Mitigating Disparate Impact of Differentially Private Learning through Bounded Adaptive Clipping

Linzh Zhao[1], Aki Rehn[1], Mikko A. Heikkilä[1], Razane Tajeddine[2], Antti Honkela[1]
[1]Department of Computer Science, University of Helsinki, Finland
[2]Department of Electrical and Computer Engineering, American University of Beirut, Lebanon
{linzh.zhao, aki.rehn, mikko.a.heikkila, antti.honkela}@helsinki.fi,
razane.tajeddine@aub.edu.lb

Abstract

Differential privacy (DP) has become an essential framework for privacy-preserving machine learning. Existing DP learning methods, however, often have disparate impacts on model predictions, e.g., for minority groups. Gradient clipping, which is often used in DP learning, can suppress larger gradients from challenging samples. We show that this problem is amplified by adaptive clipping, which will often shrink the clipping bound to tiny values to match a well-fitting majority, while significantly reducing the accuracy for others. We propose bounded adaptive clipping, which introduces a tunable lower bound to prevent excessive gradient suppression. Our method improves the accuracy of the worst-performing class on average over 10 percentage points on skewed MNIST and Fashion MNIST compared to the unbounded adaptive clipping, and over 5 percentage points over constant clipping.

1 Introduction

Differential privacy (DP; Dwork et al. 2006b; Dwork & Roth 2014) is a widely accepted framework for preserving privacy in data analysis, including during machine learning model training. While mitigating privacy issues, DP can exacerbate problems with model fairness (Bagdasaryan et al., 2019; Fioretto et al., 2022; Petersen et al., 2023). The current state-of-the-art (SOTA) solution to address fairness issues in differentially private stochastic gradient descent (DPSGD) is based on using adaptive clipping to reduce disparate impacts for underrepresented and confusable groups (Esipova et al., 2023). However, as we demonstrate in this work, the current methods (Andrew et al., 2021; Esipova et al., 2023) can actually suppress gradients from these groups when dynamically adjusting the clipping bounds, leading to very biased estimates and class-wise disparities due to decreased worst-class performance (see Figure 1).

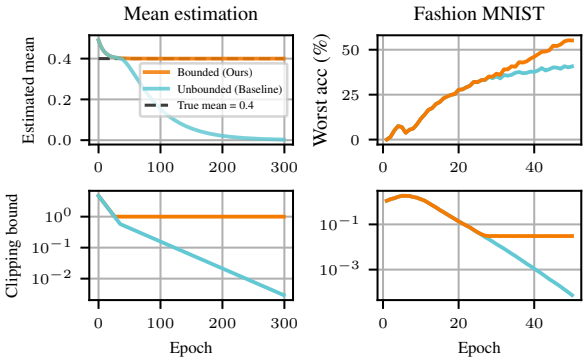


Figure 1: Existing adaptive clipping methods can lead to vanishing clipping bounds (blue), resulting in severe performance degradation for minorities and challenging examples. Setting a lower bound for clipping bound (orange) rectifies this. Left: Mean estimation in bimodal data can converge to the mean of the majority ignoring the minority, Right: DP CNN with Fashion MNIST shows significant impact in most difficult class accuracy.

Preprint. Under review.

To address the issue, we propose **lower-bounded adaptive clipping**, a mechanism aimed at mitigating the limitations of unbounded adaptive clipping. By introducing a tunable lower bound, our method preserves critical gradient updates for underrepresented and confusable groups while maintaining formal DP guarantees. We evaluate the performance of our method under both non-DP and DP hyperparameter optimization (HPO, Liu & Talwar 2019; Papernot & Steinke 2022), and demonstrate its efficiency in comparison to the current SOTA as well as robustness to HPO stochasticity across diverse datasets and model architectures.

Related work Fairness as a formal metric has received significant attention in machine learning, with various formulations proposed (Dwork et al., 2012; Kusner et al., 2017; Corbett-Davies et al., 2017). Ensuring fairness becomes even more challenging when combined with the complexities of DP. Recent work has highlighted that DP can disproportionately degrade the performance of underrepresented or confusable groups, making the mitigation of such fairness disparities a central concern in private learning (Bagdasaryan et al., 2019; Fioretto et al., 2022). In deep learning, accuracy parity, defined as achieving similar accuracy across all demographic or label groups, is considered an important metric in the realm of fairness, and it is especially sensitive to data imbalance and algorithmic design.

To address these challenges through the lens of fairness, one key research direction in recent work has explored improving clipping mechanisms in DP optimization. Starting from DPSGD with gradient clipping (Abadi et al., 2016), but focusing specifically on fairness, Tran et al. (2021) analyzed how constant gradient clipping and loss re-weighting affect model fairness. Xu et al. (2021) proposed DPSGD-Fair, which sets group-specific clipping bounds based on sample sizes and adjusts the noise levels accordingly.

Looking again at the overall performance, Andrew et al. (2021) introduced an adaptive clipping mechanism that tracks a specific quantile of gradient norms under DP, resulting in a clipping threshold that depends on the data distribution. The convergence properties of this method were later analyzed by Shulgin & Richtárik (2024), who provided theoretical guarantees on its performance and utility. To better address fairness, Esipova et al. (2023) proposed an adaptive parameterization for clipping bound updates to mitigate misalignment issues in earlier approaches.

In this work, in addition to the standard constant clipping in DP (Abadi et al., 2016), we use the adaptive clipping methods (Andrew et al., 2021; Esipova et al., 2023) as state-of-the-art baselines to evaluate our approach.

Contributions Our paper makes the following contributions:

- We identify a common failure mode with the current SOTA adaptive clipping methods leading to vanishing clipping bounds, resulting in performance and fairness issues (Figure 1). To address these issues, we propose a novel lower-bounded adaptive clipping method that introduces a tunable lower bound to protect underrepresented and confusable groups (Section 3).
- We evaluate the performance of our proposed approach comprehensively across four datasets and three models, showing its ability to achieve both strong overall performance and fairness under DP constraints compared to existing methods (Section 4.2).
- We test the performance of our method under differentially private hyperparameter optimization (DPHPO), demonstrating that our approach is robust to DPHPO stochasticity compared to existing methods (Section 4.3).

2 Preliminaries

Differential privacy (DP) DP (Dwork et al., 2006b; Dwork & Roth, 2014) is a mathematical framework for privacy preservation, centered on the principle of quantifying privacy through the comparison of output probabilities between adjacent datasets, formalized as follows.

Definition 2.1 *(Approximate DP; Dwork et al., 2006b,a) A stochastic algorithm $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ is (ε, δ)-DP if for any adjacent datasets $D, D' \in \mathcal{D}$, and for any $S \in \mathcal{R}$, it holds that*

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in S] + \delta.$$

In this work, we use sample-level add/remove adjacency, so D and D' are adjacent, if D can be turned into D' by adding or removing a single sample.

Differentially private stochastic gradient descent (DPSGD) To incorporate DP into deep learning, a common approach is to use DPSGD for optimization. DPSGD extends SGD with ℓ_2 norm gradient clipping and noise injection (Song et al., 2013; Abadi et al., 2016). In effect, clipping bounds the influence any single sample can have on the outcome, after which calibrated Gaussian noise is added to the clipped per-sample gradients to guarantee DP (Dwork et al., 2006a).

Algorithm 1 Normalized DPSGD (De et al., 2022)

Input: Iterations T, dataset D of size N, sampling rate q, expected batch size $B = qN$, clipping bound C, noise multiplier σ, initial parameters θ_0.
for iteration $t = 0, 1, \ldots, T-1$ **do**
  $\mathcal{B}_t \leftarrow$ Poisson subsample of D with rate q
  for $(x_i, y_i) \in \mathcal{B}_t$ **do**
    $g_i \leftarrow \nabla\mathcal{L}(f_{\theta_t}(x_i), y_i)$
    $\bar{g}_i \leftarrow g_i \cdot \min(\frac{1}{C}, \frac{1}{||g_i||})$
  end for
  $\tilde{g}_i \leftarrow \frac{1}{B}\left(\sum_{i\in\mathcal{B}_t} \bar{g}_i + \mathcal{N}(0, \sigma^2\mathbf{I})\right)$
  $\theta_{t+1} \leftarrow \mathrm{OptimizerUpdate}(\theta_t, \tilde{g}_i)$
end for

However, the update magnitude in standard DPSGD is influenced by two hyperparameters, the learning rate and the clipping bound, which both affect the update magnitude. Since this interdependence complicates hyperparameter tuning, De et al. (2022) proposed normalizing the learning rate by scaling all gradients by a factor of $\frac{1}{C}$, where C is the clipping bound (see Algorithm 1). This normalization decouples the learning rate and the clipping bound so the hyperparameter C exclusively controls the clipping bound without affecting the update magnitude, simplifying HPO. In the rest of this paper, DPSGD refers to DPSGD with normalization. We note that due to the standard post-processing properties of DP (Dwork & Roth, 2014), any optimizer with access only to the DP gradients, e.g., Adam (Kingma & Ba, 2015), will also satisfy DP. We therefore use the general $\mathrm{OptimizerUpdate}$ in Algorithms 1 and 2 to refer to any such optimizer.

Differentially private hyperparameter optimization (DPHPO) Finding good hyperparameters is critical for ensuring good performance, yet finding them especially under DP constraints is non-trivial due to the high computational cost of DP training (Koskela & Kulkarni, 2023) and the risk of extra privacy leakage from HPO (Liu & Talwar, 2019). Papernot & Steinke (2022) have analyzed DPHPO procedures, showing that privacy leakage can remain modest as long as each training run adheres to DP guarantees.

Considering the intersection of fairness and DP which is the focus of this work, the minority and confusable groups are often most at risk from privacy breaches (Xu et al., 2021). Hence, accounting the privacy budget throughout the entire DP pipeline including HPO can be useful to assert the risk. However, as DPHPO introduces additional randomness into the hyperparameters, it becomes more important to evaluate the robustness of any method to such stochasticity.

3 Adaptive clipping algorithms for DPSGD

While Andrew et al. (2021) first proposed adaptive clipping in the context of DP, Esipova et al. (2023) introduced mechanisms specifically designed to mitigate the disparate impact on different groups. The two algorithms are both special cases of a more general unified unbounded adaptive clipping algorithm, formalized in Algorithm 2, with $C_{\text{LB}} = 0$. This algorithm reduces to that of Andrew et al. (2021) when $\tau = 1$ and to that of Esipova et al. (2023) when $\eta_C = 1$.

3.1 Key hyperparameters

Adaptive clipping involves several hyperparameters that control its behavior and effectiveness. Below, we detail their roles and highlight key observations from prior work and our analysis.

Target quantile (γ) specifies the proportion of gradients with norms exceeding the current threshold that the algorithm aims for (Andrew et al., 2021). Notably, Esipova et al. (2023) incorrectly referred to γ as the 'clipping bound learning rate', which misrepresents its actual role. The clipping bound update will converge exponentially towards a bound where fraction $1 - \gamma$ of the gradients are clipped. The value of γ is directly linked to unfairness, as fraction $1 - \gamma$ of the gradients are ignored when

considering the clipping bound. The specific behaviour observed in the toy model of Figure 1 (left) could be avoided by setting γ to be sufficiently larger than 0.6 to also consider the minority. Things get more difficult when the size of the minority group is unknown and when it is small, because the estimation of extreme quantiles of the gradient norm distribution is less reliable, so simply using $\gamma \approx 1$ is not a silver bullet. The optimal value of γ is tightly coupled to the threshold multiplier τ that we discuss next.

Threshold multiplier (τ) for counting clipped gradients, introduced by Esipova et al. (2023), is a multiplier that determines the upper limit for identifying outlier gradient norms. Gradients with norms exceeding $\tau \cdot C_t$, where C_t is the clipping bound at the current iteration t, are treated as outliers and contribute to updating the clipping bound such that their fraction should become approximately γ. In practice, τ and γ are tightly coupled. Under a fixed gradient norm distribution, the same fixed point for clipping bound adaptation could be reached by changing τ and γ together suitably. Interestingly, both γ and τ were very stable in our experiments and the values $\gamma = 0.5$ and $\tau = 2.5$ were optimal for all datasets used in the experiments.

Algorithm 2 Unified normalized DPSGD with adaptive clipping mechanism (unbounded / lower-bounded)

Input: Iterations T, dataset D of size N, sampling rate q, expected batch size $B = qN$, initial clipping bound C_0, noise multiplier for gradients σ_{grad}, noise multiplier for clipped gradient count σ_{count}, adaptive clipping bound learning rate η_C, threshold multiplier for counting clipped gradients τ, target quantile γ, the lower-bound of adaptive clipping bound C_{LB}.
if unbounded adaptive clipping is used **then**
  Set $C_{\text{LB}} = 0$
end if
Initialize the parameters of the model θ_0 randomly
for $t = 0, 1, \ldots, T-1$ **do**
  $\mathcal{B}_t \leftarrow$ Poisson sample a batch with rate q from D
  for $(x_i, y_i) \in \mathcal{B}_t$ **do**
    $g_i \leftarrow \nabla\mathcal{L}(f_{\theta_t}(x_i), y_i)$
    $\bar{g}_i \leftarrow g_i \cdot \min(\frac{1}{C_t}, \frac{1}{||g_i||})$
  end for
  $\tilde{g}_i \leftarrow \frac{1}{B}\left(\sum_{i\in\mathcal{B}_t} \bar{g}_i + \mathcal{N}(0, \sigma^2_{\text{grad}}\mathbf{I})\right)$
  $\theta_{t+1} \leftarrow \text{OptimizerUpdate}(\theta_t, \tilde{g}_i)$
  $b_t \leftarrow |\{i : ||g_i|| > \tau C_t\}|$
  $\tilde{b}_t \leftarrow \frac{1}{B}(b_t + \mathcal{N}(0, \sigma^2_{\text{count}}))$
  $C_{t+1} \leftarrow \max\left(C_{\text{LB}}, C_t \cdot \exp\left(\eta_C(\tilde{b}_t - \gamma)\right)\right)$
end for

The other hyperparameters are similar to DP-SGD. **Initial clipping bound (C_0)** serves as the initial value for the adaptive clipping mechanism. **Clipping bound learning rate (η_C)** defines the learning rate for the updates of the adaptive clipping bound. **Noise multiplier for clipped gradient counting (σ_{count})** determines the scale of noise added to the privatized estimate of how many gradients exceed the current clipping bound. A larger σ_{count} gives more privacy budgets to the counting mechanism, but may result in less accurate estimates, potentially affecting the stability of the adaptive clipping updates. Following Andrew et al. (2021) we use $\eta_C = 0.2$ and following Esipova et al. (2023) we set $\sigma_{\text{count}} = 10\sigma_{\text{grad}}$.

3.2 A key limitation of unbounded adaptive clipping

The unbounded adaptive clipping method updates the clipping bound solely based on gradient norm statistics, without enforcing the minimum threshold. While the original theoretical analysis by Andrew et al. (2021) assumes non-changing gradient distribution, as training progresses and gradients from well-optimized samples diminish, the estimated proportion of clipped gradients, $\tilde{b}_t$, often falls below the target quantile γ, causing the bound C_t to shrink further. This iterative decay suppresses gradients from harder or underrepresented samples, limiting their influence and harming fairness.

To illustrate this problem, consider a toy (non-DP) mean estimation task as shown in Figure 1 (left). The target distribution is bimodal with 60% of points around 0 and 40% of points around 1, which implies a true mean $\mu = 0.4$. Using the loss function $\mathcal{L}(\hat{\mu}; x) := \frac{1}{2}\sum_i (x_i - \hat{\mu})^2$, the per-sample gradient is $g_i = x_i - \hat{\mu}_t$. We use $\tau = 1$ and target quantile $\gamma = 0.5$. Early in training both classes contribute sizable gradients, but once $\hat{\mu}_t$ approaches the majority value 0, the majority gradients fall beneath the current bound while the minority (ones) still exceed it. Because $\tilde{b}_t < \gamma$, the unbounded rule keeps shrinking C_t until every minority gradient is clipped to the same tiny magnitude, effectively turning the update into a *majority vote*. The estimate is then driven all the way to 0, as traced by the blue curve in Figure 1. In contrast, our bounded scheme (orange) halts the decay of C_t, preserving the influence of the minority gradients, and converges to approximately the correct mean. We observe

the same pattern on the higher-dimensional Fashion-MNIST benchmark in Figure 1 (right), where bounded adaptive clipping consistently yields higher accuracy of the worst-performing class than its unbounded counterpart.

3.3 Bounded adaptive clipping: mitigating disparate impact

To address the limitations of unbounded adaptive clipping, which often results in excessively small clipping bounds, we propose a bounded adaptive clipping mechanism with a tunable lower bound C_{LB}, as described in the lower-bounded version of Algorithm 2. This mechanism ensures that the clipping bound does not shrink below a specified minimum value, allowing the gradients from the challenging samples to continue contributing to the learning.

Returning to the example in Figure 1, we see that bounded adaptive clipping (orange) effectively prevents the exponential decay of the clipping bound seen in unbounded methods (blue) during later stages of training. By enforcing a lower bound, it avoids excessively suppressing the gradients of underrepresented or confusable classes, ensuring that these gradients contribute efficiently to the accumulated updates. This is particularly critical in later epochs, where the majority of samples become well-optimized, leading to smaller gradients. Without a lower bound, gradients from challenging groups risk being overwhelmed by those of well-optimized samples, halting further optimization for these groups.

3.4 Privacy of adaptive clipping

Achieving DP with adaptive clipping requires accounting for the two accesses to the data for the gradients used in the update as well as the counting query needed for adapting the clipping bound. As both of these are based on the Gaussian mechanism, we can obtain their exact composition using Gaussian DP (Dong et al., 2022).

Lemma 3.1 *A composition of two Gaussian mechanisms with sensitivities $\Delta_1 = \Delta_2 = 1$ and noise multipliers σ_1 and σ_2 has exactly same privacy properties as a Gaussian mechanism with sensitivity 1 and noise multiplier*

$$\sigma = \left(\sigma_1^{-2} + \sigma_2^{-2}\right)^{-1/2}. \tag{1}$$

This allows us to evaluate the privacy using standard privacy accountants using the following theorem.

Theorem 3.2 *The adaptive clipping algorithm in Algorithm 2 is (ε, δ)-DP with privacy parameters returned by a privacy accountant using $T = T$, $q = q$, $\sigma = \left(\sigma_{grad}^{-2} + \sigma_{count}^{-2}\right)^{-1/2}$.*

Implementation details are provided in Appendix A.5.

4 Experimental results

The code for replicating all the results will be released with the published version of the paper.

We focus on evaluating the performance of our proposed bounded adaptive clipping method under two main settings: (i) with optimal hyperparameters, i.e., with the hyperparameter values derived from extensive non-DP HPO (Section 4.2), and (ii) with hyperparameters resulting from proper DPHPO (Section 4.3).

Under setting (i) with optimal hyperparameters, we want to show that our proposed method outperforms the baselines when each algorithm is optimally tuned for the task. With setting (ii) using hyperparameters resulting from DPHPO, we want to demonstrate that our method is more robust to the stochasticity in the hyperparameters compared to the baselines.

4.1 Methodology

Grid search In both cases, the joint grid search includes two key hyperparameters: the learning rate η and the clipping parameter: either the fixed bound C for constant clipping, or the lower bound C_{LB} for the bounded adaptive schemes. For the unbounded method, we set $C_{\text{LB}} = 0$. Moreover, Macro

accuracy is adopted as the objective function. To report performance under optimal hyperparameters, we fix the batch size to a near-optimal value in order to control computational cost. In contrast, the DPHPO setting incorporates batch size as an additional tunable parameter. This is enabled by the use of randomized search rather than full grid evaluation, allowing exploration of a higher-dimensional hyperparameter space at reduced cost. Furthermore, our DPHPO setting explicitly accounts for the privacy cost incurred during hyperparameter search. This ensures that the final privacy guarantee reflects both model training and hyperparameter selection, adhering to DP principles.

For adaptive methods, we fix other hyperparameters: target quantile $\gamma = 0.5$, multiplier $\tau = 2.5$, and clipping bound learning rate $\eta_C = 0.2$, based on preliminary sensitivity analysis. Appendix A.3 details the full search protocol.

Models For image recognition, we use ResNet-18 (He et al., 2016), implemented in Timm (Wightman, 2019), with Batch Normalization replaced by Group Normalization (Wu & He, 2020) as is standard in DP training (Maaten & Hannun, 2020). We also include a simple two-layer convolutional neural network (CNN); see Appendix A.5 for details. For tabular datasets, we adopt logistic regression.

Datasets We use two image datasets and two tabular datasets in the evaluations:

Fashion MNIST (Xiao et al., 2017) contains grayscale images of fashion items from 10 balanced categories. The dataset includes 60,000 training samples and 10,000 test samples. This task is particularly challenging due to visual similarity between certain classes, leading to frequent misclassifications.

Skewed MNIST (LeCun & Cortes, 2010; Bagdasaryan et al., 2019) In the skewed MNIST dataset (Bagdasaryan et al., 2019; Xu et al., 2021), class 8 is artificially subsampled to 10% of its original size, leaving approximately 600 samples in the training set, compared to 6,000 samples for the other classes.

Dutch (Van der Laan, 2000) and *Adult* (Becker & Kohavi, 1996) are tabular census datasets. In both cases, we treat "gender" as the protected attribute and balance the dataset to ensure equal representation across genders, following the setup of Esipova et al. (2023).

Appendix A.1 provides more details on the dataset configurations, including the class subsampling process for skewed MNIST, the definition of confusable classes in Fashion MNIST, and our data pre-processing methods.

Metrics For image classification tasks, we report macro-average accuracy (Macro acc (%)) and worst-class accuracy (Worst acc (%)).

Unlike standard (micro-)average accuracy dominated by the majority classes, macro-average accuracy gives equal weight to each class, providing a more balanced perspective on both overall utility and the minimization of performance disparities. Hence, we report macro-average accuracy to assess the overall performance, which is defined as

$$\text{Macro} = \frac{1}{K} \sum_{k=1}^{K} \frac{\text{TP}_k}{N_k}, \tag{2}$$

where K is the number of classes, and TP_k, N_k denote the number of true positives and the total number of samples for class k, respectively.

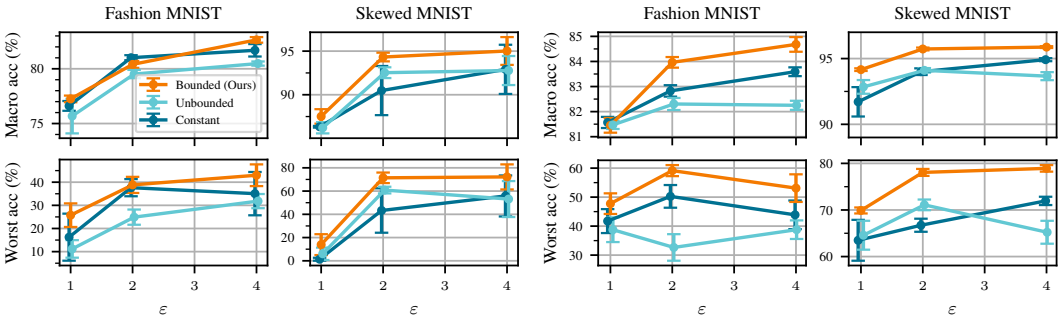
Worst-case accuracy measures the performance of the least accurately predicted class. This metric directly aligns with our goal of mitigating disparate impacts by reducing disparities across classes (Zafar et al., 2017).

For tabular datasets with binary prediction tasks, we report accuracy separately for each sensitive class (Female/Male acc (%)).

Baselines We compare our proposed method with i) *constant clipping* with tuned clipping bound (Abadi et al., 2016), which is the standard clipping method in DP learning, and ii) *unbounded adaptive clipping* (Algorithm 2), which represents the current SOTA for mitigating fairness problems with DPSGD.

4.2 Results with optimal hyperparameters

We first assess the performance of each clipping strategy under optimal hyperparameter settings. This isolates the capability of each algorithm when tuning is unconstrained.



(a) Training ResNet-18 with DP from scratch

(b) Training two-layer CNN with DP from scratch

Figure 2: Comparison of macro accuracy (top) and worst-class accuracy (bottom) across privacy budgets ($\varepsilon, \delta = 10^{-5}$) using optimal hyperparameters. We evaluate three clipping strategies: constant clipping, unbounded adaptive clipping, and our proposed bounded adaptive clipping. Our bounded adaptive clipping consistently outperforms the baselines in both metrics across ε. Results are averaged over 10 seeds, error bars indicate their standard errors. Clusters are slightly shifted for readability.

Image datasets As shown in Figure 2, bounded adaptive clipping consistently achieves the best or near-best performance in both macro accuracy and worst-class accuracy across all privacy budgets ε.

The gains in worst-class accuracy are particularly notable, affirming the fairness benefits of our bounded clipping method. However, we observe that worst-class accuracy does not always increase smoothly with ε, especially for the two-layer CNN on Fashion MNIST (Figure 2b). As the noise level drops, the optimizer increasingly focuses on majority classes, occasionally neglecting minority ones, flattening or even degrading worst-class performance.

These patterns reflect inherent trade-offs in clipping strategies. Constant clipping lacks adaptability and underperforms across both metrics. Unbounded adaptive clipping adjusts to gradient norms, but often shrinks the bound too aggressively, suppressing gradients from underrepresented classes. In contrast, our bounded variant prevents this excessive suppression via a lower-bound constraint. As a result, our algorithm achieves a better balance between macro and worst-class accuracy, and demonstrates robust, stable performance across different privacy budgets.

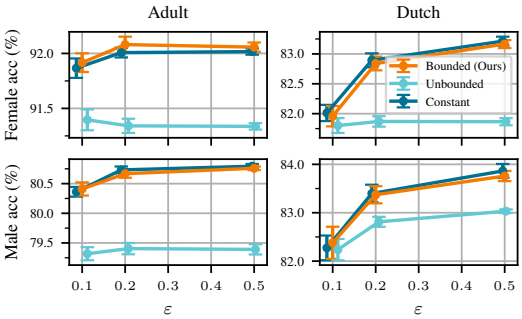


Figure 3: Gender-specific accuracy on Adult and Dutch datasets using logistic regression under DP training. These tabular tasks are relatively simple, and models often reach high accuracy even at small privacy budgets ($\varepsilon, \delta = 10^{-5}$), making it difficult to differentiate methods based on overall accuracy alone. Despite this, bounded adaptive clipping consistently achieves the best or near-best performance across privacy levels and groups. Results are averaged over 10 seeds, with bars indicating standard errors.

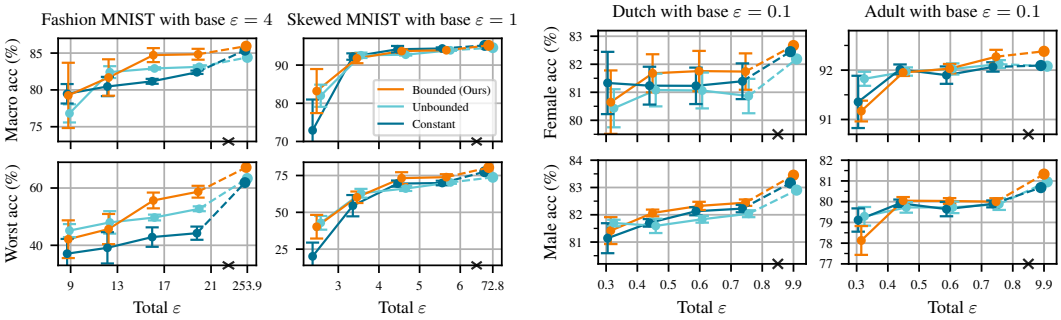
Tabular datasets Bounded adaptive clipping consistently matches or exceeds the best performance across genders and ε levels in Figure 3. Both bounded adaptive clipping and constant clipping can achieve comparable performance, while unbounded adaptive clipping performs noticeably worse. This performance gap can be attributed to the tendency of the unbounded method to shrink the clipping bound excessively, which suppresses gradients, thus reducing subgroup performance. Appendix B.2 provides full performance landscapes across grid points, further confirming the stability of bounded clipping across tuning settings.

Together with the image classification experiments, these results highlight the generality of our approach: it performs well not only in complex, high-variance deep learning tasks, but also in constrained, low-capacity settings typical of DP learning with tabular data.

4.3 Results with DPHPO

To assess our adaptive clipping method within a differentially private hyper-parameter optimization (DPHPO) setting, we begin with a fixed Cartesian grid over the learning rate and clipping-related hyperparameters, then follow Theorem 2 of Papernot & Steinke (2022): the number of grid points we actually evaluate is drawn from a truncated negative-binomial distribution. This randomized stopping rule makes the HPO stage itself differentially private and allows us to account precisely for the privacy budget consumed during tuning.

For the final evaluation of the selected hyperparameters, we conservatively charge privacy based on the total number of grid cells, i.e., the expected number of trials in the DPHPO process is treated as the full grid size for accounting purposes. This aligns with the conservative interpretation used in prior work and ensures the validity of the overall privacy guarantee.



(a) Macro and worst-class accuracy on Fashion-MNIST and skewed MNIST using a two-layer CNN. Bounded adaptive clipping yields consistently better worst-class accuracy.

(b) Gender-specific accuracy on Adult and Dutch datasets using logistic regression. Our method achieves near-optimal performance with smaller extra budgets.

Figure 4: Performance comparison across privacy budgets ($\varepsilon, \delta = 10^{-5}$) with DPHPO. The rightmost symbol in every curve marks the best accuracy obtained anywhere on the grid, and the corresponding ε combines training privacy with the extra cost of evaluating that grid. Our bounded adaptive clipping (orange) consistently preserves worst-class or subgroup accuracy while matching or exceeding the macro-level performance of unbounded and constant clipping. Each line shows the mean over 20 runs; bars indicate standard errors. Clusters are slightly shifted for readability.

Image datasets As shown in Figure 4a, bounded adaptive clipping performs comparably to, and in some cases better than, constant and unbounded adaptive clipping in DP learning scenarios. The inherent noise and the limited number of trials permitted under privacy constraints often hinder DPHPO from reaching optimal performance, making stable methods such as bounded adaptive clipping particularly appealing. Despite these challenges, bounded adaptive clipping consistently achieves better or comparable worst-class accuracy, especially under small ε, compared to the best-performing baselines within the same privacy budget.

For the Fashion MNIST dataset, macro accuracy exhibits minimal differences across the three algorithms. However, in terms of worst-class accuracy, bounded adaptive clipping demonstrates a clear advantage across a range of privacy budgets, ensuring algorithmic fairness by improving the performance of underrepresented classes.

For skewed MNIST, our bounded adaptive clipping achieves macro accuracy that is consistently on par with, or slightly better than, constant clipping. In addition, it outperforms unbounded adaptive clipping, particularly under lower privacy budgets. The higher worst-class accuracy further reinforces the effectiveness of bounded adaptive clipping, as it achieves near-optimal accuracy even with smaller ε. These results highlight the robustness of bounded adaptive clipping in balancing fairness and

utility under various privacy constraints, making it a practical and effective approach for DP training scenarios.

Tabular datasets As shown in Figure 4b, for the Dutch and Adult datasets, bounded adaptive clipping demonstrates improved performance across gender-specific metrics, outperforming unbounded adaptive clipping and constant clipping in most cases. This advantage is more evident under moderate total ε, where bounded adaptive clipping often approaches optimal performance with fewer HPO trials. However, at very low total ε values, the benefits are less consistent.

5 Conclusion

This work studies the intersection of DP and fairness. By its very nature, the DP definition enforces a certain level of unfairness as a single individual, or by group privacy a small group of individuals, will only have a limited impact on the outcome of a DP algorithm. However, the group privacy bounds weaken exponentially with the size of the group and quickly become vacuous. As such, they are not sufficient to explain the unfairness seen in most practical applications.

To improve practical fairness of DP learning, this work introduces bounded adaptive clipping, a novel mechanism aimed at mitigating the disparate impacts caused by unbounded adaptive clipping or constant clipping under differential privacy. By introducing a tunable lower bound for clipping, our method reduces excessive suppression of gradients from underrepresented and confusable groups, alleviating disparate impacts and improving worst-class performance. Extensive experiments across image and tabular datasets show that bounded adaptive clipping outperforms existing methods, demonstrating a robust balance between privacy, fairness, and utility.

Our key findings highlight the advantages of bounded adaptive clipping, including significant improvements in worst-class accuracy and improved robustness, both with optimal hyperparameters and during differentially private hyperparameter tuning. By providing smoother hyperparameter landscapes and achieving competitive performance with small total ε, our approach alleviates the challenges associated with HPO under privacy constraints.

Limitations and Future Directions While our approach demonstrates clear benefits in both macro and worst-class accuracy, especially under tight privacy budgets, several aspects merit further investigation. Key hyperparameters, such as the lower bound, target quantile, and clipping threshold, could be tuned or adapted dynamically across tasks to further enhance performance. Moreover, although we adopt worst-class and macro accuracy to reflect fairness under DP, future work could explore alternative fairness definitions and metrics that are compatible with DP guarantees, potentially uncovering deeper trade-offs between privacy, utility, and representational equity.

Impact statement

This paper aims to contribute to the advancement of machine learning while striving for positive societal impact. By addressing disparate impacts on underrepresented and confusable groups, our work seeks to enhance accuracy parity, thereby promoting greater fairness and equity in machine learning. Through this, we aim to support broader goals of fairness and social justice.

Acknowledgments

This work was supported by the Research Council of Finland (Flagship programme: Finnish Center for Artificial Intelligence, FCAI, Grant 356499 and Grant 359111), the Strategic Research Council at the Research Council of Finland (Grant 358247) as well as the European Union (Project 101070617). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them. The authors wish to thank the CSC – IT Center for Science, Finland for supporting this project with computational and data storage resources.

References

Abadi, M., Chu, A., Goodfellow, I. J., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pp. 308–318. ACM, 2016.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pp. 2623–2631. ACM, 2019.

Andrew, G., Thakkar, O., McMahan, B., and Ramaswamy, S. Differentially private learning with adaptive clipping. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 17455–17466, 2021.

Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 15453–15462, 2019.

Becker, B. and Kohavi, R. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pp. 797–806. ACM, 2017.

De, S., Berrada, L., Hayes, J., Smith, S. L., and Balle, B. Unlocking high-accuracy differentially private image classification through scale. *CoRR*, abs/2204.13650, 2022. doi: 10.48550/ARXIV.2204.13650. URL `https://doi.org/10.48550/arXiv.2204.13650`.

Dong, J., Roth, A., and Su, W. J. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37, 2022. doi: 10.1111/rssb.12454.

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014. doi: 10.1561/0400000042. URL `https://doi.org/10.1561/0400000042`.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology - EUROCRYPT 2006, 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28 - June 1, 2006, Proceedings*, volume 4004 of *Lecture Notes in Computer Science*, pp. 486–503. Springer, 2006a.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. D. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, volume 3876 of *Lecture Notes in Computer Science*, pp. 265–284. Springer, 2006b.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pp. 214–226. ACM, 2012.

Esipova, M. S., Ghomi, A. A., Luo, Y., and Cresswell, J. C. Disparate impact in differential privacy from gradient misalignment. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Fioretto, F., Tran, C., Hentenryck, P. V., and Zhu, K. Differential privacy and fairness in decisions and learning tasks: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pp. 5470–5477. ijcai.org, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Koskela, A. and Kulkarni, T. D. Practical differentially private hyperparameter tuning with subsampling. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4066–4076, 2017.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

Liu, J. and Talwar, K. Private selection from private candidates. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019*, pp. 298–309. ACM, 2019.

Maaten, L. v. d. and Hannun, A. The Trade-Offs of Private Prediction, 2020.

Papernot, N. and Steinke, T. Hyperparameter tuning with renyi differential privacy. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

Petersen, E., Ganz, M., Holm, S. H., and Feragen, A. On (assessing) the fairness of risk score models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023*, pp. 817–829. ACM, 2023.

Shulgin, E. and Richtárik, P. On the convergence of DP-SGD with adaptive clipping. In *OPT 2024: Optimization for Machine Learning*, 2024.

Song, S., Chaudhuri, K., and Sarwate, A. D. Stochastic gradient descent with differentially private updates. In *IEEE Global Conference on Signal and Information Processing, GlobalSIP 2013, Austin, TX, USA, December 3-5, 2013*, pp. 245–248. IEEE, 2013.

Tran, C., Dinh, M. H., and Fioretto, F. Differentially private empirical risk minimization under the fairness lens. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 27555–27565, 2021.

Van der Laan, P. Integrating administrative registers and household surveys. *Netherlands Official Statistics*, 15(2):7–15, 2000.

Wightman, R. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Wu, Y. and He, K. Group Normalization. *International Journal of Computer Vision*, 128(3):742–755, 2020. doi: 10.1007/s11263-019-01198-w.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017. URL `http://arxiv.org/abs/1708.07747`.

Xu, D., Du, W., and Wu, X. Removing disparate impact on model accuracy in differentially private stochastic gradient descent. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pp. 1924–1932. ACM, 2021.

Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J., Cormode, G., and Mironov, I. Opacus: User-friendly differential privacy library in pytorch. *CoRR*, abs/2109.12298, 2021. URL `https://arxiv.org/abs/2109.12298`.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pp. 962–970. PMLR, 2017.

A Experimental details

A.1 Datasets and pre-processsing

In this section, we describe the datasets used in our experiments and the pre-processing steps applied to ensure compatibility with our evaluation objectives.

Skewed MNIST

The skewed MNIST dataset is utilized to investigate scenarios where certain groups are underrepresented. Specifically, following prior work (Bagdasaryan et al., 2019; Xu et al., 2021; Esipova et al., 2023), we create an imbalanced training set by sampling only about 9% of the examples from class 8, while and retaining the standard (balanced) test set. The protected feature in this dataset is the class label, which allows us to study fairness across different classes.

Fashion MNIST

The Fashion MNIST dataset is chosen to study scenarios where some classes are confusable due to feature overlap. Notably, the “Pullover," “Coat," and “Shirt" classes have the highest false positive and false negative rates. This dataset is balanced across classes, making it suitable for evaluating how adaptive clipping mechanisms handle class-specific confusion.

Adult

The Adult dataset is used to evaluate fairness with respect to sensitive attributes rather than class accuracy. Following the pre-processing steps outlined in (Xu et al., 2021; Esipova et al., 2023), we remove the “final-weight" feature and simplify the “race" attribute to a binary feature (white, non-white). Numerical features are normalized, and categorical features are one-hot encoded. The protected attribute for this dataset is “gender", while the target variable is binary income classification (above or below $50,000).

Dutch

The Dutch dataset (Van der Laan, 2000) is used to examine fairness in predictions concerning the protected attribute "gender." Pre-processing involves removing underage samples and the "weight" feature, along with filtering out "unemployed" samples and those with missing or middle-level occupation values. Occupation levels are binarized, with codes 4, 5, and 9 classified as low-level professions and codes 1 and 2 as high-level professions. The task is to predict occupation categories based on remaining features.

A.2 Experiment environment and settings

Our experiments are performed on the clusters, which equipped with AMD EPYC Trento CPU and AMD MI250x GPU for experiments with image dataset, and Xeon Gold 6230 CPU and Nvidia V100 GPU for tabular datasets.

For image datasets, a single training-from-scratch task costs about 10 minutes with one GPU. While for tabular datasets, about 6 minutes are required to execute one training-from-scratch task.

A.3 Hyperparameter optimization protocol and selection of fixed parameters

The sensitivity of hyperparameters For adaptive clipping in particular, we used the Optuna framework (Akiba et al., 2019) to perform hyperparameter optimization under different values of the clipping bound lower bound C_{LB}, including 0, 10^{-4}, and 10^{-2}. Each configuration was optimized over 20 trials with 10 repetitions per trial. We found that optimized values for most hyperparameters, such as the target quantile γ, the clipping threshold τ, and the clipping bound learning rate η_C remained stable between different C_{LB} values. The optimized values are summarized in Table 1, with variation captured by their standard deviations.

Table 1: Mean and standard deviation of adaptive-specific hyperparameters after tuning, across four datasets.

Hyperparameter	Dataset	Mean	Std
Batch size	Skewed MNIST	15070.6	5054.9
	Fashion MNIST	5992.9	1641.3
	Dutch	8556.6	2775.0
	Adult	12320.0	4954.8
Target quantile (γ)	Skewed MNIST	0.5150	0.1691
	Fashion MNIST	0.4828	0.2744
	Dutch	0.6429	0.1713
	Adult	0.5087	0.3267
Clipping multiplier (τ)	Skewed MNIST	2.3002	1.6107
	Fashion MNIST	2.7438	3.1248
	Dutch	2.3864	2.0281
	Adult	2.8830	2.0834
Clipping bound LR (η_C)	Skewed MNIST	0.4250	0.3932
	Fashion MNIST	0.4955	0.4067
	Dutch	0.1402	0.1488
	Adult	0.1691	0.1610

This empirical robustness supports our decision to treat C_{LB} as a separate tuning dimension while fixing the remaining hyperparameters in subsequent experiments. Specifically, we fixed $\gamma = 0.5$, $\tau = 2.5$, and $\eta_C = 0.2$ across all datasets, based on their convergence and consistency. The influence of C_{LB} on training dynamics is further explored in Appendix B.2.

Note: The learning rate and clipping bound lower bound (C_{LB}) were treated as grid search parameters in our experiments and were not included in the table. Specifically, the learning rate was tuned extensively as it directly influences optimization dynamics, and C_{LB} was tested for its impact on adaptive clipping. Their role as search dimensions is further explored in Appendix B.2.

Seeds We selected seeds starting from 1. For instance, if 5 seeds were used, the set of seeds would be 1, 2, 3, 4, 5.

Table 2: Hyperparameter values used in experiments for each dataset. The table specifies the Batch Size, Target Quantile (γ), Clipping Threshold (τ), and Clipping Bound Learning Rate (η_C) for each dataset.

Hyperparameter	Skewed MNIST	Fashion MNIST	Dutch	Adult
Epochs	50	50	40	40
Batch Size	12,000	6,000	10,000	Full-batch
Target Quantile (γ)	0.5	0.5	0.5	0.5
Clipping Threshold (τ)	2.5	2.5	2.5	2.5
Clipping Bound Learning Rate (η_C)	0.2	0.2	0.2	0.2

Privacy accounting To report the (ε, δ)-DP guarantees, δ was fixed at 10^{-5} across all datasets. For Rènyi Differential Privacy (RDP) computations, we used the hyperparameters outlined in Table 2 and relied on Opacus's implementation (Yousefpour et al., 2021).

Initial value of clipping bound The initialization of the clipping bound is a critical aspect of adaptive clipping. Due to the geometric update mechanism introduced by Andrew et al. (2021), the adaptive clipping bound can adjust dynamically across several orders of magnitude during training. This allows the algorithm to efficiently adapt the clipping bound based on the gradient norms observed at each step, accommodating varying distributions of gradient magnitudes. To simplify the process and ensure stability during the initial phases of training, we set the initial clipping bound to 1 across all experiments. This choice strikes a balance between simplicity and generality, as the geometric

updates quickly adapt the clipping bound to appropriate levels during training. Our experiments demonstrated that this initialization works effectively across diverse datasets and hyperparameter settings, further highlighting the robustness of the adaptive clipping mechanism. By keeping the initialization consistent, we also reduce the number of hyperparameters requiring fine-tuning, making the method more practical for real-world applications.

A.3.1 Hyperparameters used in experiments

Based on the evidence and discussion about the sensitivity of some hyperparameters, we used a fixed set of hyperparameters in our experiments to ensure consistency across different datasets. These values include epochs, batch size, target quantile (γ), clipping threshold (τ), and clipping bound learning rate (η_C). Table 2 lists the hyperparameter values used for skewed MNIST, Fashion MNIST, Dutch, and Adult datasets.

It is worth noting that the grid search focused on the learning rate and clipping bound lower bound (C_{LB}). These parameters were excluded from the table as their selection involved a separate evaluation to identify optimal ranges for different datasets. The impact of these parameters is detailed in Appendix B.2.

A.4 Grid design

A.4.1 Grid for reporting performances with optimal hyperparameters

To report the true performances of different algorithms, the experiments should reduce the randomness from hyperparameter optimization. Specifically, all of the reasonable combination of hyperparameters should be tested. Nonetheless, considering the computation cost, we fixed the less sensitive hyperparameters

Table 3: Hyperparameter search space used during tuning. All parameters were treated as categorical.

Hyperparameter	**Ranges**
Learning rate	1.0000, 1.2915, 1.6681, 2.1544, 2.7826, 3.5938, 4.6416, 5.9948, 7.7426, 10.0000
Clipping bound	0.0010, 0.0018, 0.0031, 0.0055, 0.0098, 0.0172, 0.0305, 0.0539, 0.0952, 0.1682, 0.2973, 0.5254, 0.9285, 1.6409, 2.9000, 5.1252, 9.0579, 16.0082, 28.2915, 50.0000

A.4.2 Grid for random search in DPHPO

Moreover, random search is widely used to enable privacy accounting in DPHPO (Liu & Talwar, 2019; Papernot & Steinke, 2022). This approach also requires a predefined grid, from which random samples are drawn.

Table 4: Hyperparameter search space used during tuning. All parameters were treated as categorical.

Hyperparameter	**Ranges**
Batch size	1024, 2048, 4096, 8192, 16384, 32768
Learning rate	1.0000, 1.2915, 1.6681, 2.1544, 2.7826, 3.5938, 4.6416, 5.9948, 7.7426, 10.0000
Clipping bound	0.0010, 0.0018, 0.0031, 0.0055, 0.0098, 0.0172, 0.0305, 0.0539, 0.0952, 0.1682, 0.2973, 0.5254, 0.9285, 1.6409, 2.9000, 5.1252, 9.0579, 16.0082, 28.2915, 50.0000

A.5 Implementation details

We build our work on Opacus (Yousefpour et al., 2021), a framework designed for training models with differential privacy. Below, we describe the network architectures used in our experiments.

Logistic regression

The logistic regression model consists of a single linear layer that maps the input features directly to the output. The output is passed through a sigmoid activation function to produce probabilities for binary classification tasks.

Convolutional neural network (CNN)

The CNN model has the following structure:

- Two convolutional layers, each followed by a max-pooling layer. The first convolutional layer has 64 filters with a kernel size of 3, followed by a max-pooling layer with a kernel size of 3 and stride 2. The second convolutional layer also has 64 filters with similar configurations.
- Three fully connected layers: the first two layers have 500 units each, and the final fully connected layer outputs predictions for the number of classes in the task.
- ReLU activation functions are used between layers to introduce non-linearity.

ResNet-18

We use the ResNet-18 architecture (He et al., 2016) for image classification tasks, implemented via the `timm` library (Wightman, 2019). To comply with standard practices in differentially private training, we replace all Batch Normalization layers with Group Normalization (Wu & He, 2020). All ResNet models are trained from scratch, without any use of pretrained weights.

These architectures are optimized for DP training, ensuring compatibility with privacy constraints while maintaining competitive performance.

B Full result of experiments

B.1 Data summary

The data presented in the Figure 2 are reported in detail as tables in Tables 5 and 6.

Table 5: Comparison of macro accuracy and worst-class accuracy across algorithms on the Fashion MNIST and Skewed MNIST datasets under varying privacy budgets (ε). Bounded adaptive clipping consistently achieves higher worst-class accuracy while maintaining competitive macro accuracy.

Dataset	ε	**Algorithm**	**Macro Acc.**	**Worst-Class Acc.**
Fashion MNIST	1.0	Bounded Adaptive	**0.7725** $\pm$ 0.0030	**0.2575** $\pm$ 0.0510
		Constant Clipping	0.7661 $\pm$ 0.0045	0.1625 $\pm$ 0.1016
		Unbounded Adaptive	0.7568 $\pm$ 0.0158	0.1117 $\pm$ 0.0379
	2.0	Bounded Adaptive	0.8040 $\pm$ 0.0032	**0.3882** $\pm$ 0.0348
		Constant Clipping	**0.8101** $\pm$ 0.0023	0.3763 $\pm$ 0.0371
		Unbounded Adaptive	0.7952 $\pm$ 0.0038	0.2490 $\pm$ 0.0328
	4.0	Bounded Adaptive	**0.8264** $\pm$ 0.0025	**0.4300** $\pm$ 0.0469
		Constant Clipping	0.8169 $\pm$ 0.0057	0.3505 $\pm$ 0.0935
		Unbounded Adaptive	0.8048 $\pm$ 0.0019	0.3183 $\pm$ 0.0306
MNIST	1.0	Bounded Adaptive	**0.8752** $\pm$ 0.0083	**0.1391** $\pm$ 0.0898
		Constant Clipping	0.8633 $\pm$ 0.0010	0.0136 $\pm$ 0.0126
		Unbounded Adaptive	0.8621 $\pm$ 0.0064	0.0603 $\pm$ 0.0531
	2.0	Bounded Adaptive	**0.9432** $\pm$ 0.0050	**0.7141** $\pm$ 0.0453
		Constant Clipping	0.9048 $\pm$ 0.0282	0.4328 $\pm$ 0.1912
		Unbounded Adaptive	0.9253 $\pm$ 0.0062	0.6086 $\pm$ 0.0288
	4.0	Bounded Adaptive	**0.9502** $\pm$ 0.0159	**0.7219** $\pm$ 0.1078
		Constant Clipping	0.9291 $\pm$ 0.0282	0.5580 $\pm$ 0.1765
		Unbounded Adaptive	0.9277 $\pm$ 0.0166	0.5313 $\pm$ 0.1548

Table 6: Accuracy for female and male groups across tabular datasets, algorithms, and privacy budgets (ε). As the differences between our bounded adaptive clipping and the constant clipping baseline are not statistically significant, no values are highlighted.

Dataset	ε	**Algorithm**	**Female Acc.**	**Male Acc.**
Adult	0.1	Bounded Adaptive	0.9192 $\pm$ 0.0009	0.8041 $\pm$ 0.0011
		Constant Clipping	0.9187 $\pm$ 0.0009	0.8036 $\pm$ 0.0008
		Unbounded Adaptive	0.9140 $\pm$ 0.0010	0.7932 $\pm$ 0.0011
	0.2	Bounded Adaptive	0.9208 $\pm$ 0.0007	0.8067 $\pm$ 0.0007
		Constant Clipping	0.9201 $\pm$ 0.0004	0.8073 $\pm$ 0.0006
		Unbounded Adaptive	0.9134 $\pm$ 0.0006	0.7940 $\pm$ 0.0010
	0.5	Bounded Adaptive	0.9206 $\pm$ 0.0004	0.8076 $\pm$ 0.0003
		Constant Clipping	0.9202 $\pm$ 0.0003	0.8079 $\pm$ 0.0004
		Unbounded Adaptive	0.9134 $\pm$ 0.0003	0.7939 $\pm$ 0.0009
Dutch	0.1	Bounded Adaptive	0.8196 $\pm$ 0.0017	0.8238 $\pm$ 0.0033
		Constant Clipping	0.8201 $\pm$ 0.0014	0.8227 $\pm$ 0.0025
		Unbounded Adaptive	0.8180 $\pm$ 0.0012	0.8224 $\pm$ 0.0022
	0.2	Bounded Adaptive	0.8284 $\pm$ 0.0011	0.8337 $\pm$ 0.0018
		Constant Clipping	0.8289 $\pm$ 0.0012	0.8340 $\pm$ 0.0018
		Unbounded Adaptive	0.8187 $\pm$ 0.0009	0.8281 $\pm$ 0.0010
	0.5	Bounded Adaptive	0.8316 $\pm$ 0.0007	0.8376 $\pm$ 0.0011
		Constant Clipping	0.8321 $\pm$ 0.0008	0.8386 $\pm$ 0.0014
		Unbounded Adaptive	0.8187 $\pm$ 0.0006	0.8303 $\pm$ 0.0004

B.2 Heat-map of the landscape among different metrics on datasets

In this subsection, we provide heatmaps to visualize the landscape of hyperparameter optimization across different metrics for the datasets used in our experiments. The heatmaps illustrate how the learning rate and the clipping bound lower bound (C_{LB}) interact to influence performance across various metrics. Each dataset is analyzed under its specific privacy budget (ε), and the metrics reported are tailored to the characteristics of the dataset.

For Fashion MNIST, a balanced dataset, we report macro accuracy, worst-class accuracy, and loss, omitting micro accuracy as it is nearly identical to macro accuracy, which are shown in Figure 5. The heatmap shows that the hyperparameter landscape for macro and worst-class accuracy largely overlaps, indicating robust performance across different objectives.

For skewed MNIST, a class-imbalanced dataset, we report macro accuracy, worst-class accuracy, micro accuracy, and loss in Figure 6. The inclusion of micro accuracy highlights the discrepancies between class-weighted metrics (macro) and sample-weighted metrics (micro), showcasing how imbalance affects the optimization landscape. The heatmap reveals that the optimal regions for macro and micro accuracy are closely aligned, but worst-class accuracy demonstrates a more restrictive optimal range, indicating its sensitivity to hyperparameters.

For Adult and Dutch datasets, we focus on the accuracy gap between genders and the overall loss in Figures 7 and 8. These datasets are used to evaluate fairness-related metrics, with male and female accuracies reported separately. The heatmaps highlight how hyperparameters influence gender disparities in accuracy. While minimizing loss generally aligns with optimizing male and female accuracies, the gender gap exhibits a more nuanced response, requiring careful hyperparameter tuning to ensure fairness.

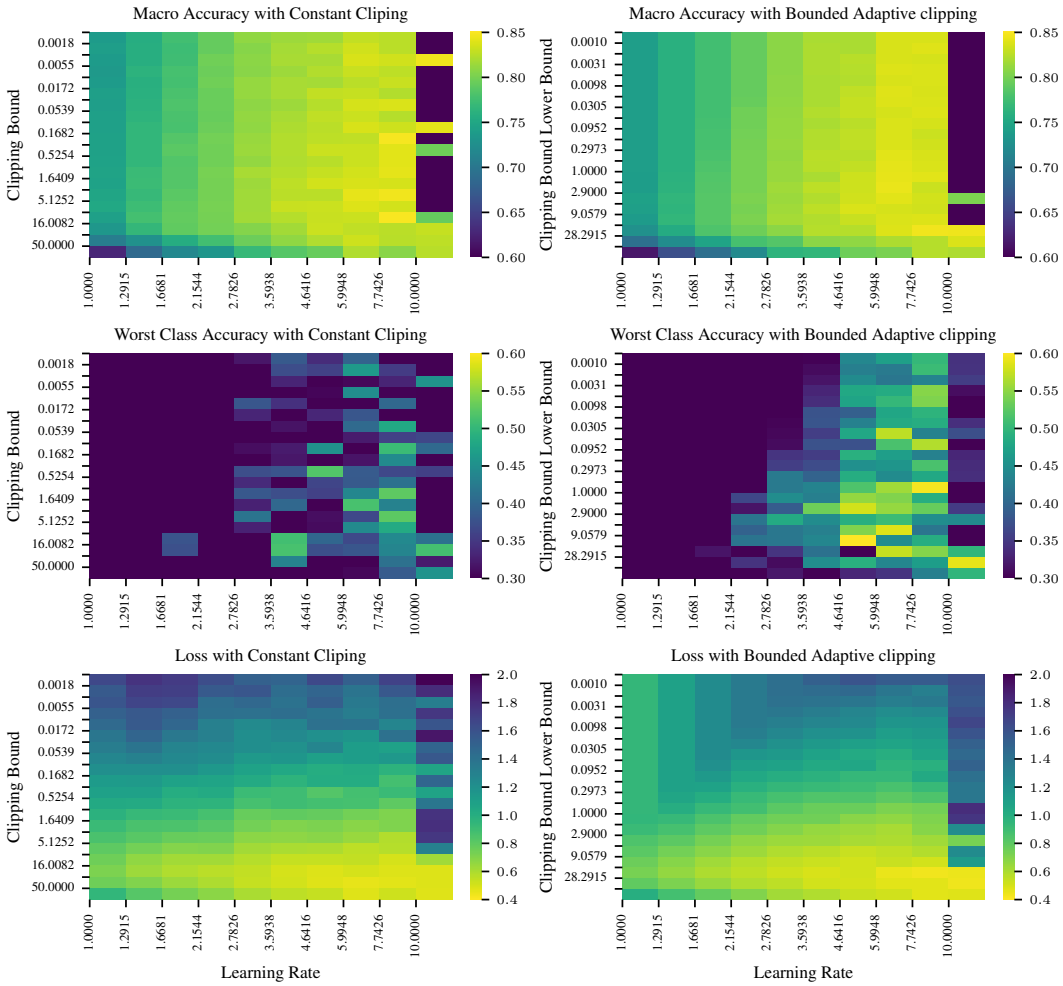


Figure 5: The heatmap of macro accuracy on Fashion MNIST with $\varepsilon = 4.0$.

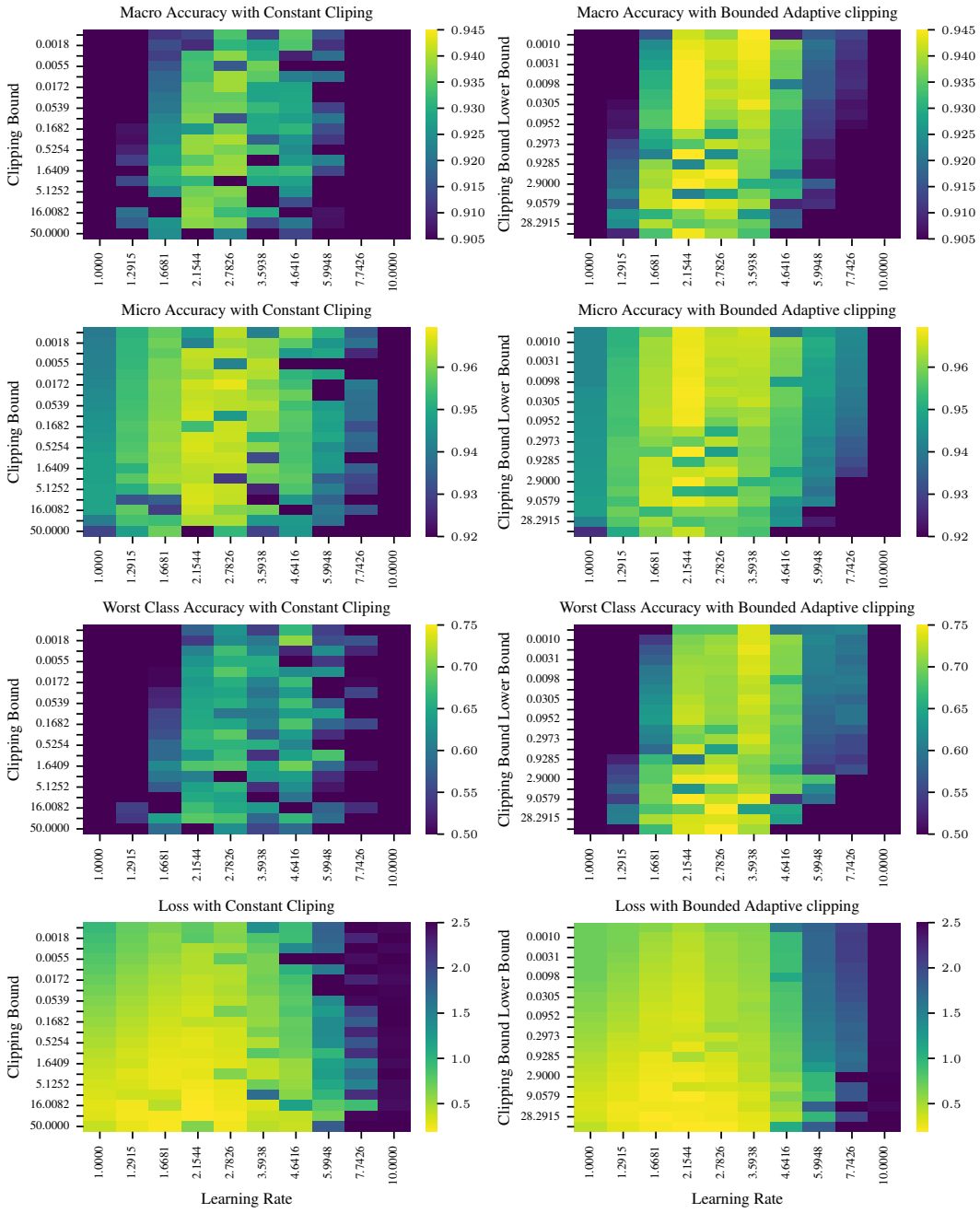


Figure 6: The heatmap of macro accuracy on skewed MNIST with $\varepsilon = 1.0$.

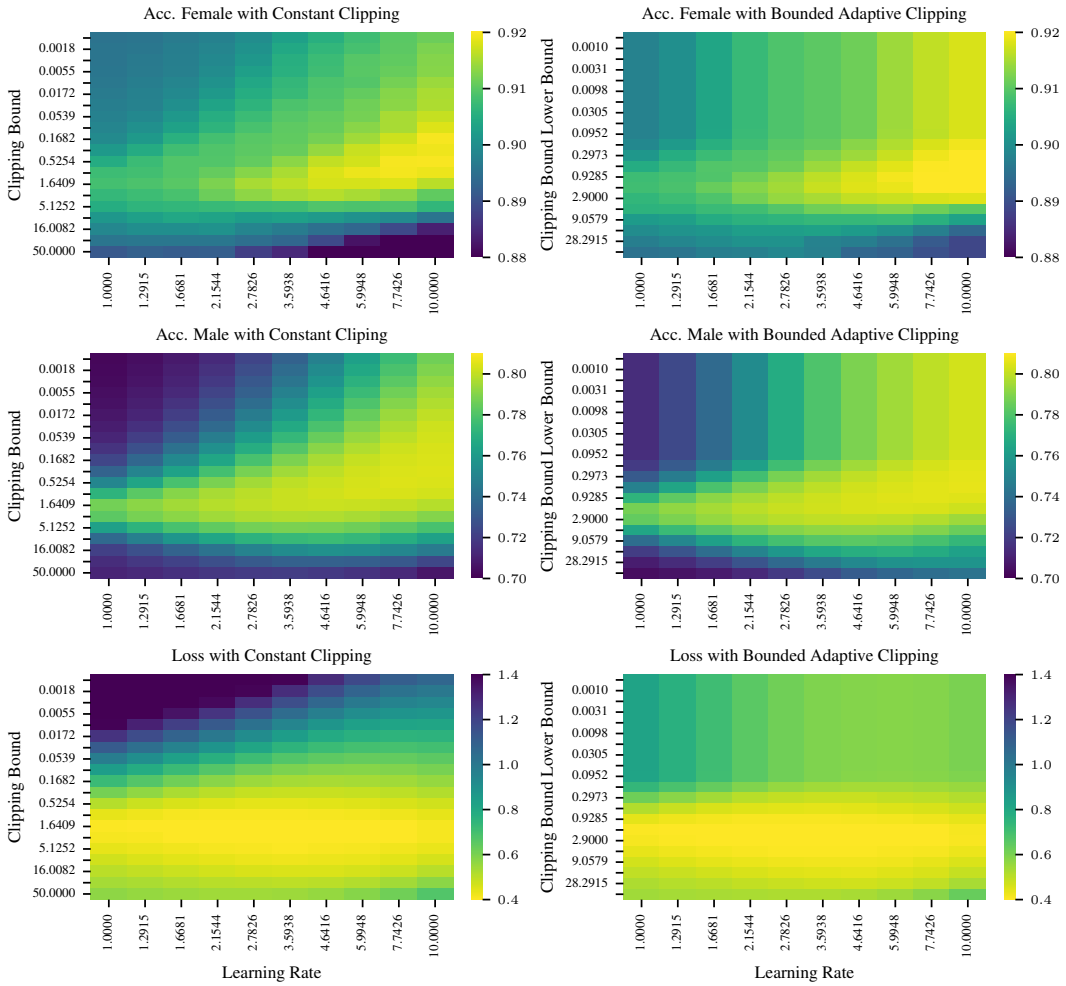


Figure 7: The heatmap of macro accuracy on Adult with $\varepsilon = 0.1$.

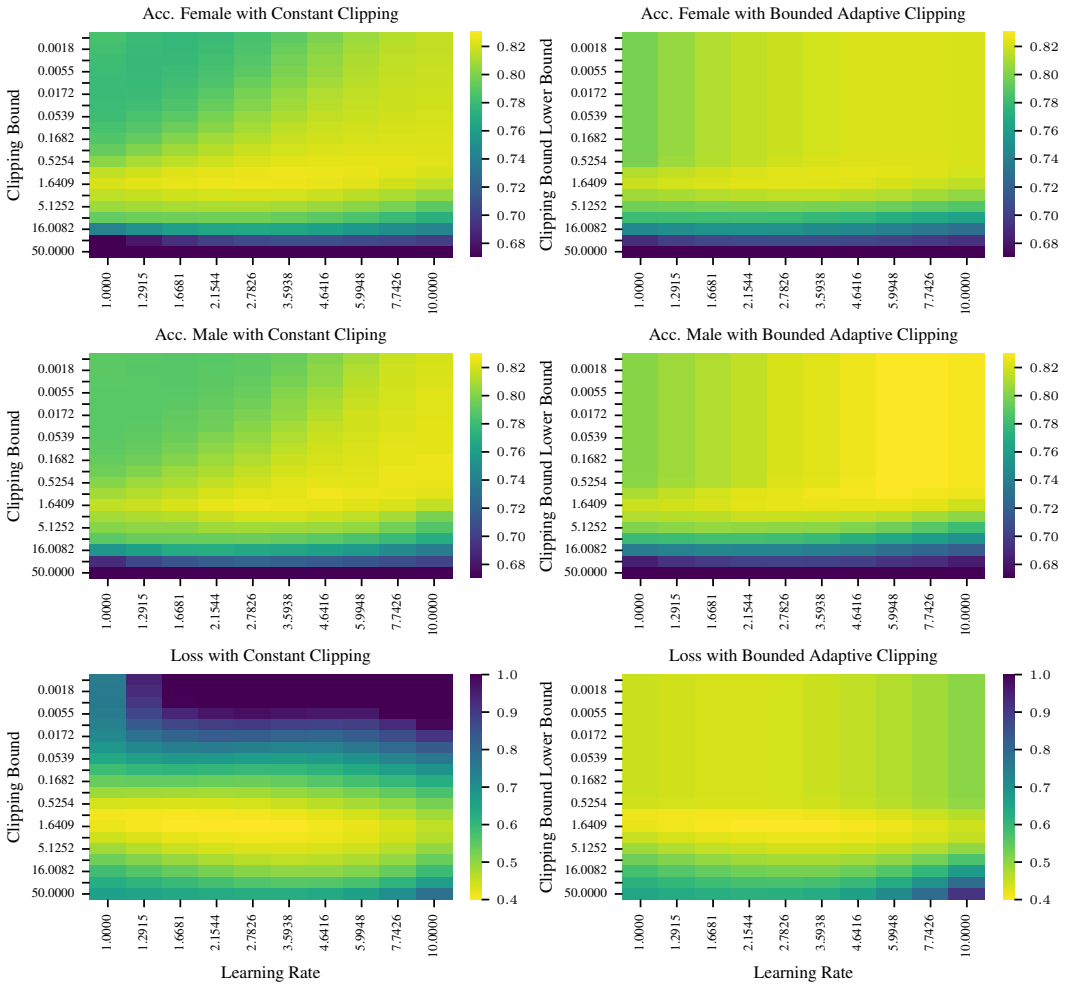


Figure 8: The heatmap of macro accuracy on Dutch with $\varepsilon = 0.1$.