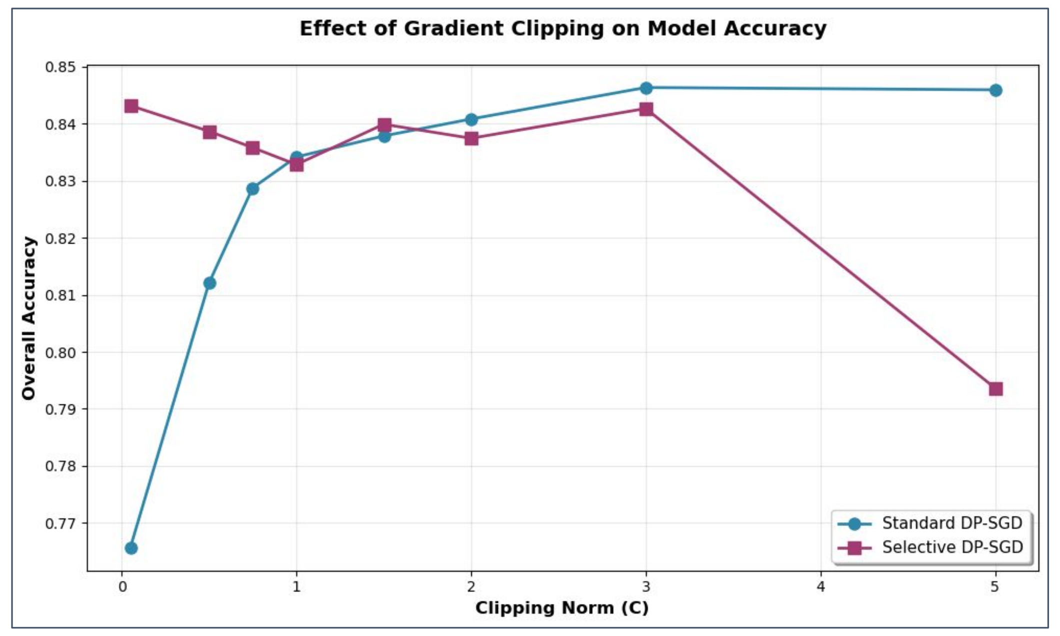
Key Finding: S-DP-SGD **recovers** 68% of utility loss from DP-SGD

	Sex Fairness			Race Fairness		
Model	DI	TPR difference	FPR difference	DI	PR difference	FPR difference
Baseline	0.298	-0.102	-0.074	0.537	-0.106	-0.032
DP-SGD	0.149	-0.305	-0.084	0.704	+0.009	-0.007
S-DP-SGD	0.304	-0.137	-0.079	0.449	-0.449	-0.038

Key Finding: S-DP-SGD **improves** sex fairness but **worsens** race fairness

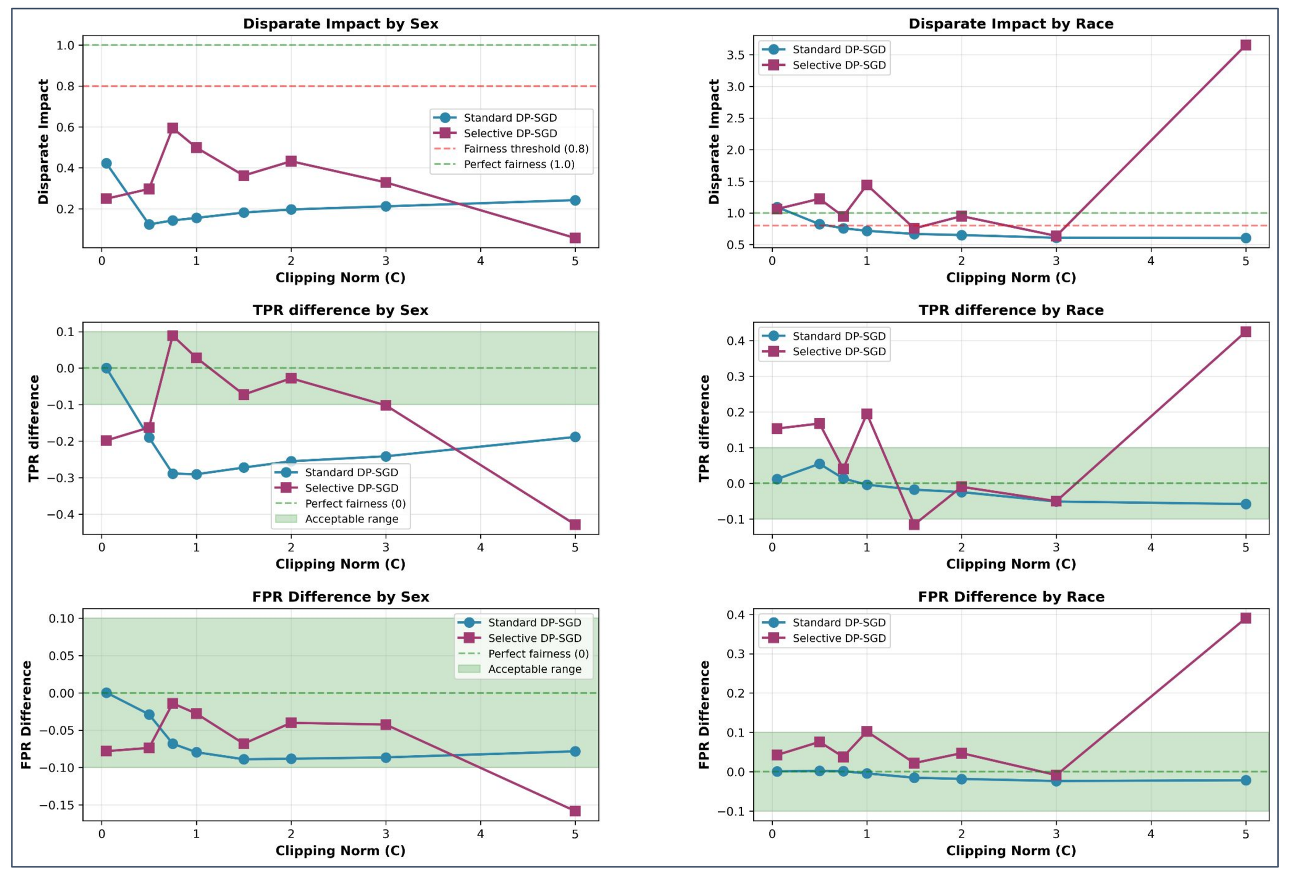
Analysis 2 (Sensitivity Analysis)

C (Clipping Norm)



DP-SGD: Severely hurt by tight clipping (76.8% at C=0.05), gradually recovers to 85% at C=5.0
S-DP-SGD: Starts strong (84% at C=0.05), stays stable, then *unexpectedly drops* at high C

Winner: S-DP-SGD dominates at strict privacy (low C)



Sex Fairness:

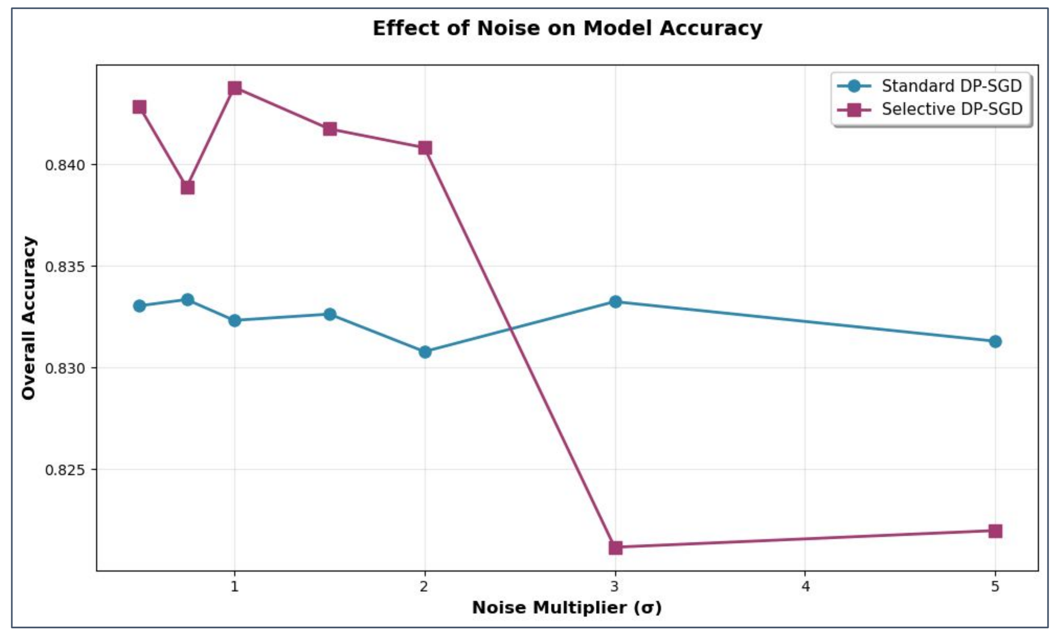
- Both models struggle (DI < 0.8 across all C)
- **S-DP-SGD** more variable but better at moderate C (0.5-3.0)

Race Fairness:

- **DP-SGD:** Rock-solid stable (DI ≈ 0.7-1.1)
- **S-DP-SGD:** Volatile and catastrophic failure at C=5.0 (DI spikes to 3.5+, TPR diff reaches +0.4)

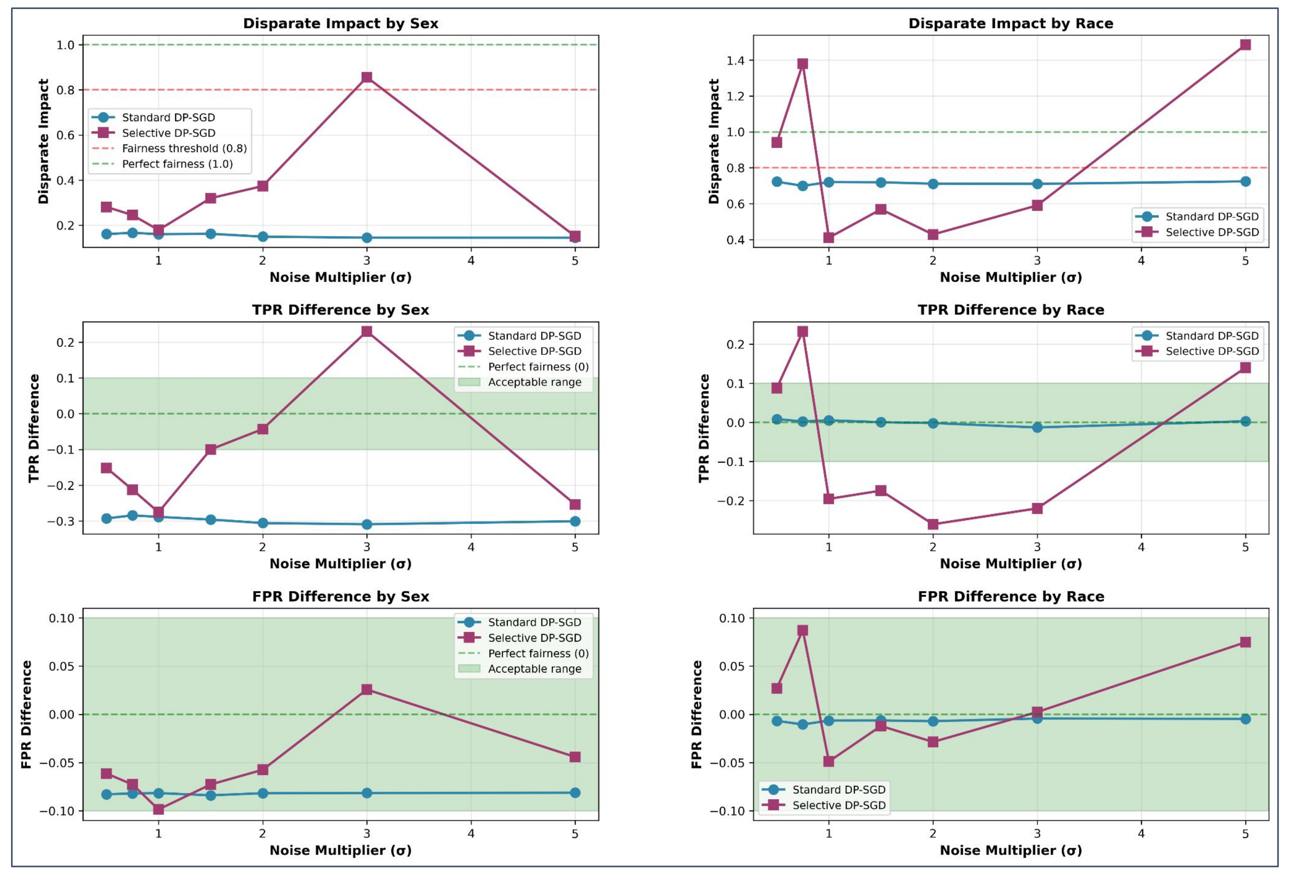
Winner: DP-SGD for stability; S-DP-SGD only safe at C ≤ 2.0

σ (Noise Multiplier)



DP-SGD: Remarkably stable (83.0-83.3%) across ALL noise levels
S-DP-SGD: Strong until σ=2.0 (84.3%), then **crashes** to 82.2% at σ≥3.0 (−2pp drop)

Winner: DP-SGD is more stable and higher accuracy for high noise.



Sex Fairness:

- **DP-SGD:** Poor but predictable (DI ≈ 0.15-0.20)
- **S-DP-SGD:** Erratic (0.16 → 0.87 → 0.16), briefly improves at σ=3.0

Race Fairness:

- **DP-SGD:** Near-perfect stability (DI ≈ 0.70-0.75, TPR diff ±0.02)
- **S-DP-SGD:** Wildly unpredictable (DI swings 0.43 → 1.50, TPR diff all over)

Winner: DP-SGD dominates (especially for race fairness)

Conclusion

Answer: It depends. S-DP-SGD works better ONLY under carefully tuned, moderate privacy settings (**C** ≤ 2.0, **σ** ≤ 1.5).

Implications:

- No universal "best" approach → depends on context
- Privacy-fairness-FAIRNESS triple trade-off (sex vs. race)
- Need careful validation before deployment

The right choice depends on the **fairness priorities** and **stability tolerance.**

Pitfalls Avoided

Pitfall	How We Avoided It
Data leakage	Same train/test split (80/20, seed=42) across ALL models
Unfair comparison	Identical preprocessing, Logistic Regression architecture, training params
Cherry-picking results	Tested across 8 clipping norms × 7 noise levels = 56 configs
Ignoring implementation details	Used per-sample gradients (not averaged), proper privacy accounting via RDP
Single fairness metric	Evaluated 3 metrics (DI, TPR, FPR) × 2 groups (sex, race) = 6 dimensions
Overstating findings	Acknowledged S-DP-SGD's instability and race fairness failures

Limitations

Theoretical Limitations

Policy Function Choice
We hardcoded race/sex as sensitive, but optimal policy is context-dependent

Feature Correlation Ignored
Non-sensitive features (e.g., occupation) correlate with race/sex → potential indirect leakage

Privacy Accounting Uncertainty
Our ε calculation for S-DP-SGD uses standard RDP accountant, which may overestimate privacy cost, so not directly comparable to ε for DP-SGD

Experimental Limitations

Single Dataset
Results may not generalize to other domains/distributions

Linear Model Only
Deep networks may show different patterns

No Membership Inference Attacks
Didn't empirically validate privacy claims via attacks

Fairness Limitations

Group Fairness Only
Didn't evaluate individual fairness or intersectional groups (e.g., Black women)

Binary Race Grouping
Collapsed to White/Non-White loses nuance (e.g., Asian, Native American subgroups)

Fixed Fairness Definitions
Only tested on DI/TPR/FPR, other fairness metrics might be needed

Future Work:

1. **Adaptive policy functions** that learn which features to protect
2. **Expanded fairness analysis** beyod DI, TPR diff, and FPR diff
3. **Deep learning extensions** and intersectional analysis (Black women, etc.)
4. **Empirical privacy testing** via membership inference attacks

References

Ferdinando Fioretto et al. "Differential privacy and fairness in decisions and learning tasks: A survey". In: arXiv preprint arXiv:2202.08187 (2022).
Weiyan Shi et al. "Selective differential privacy for language modeling". In: arXiv preprint arXiv:2108.12944 (2021).
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.
Nicolas Papernot et al. "Tempered sigmoid activations for deep learning with differential privacy". In: arXiv preprint arXiv:2007.14191 (2020).
Om Thakkar, Galen Andrew, and H Brendan McMahan. "Differentially private learning with adaptive clipping". In: arXiv preprint arXiv:1905.03871 (2019).
Additional Sources referenced for our paper