Disparate Impact in Differential Privacy from Gradient Misalignment

Maria S. Esipova
Atiyeh Ashari Ghomi
Yaqiao (Emily) Luo
Jesse C. Cresswell

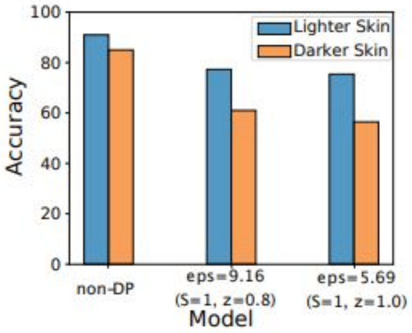
arXiv: 2206.07737

layer6

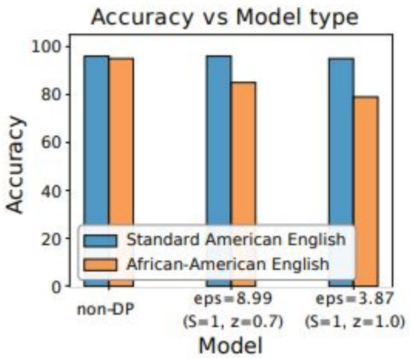
Privacy and Fairness in ML

As machine learning is increasingly applied throughout society, **fairness** and **privacy** become more important concerns.

Privacy and fairness have been extensively studied separately, but their interactions have only come into focus recently



Gender classification



Sentiment analysis

[Bagdasaryan et al. NeurIPS 2019]

layer 6

DP-SGD

Differential Privacy is the most widely used framework for providing rigorous privacy guarantees.

One of the most commonly used private algorithms is **DP-SGD**, which allows training ML models with a DP guarantee.

Take a batch of data → Compute the average loss → Get the gradient of loss for each datapoint → Clip each gradient → Aggregate the clipped gradients → Add noise to the aggregated gradients → Update the parameters

Clipping: each gradient vector is clipped if its norm is greater than a clipping bound

Noise addition: Gaussian noise is added with standard deviation a multiple of the clipping bound

Gradient clipping causes disparate impact

Positive feedback loop of unfairness:

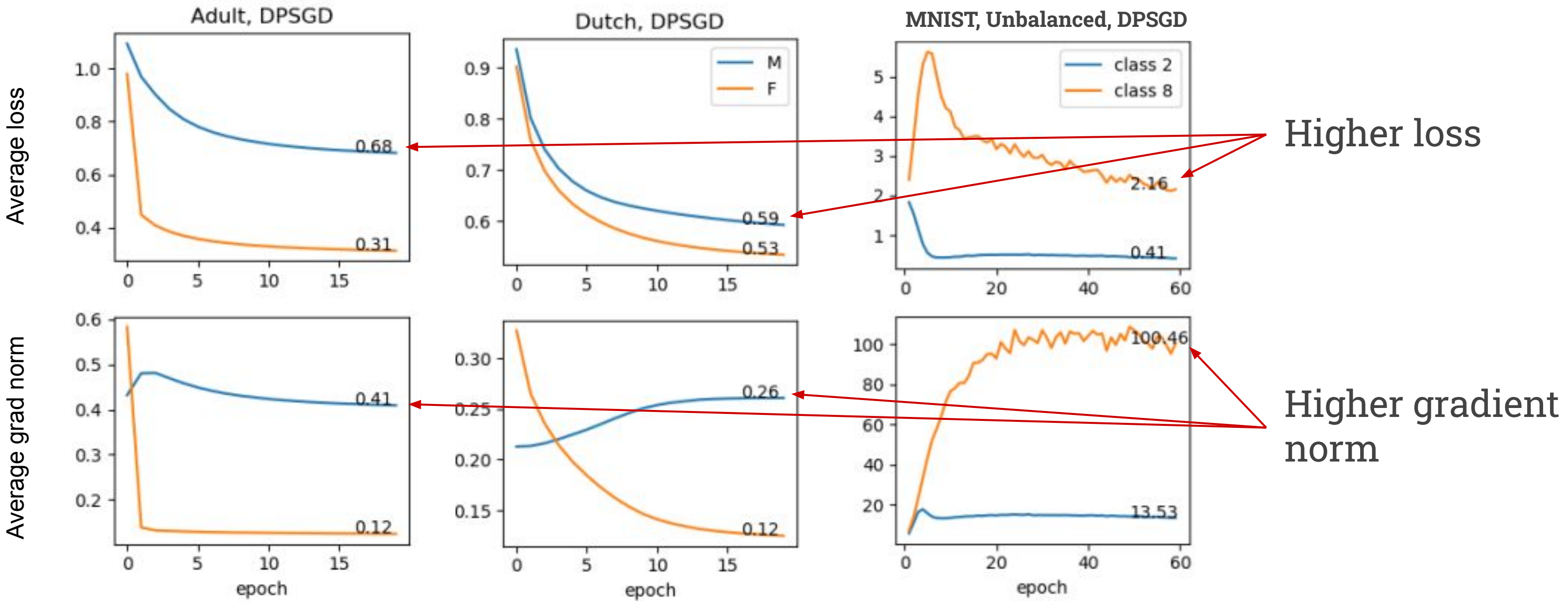
Models need more exposure to data from **underrepresented or complex** groups

These datapoints tend to have **larger gradients**

Hence they are more likely to be **clipped**

Clipping **decreases influence** in the model

Gradient clipping has disparate impact

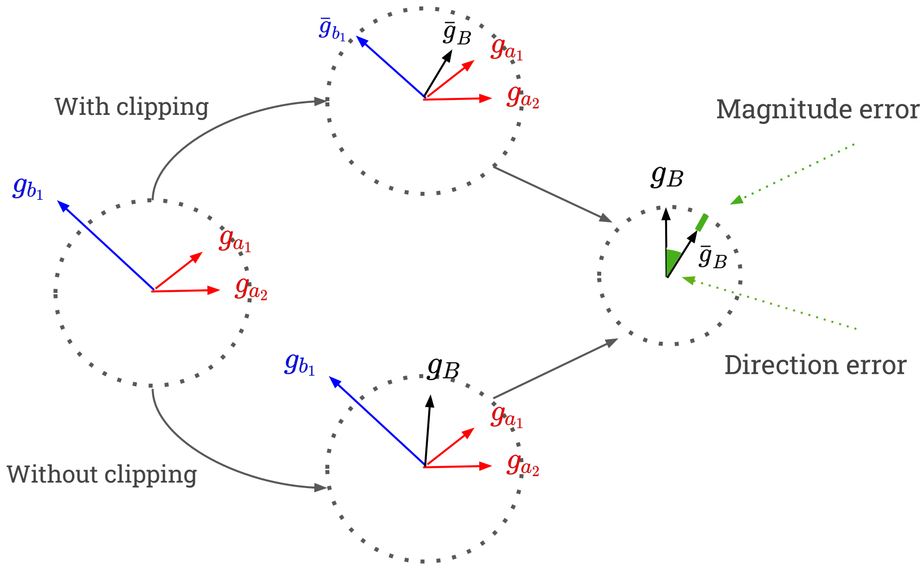


[Xu et al. SIGKDD 2021]

See also [Tran et al. NeurIPS 2021]

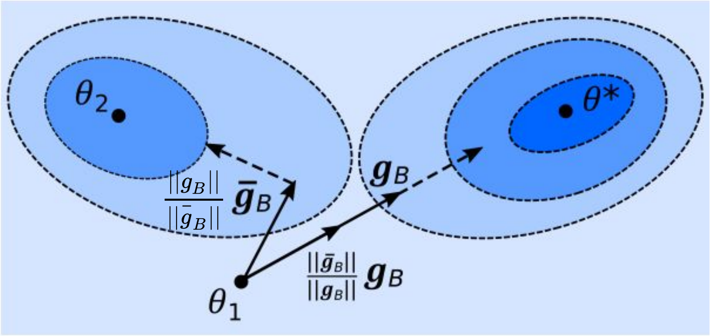
Clipping causes direction errors and magnitude errors

We break down the error due to clipping into direction and magnitude error, and theoretically quantify each.



Direction error is worse than magnitude error

Direction error can result in a worse local optimum. Magnitude error only affects convergence rate.



Effect of direction vs. magnitude error on MNIST with class 8 underrepresented.

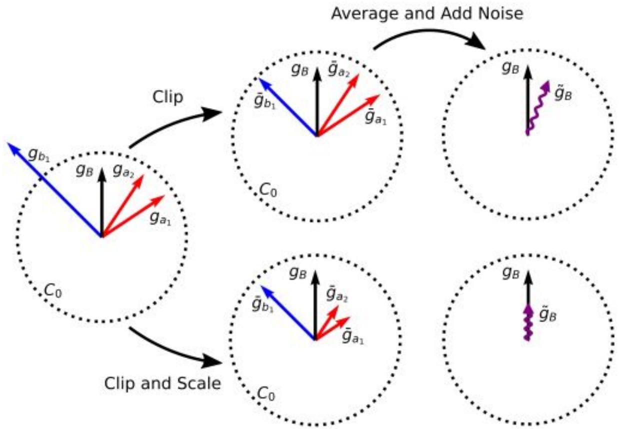
TYPE OF ERROR	ACC 2	ACC 8	LOSS 2	LOSS 8
MAGNITUDE	99.0	93.5	0.002	0.005
DIRECTION	96.8	84.1	0.076	0.518

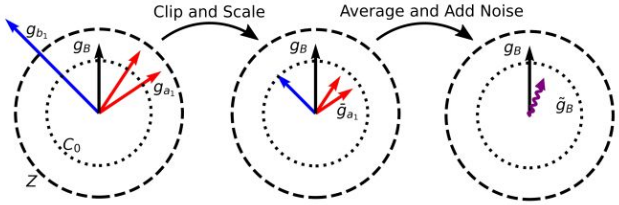
We experimentally isolated the effects of magnitude vs. direction error.

DPSGD-Global and DPSGD-Global-Adapt

Global clipping scales all gradients in order to preserve the direction of the average gradient vector

Global-adaptive clipping finds the best scaling factor adaptively





Top: DP-SGD
Bottom: Global clipping [Bu et al. 2106.07830]

Global adaptive clipping [Ours]

layer 6

Fairness metrics

Adding privacy guarantees negatively affects utility.

A fair private model should have the **cost of privacy shared equally** across groups (no disparate impact).

Excessive risk for group k is $R_k = \mathbb{E}_{\tilde{\theta}}\left[L(\tilde{\theta}; D_k)\right] - L(\theta^*; D_k)$

Excessive risk gap is defined as $R_{a,b} = |R_a - R_b|$

Privacy cost $\pi_k = acc(\theta^*; D_k) - \mathbb{E}_{\tilde{\theta}}\left[acc(\tilde{\theta}; D_k)\right]$

Privacy cost gap $\pi_{a,b} = |\pi_a - \pi_b|$

layer 6

Experimental results

CelebA: Binary classification on gender, with protected groups with/without glasses.

Performance and Fairness metrics for CelebA

Method	Acc W/O	Acc W	$\pi_{W/O}$	π_{W}	$\pi_{W/O, W}$	Loss W/O	Loss W	$R_{W/O}$	R_{W}	$R_{W/O, W}$
Non Private	95.8±0.1	89.7±0.4	-	-	-	0.11±0.00	0.24±0.01	-	-	-
DPSGD	86.5±0.2	74.0±0.6	9.3±0.3	15.7±0.6	6.4±0.7	0.60±0.01	1.34±0.05	0.49±0.01	1.10±0.05	0.61±0.05
DPSGD-F	91.8±0.2	79.7±0.5	4.0±0.2	10.0±0.6	6.0±0.6	0.32±0.01	0.97±0.04	0.21±0.01	0.73±0.04	0.52±0.04
DPSGD-G.	93.1±0.3	82.5±0.5	2.7±0.3	7.2±0.6	4.5±0.5	0.21±0.01	0.57±0.05	0.10±0.01	0.33±0.05	0.24±0.04
DPSGD-G.-A.	94.2±0.1	84.5±0.2	1.6±0.2	5.2±0.5	3.6±0.4	0.17±0.00	0.45±0.01	0.06±0.00	0.21±0.01	0.15±0.01

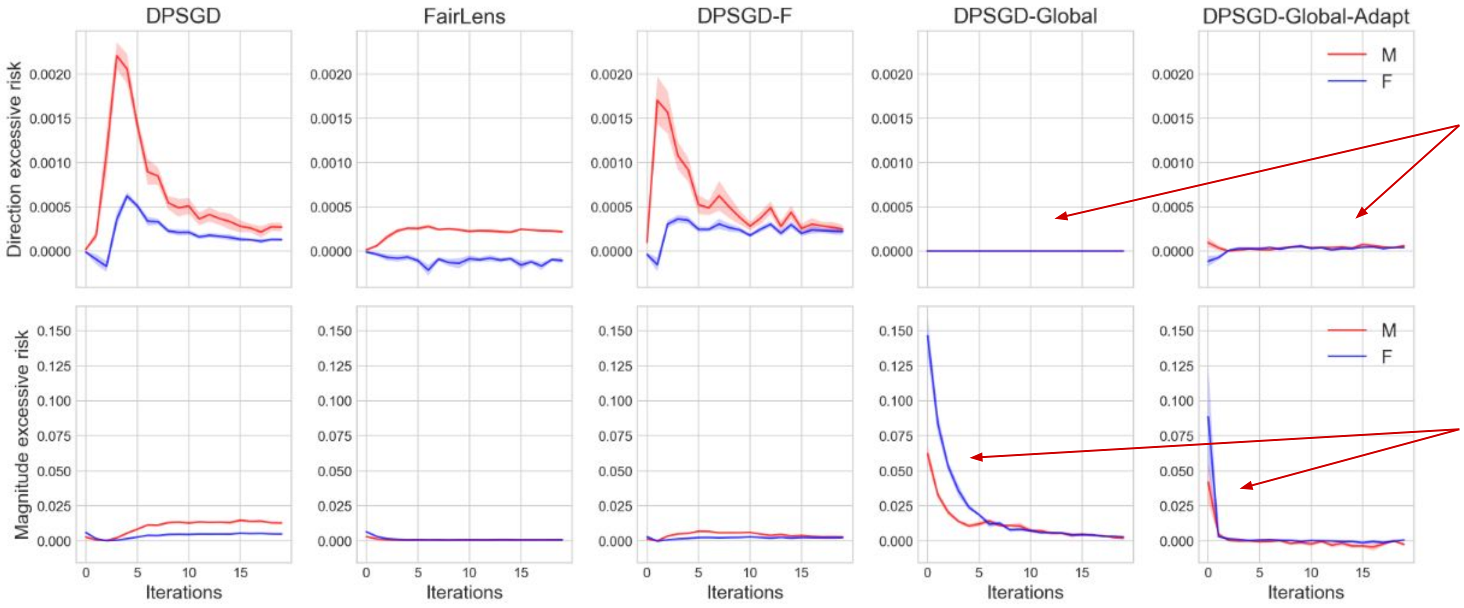
DPSGD-F [Xu et al. SIGKDD 2021] uses group-label information to clip groups differently (disparate treatment).

Global clipping achieves statistically significant improvements in fairness over previous methods. Our adaptive global clipping further improves utility and fairness.

layer 6

Direction and magnitude excessive risk

Adult Dataset



Direction error is eliminated with global clipping

Magnitude error is increased, but this has less impact on final model fairness

layer 6

Conclusion

We derived and quantified that **direction errors are the main source of unfairness** in DPSGD.

We identified that **global clipping minimizes direction errors**, and verified experimentally that it results in more fair models.

We **improved the utility** of global clipping with DPSGD-Global-Adapt.

Unlike previous fair methods, global clipping **does not require group labels** during training. The collection of group labels exposes people to greater privacy risks, and may even be prohibited by laws and regulations.

layer 6