# DP-AdamW: Investigating Decoupled Weight Decay and Bias Correction in Private Deep Learning

**Jay Chooi** [*1] **Kevin Cong** [*1] **Russell Li** [*1] **Lillian Sun** [*1]

## Abstract

As deep learning methods increasingly utilize sensitive data on a widespread scale, differential privacy (DP) offers formal guarantees to protect against information leakage during model training. A significant challenge remains in implementing DP optimizers that retain strong performance while preserving privacy. Recent advances introduced ever more efficient optimizers, with AdamW being a popular choice for training deep learning models because of strong empirical performance. We study *DP-AdamW* and introduce *DP-AdamW-BC*, a differentially private variant of the AdamW optimizer with DP bias correction for the second moment estimator. We start by showing theoretical results for privacy and convergence guarantees of DP-AdamW and DP-AdamW-BC. Then, we empirically analyze the behavior of both optimizers across multiple privacy budgets ($\epsilon = 1, 3, 7$). We find that DP-AdamW outperforms existing state-of-the-art differentially private optimizers like DP-SGD, DP-Adam, and DP-AdamBC, scoring over 15% higher on text classification, up to 5% higher on image classification, and consistently 1% higher on graph node classification. Moreover, we empirically show that incorporating bias correction in DP-AdamW (DP-AdamW-BC) consistently decreases accuracy, in contrast to the improvement of DP-AdamBC improvement over DP-Adam.

## 1. Introduction

In recent years, deep learning has achieved widespread adoption, with applications ranging from natural language processing to image generation (Brown et al., 2020; Ho et al., 2020). However, models trained on large and sensitive datasets have been shown to be vulnerable to privacy attacks on their training data, raising significant privacy conerns (Carlini et al., 2021; 2022; Balle et al., 2022). Differential Privacy (DP) provides a rigorous mathematical framework to address these concerns, enabling the development of algorithms with provable guarantees against leaking individual-specific information (Dwork et al., 2006). In stochastic optimization, this has been most commonly realized by the *DP-SGD* algorithm introduced by Abadi et al. (2016). The first application of DP in the deep learning model training pipeline, this approach adds Gaussian noise of variance $\sigma^2 C^2$ after clipping per-sample gradients to a radius $C$. While DP-SGD offers strong theoretical guarantees, its empirical performance, as with vanilla SGD, often lags behind the adaptive Adam family of optimizers that dominate non-private deep learning. Subsequent work therefore lifted Adam into the private setting; specific differentially private optimizers that were developed to potentially improve the privacy-utility tradeoff include DP-GD, DP-RMSprop, and DP-Adam (Zhou et al., 2020; Li et al., 2022a).

Motivated by this line of work, we study differentially private variants of the AdamW optimizer which has been empirically shown to achieve improved generalization performance over Adam, particularly enabling competitive performance with SGD on image classification tasks (Loshchilov & Hutter, 2017). To provide a differentially private implementation of AdamW, we introduce the *DP-AdamW* algorithm by decoupling the weight decay from the gradient update in the DP-Adam algorithm (Tang et al., 2023). Additionally, we introduce *DP-AdamW-BC*, a variant of DP-AdamW that corrects for the DP bias in the second moment estimator, which Tang et al. (2023) characterized and corrected for DP-Adam, resulting in DP-AdamBC.

We provide theoretical results on the privacy and convergence guarantees of the new optimizers and in particular show that DP-AdamW maintains similar privacy bounds and convergence guarantees to DP-AdamBC. We then compare the performance of our DP-AdamW and DP-AdamW-BC to previous optimizers, empirically finding that they consistently improve on the privacy-accuracy tradeoff compared

[*]Equal contribution, alphabetcally ordered. [1]Harvard University, Cambridge, MA, USA. Correspondence to: Lillian Sun <lilliansun@college.harvard.edu>, Kevin Cong <kcong@college.harvard.edu>.

*Accepted at the ICML 2025 Workshop on Collaborative and Federated Agentic Workflows (CFAgentic@ICML'25), Vancouver, Canada. July 19, 2025.* 

to their DP-Adam counterparts. In contrast to the findings of Tang et al. (2023), we show that incorporating bias correction in DP-AdamW-BC consistently decreases accuracy across diverse tasks. We demonstrate that DP-AdamW outperforms DP-SGD even in image classification, for which Adam is known to not generalize as well (Loshchilov & Hutter, 2017).

The remainder of our paper proceeds as follows. In Section 3, we present the DP-AdamW and DP-AdamW-BC algorithms and their theoretical guarantees. In Section 4, we detail our experiments and present the results of using the DP-AdamW and DP-AdamW-BC optimizers in comparison to DP-Adam and DP-AdamW. In Section 5, we discuss these results, and in Sections F and 6, we conclude and offer further avenues of work.

## 2. Motivation and Related Work

Differentially private optimization was first studied by (Abadi et al., 2016), who introduced Differentially Private Stochastic Gradient Descent (DP-SGD). DP-SGD achieved provable privacy guarantees via adding noise to the gradients during gradient descent. However, the additional noise led to significantly worse performance in comparison to non-private models. This led to the work of (Li et al., 2022b), who showed that finetuning pretrained models under DP-Adam achieved strong performance on par with non-private models. Moreover, they demonstrated a scaling law of DP-models by parameters and empirically invalidated the hypothesis that DP-ML suffers from dimension-dependent performance degradation. This set the stage for practical applications of DP on deep learning. More recently, (Tang et al., 2023) proposed an improvement to DP-Adam, namely DP-AdamBC, adding a correction term in DP-Adam to remove the bias in the second moment estimate arising from the added gaussian noise. DP-AdamBC achieved an improvement over DP-Adam across several different target privacy budgets on image, text and graph node classification tasks.

Motivated by these developments, our present work aims to improve the performance of DP-Adam and DP-AdamBC via adding weight decay, thereby yielding DP-AdamW and DP-AdamW-BC. We finally evaluate the performance of these optimizers using the same suite of tasks in their paper.

## 3. DP-AdamW: Algorithm and Theoretical Guarantees

In this section, we introduce the DP-AdamW algorithm and its variant, DP-AdamW-BC, and present theoretical results on their privacy and convergence guarantees

### 3.1. DP-AdamW: The Algorithm

We now introduce the DP-AdamW algorithm. The main point is to add Gaussian noise to the gradients in the AdamW algorithm. Given a batch of gradients $g_i$, we compute the average gradient $\overline{g}$. Define the noised gradient

$$\tilde{g} = \frac{1}{B}\sum_i \frac{g_i}{\max(1, \frac{\|g_i\|_2}{C})} + \frac{1}{B}\mathcal{N}(0, \sigma^2 C^2 \mathbf{I}).$$

We then use the noised gradient $\tilde{g}$ as the input gradient into the AdamW algorithm of (Loshchilov & Hutter, 2017).[1] The resulting DP-AdamW algorithm is given in Algorithm 1.

**Algorithm 1** DP-AdamW

**Require:** total steps $T$, learning-rate schedule $\{\eta_t\}_{t=1}^T$, clip norm $C$, weight-decay $\lambda$, hyper-parameters $\alpha$, $\beta_1$, $\beta_2$, noise multiplier $\sigma$, initial parameters $\theta_0$, numerical stability constant $\epsilon_0$

$m_0 \leftarrow 0$, $v_0 \leftarrow 0$
**for** $t \leftarrow 1$ **to** $T$ **do**
  $g_t \leftarrow \nabla f(\theta_{t-1})$
  $\tilde{g}_t \leftarrow \frac{1}{B}\sum_i \frac{g_i}{\max\left(1, \frac{\|g_i\|_2}{C}\right)} + \frac{1}{B}\mathcal{N}(0, \sigma^2 C^2 I)$
  $m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1)\,\tilde{g}_t$
  $v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2)\,\tilde{g}_t^2$
  $\hat{m}_t \leftarrow \frac{m_t}{1-\beta_1^t}$
  $\hat{v}_t \leftarrow \frac{v_t}{1-\beta_2^t}$
  $\theta_t \leftarrow \theta_{t-1} - \eta_t\left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t + \epsilon_0}} + \lambda\,\theta_{t-1}\right)$
**end for**

Moreover, (Tang et al., 2023) observe that within the Adam update, the estimate of the second moment is biased due to the use of a noised gradient. They offer a method of bias-correction, in particular by replacing the update term $\frac{\hat{m}_t}{\sqrt{\hat{v}_t}}$ with

$$\frac{\hat{m}_t}{\sqrt{\hat{v}_t - \left(\frac{\sigma C}{B}\right)^2}} = \frac{\hat{m}_t}{\sqrt{\hat{v}_t - \Phi}}.$$

We also consider this modification, which is given in Algorithm 2.

### 3.2. Privacy Guarantees

Both algorithms carry the same privacy guarantees as DP-SGD and DP-AdamBC. Formally, we have the following result, analogous to the privacy guarantees in Abadi et al. (2016) and (Tang et al., 2023).

[1] Alternatively, one can view this as adding decoupled weight decay into the DP-Adam algorithm of (Tang et al., 2023)

**Algorithm 2** DP-AdamW-BC

**Require:** total steps $T$, learning-rate schedule $\{\eta_t\}_{t=1}^T$, clip norm $C$, weight-decay $\lambda$, hyper-parameters $\alpha$, $\beta_1$, $\beta_2$, noise multiplier $\sigma$, initial parameters $\theta_0$, numerical stability constant $\gamma$

$m_0 \leftarrow 0$, $v_0 \leftarrow 0$

**for** $t \leftarrow 1$ **to** $T$ **do**

$g_t \leftarrow \nabla f(\theta_{t-1})$

$\tilde{g}_t \leftarrow \frac{1}{B}\sum_i \frac{g_i}{\max\left(1, \frac{\|g_i\|_2}{C}\right)} + \frac{1}{B}\mathcal{N}(0, \sigma^2 C^2 I)$

$m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1)\tilde{g}_t$

$v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2)\tilde{g}_t^2$

$\hat{m}_t \leftarrow \frac{m_t}{1-\beta_1^t}$

$\hat{v}_t \leftarrow \frac{v_t}{1-\beta_2^t}$

$\theta_t \leftarrow \theta_{t-1} - \eta_t\left(\frac{\hat{m}_t}{\sqrt{\max(\hat{v}_t - (\frac{\sigma C}{B})^2, \gamma)}} + \lambda\theta_{t-1}\right)$

**end for**

**Theorem 3.1** (cf. Proposition 1 of (Tang et al., 2023))**.** *Suppose that the DP-SGD optimizer* $DP-SGD(\theta, X, y, C, \sigma, B)$ *satisfies* $(\epsilon, \delta)$*-DP with privacy analysis* $\phi(T, \theta_i)$*. Then both* $DP-AdamW(\theta, X, y, C, \sigma, B)$ *and* $DP-AdamW-BC(\theta, X, y, C, \sigma, B)$ *satisfy* $(\epsilon, \delta)$*-DP with the same privacy analysis* $\phi(T, \theta_i)$*.*

The proof of this theorem is given in Appendix A; it follows the outline of Proposition 1 of (Tang et al., 2023).

### 3.3. Convergence Guarantees

We now show that the DP-AdamW and DP-AdamW-BC optimizers converge[2] under reasonable assumptions. Our results are based on and take similar form to those of (Défossez et al., 2022), who proved convergence guarantees for vanilla Adam, and (Tang et al., 2023), who proved analogous guarantees for DP-AdamBC. Our main contribution is to address the weight decay term; for this, the main idea is roughly to show that the parameters $\|\theta_t\|$ are bounded and apply standard inequalities.

Throughout this subsection, we will work under the following natural assumptions, found in (Tang et al., 2023). Let $F : \mathbb{R}^d \to \mathbb{R}$ denote the objective function and $f : \mathbb{R}^d \to \mathbb{R}$ denote a stochastic function with $\mathbb{E}(\nabla f(\theta)) = \nabla F(\theta)$.[3] Let $\|\cdot\|$ denote the $L_2$-norm. We then have the following.

[2]Here, convergence is in the sense of 'average gradient of the true objective function is small'. See Theorems 3.5 and 3.6 for details.

[3]That is, $F$ is the true objective function, while $f$ is the estimate of the objective function. For instance, if one is using a squared-error loss, then $f$ is the empirical MSE, while $F$ is the expected squared-error loss.

**Assumption 3.2.** $F$ is bounded below: $F(\theta) \geq F_*$ for all $\theta$.

**Assumption 3.3.** The gradients $\|\nabla f_t(\theta)\| \leq C_1 \leq C$ are uniformly almost surely bounded.

**Assumption 3.4.** The gradient of $F$ is $L$-Lipschitz continuous: $\|\nabla F(\theta) - \nabla F(\theta')\| \leq L\|\theta - \theta'\|$.

*Remark.* These are the same assumptions as found in (Défossez et al., 2022) and (Tang et al., 2023), but we will provide some further intuition. Assumption 3.2 says that $F$ *can* be optimized; this is clearly necessary and moreover holds in practice (for instance, $F_* = 0$ holds for any squared-error loss). Assumption 3.3 assumes that the gradients do not blow up. This assumption is necessary for theoretical results, but does not always hold in practice; exploding gradients often occur in empirical studies. Note that we additionally assume $C_1 \leq C$ for simplicity, since bounded gradients implies that gradient clipping is not necessary. Lastly, Assumption 3.4 is again necessary to attain effective theoretical bounds. It is generally true if $F$ is a reasonably 'smooth' function.

We divide our results into two settings. We first have the following guarantees on DP-AdamW and DP-AdamW-BC without momentum, i.e. the regime of $\beta_1 = 0$.

**Theorem 3.5.** *Under Assumptions 3.2, 3.3, and 3.4, suppose that* $\beta_1 = 0$, $0 < \beta_2 < 1$, $\alpha \in (0, 1)$*, and the learning rate follows* $\eta_t = \eta\sqrt{\frac{1-\beta_2^t}{1-\beta_2}}$*. Let* $\Phi = \left(\frac{\sigma C}{B}\right)^2$ *denote the bias correction term and let* $\mu^* = \frac{\beta_2(1-\beta_2^T)}{1-\beta_2}[(\Phi - \frac{2\Phi}{\pi}) + \left(C + \sqrt{\frac{2\Phi}{\pi}}\right)^2]$*,* $\nu^* = 2\beta_2^2\Phi\sqrt{\frac{1-\beta_2^{2T}}{1-\beta_2^2}}$*, and* $b^* = 4\beta_2\Phi$ *be constants. Then there exists a constant* $c(\lambda) = c(\beta_1, \beta_2, \lambda, \eta, C_1, L, \theta_0, \epsilon_0, \Phi)$ *such that whenever*

$$\delta_0 \geq \begin{cases} \mu^* + \sqrt{\ln(1/\frac{\alpha}{2T})(2(\nu^*)^2)} & 0 \leq \delta_0 \leq \frac{(\nu^*)^2}{b^*} \\ \mu^* + \ln(1/\frac{\alpha}{2T})2b^* & \delta_0 \geq \frac{(\nu^*)^2}{b^*}, \end{cases}$$

*we have with probability at least* $1 - \alpha$ *that for DP-AdamW,*

$$\frac{1}{T}\sum_0^{T-1}\mathbb{E}\|\nabla F(\theta_{t-1})\|^2 \leq \frac{2(\delta_0 + C_1)(F(\theta_0) - F_*)}{\eta T} + \left(\frac{4d(C^2+\Phi)}{\sqrt{1-\beta_2}} + \frac{\eta d L\sqrt{C^2+\Phi}(1+\lambda)}{1-\beta_2}\right)\cdot\frac{R}{T} + \frac{1}{2T}\left(C_1^2 + \|\theta_0\|^2 + c(\lambda)\max_t(\eta_t + \eta_t^2)R\right)\left(\lambda\sum_{t=1}^T\eta_t + \frac{L}{2}(\lambda+\lambda^2)\sum_{t=1}^T\eta_t^2\right),$$

*where* $R = d\left(\ln\left(1 + \frac{C^2+\Phi}{(1-\beta_2)\epsilon_0}\right) - T\ln\beta_2\right)$*,*

*and for DP-AdamW-BC, we have* $\frac{1}{T}\sum_{i=0}^{T-1}\mathbb{E}[\|\nabla F(\theta_i)\|^2]$

$$\leq \frac{2\sqrt{(\delta_0 + C_1)^2 - \Phi}(F(\theta_0) - F_*)}{\eta T}$$

$$+\left(\frac{4dC^2}{\sqrt{1-\beta_2}}+\frac{\eta dL(1+\lambda)C}{1-\beta_2}\right)\cdot\frac{R_{BC}}{T}$$
$$+\frac{1}{2T}\left(C_1^2+\|\theta_0\|^2+c(\lambda)\max_t(\eta_t+\eta_t^2)R_{BC}\right)$$
$$\left(\lambda\sum_{t=1}^T\eta_t+\frac{L}{2}(\lambda+\lambda^2)\sum_{t=1}^T\eta_t^2\right),$$

*where* $R_{BC}=d\left(\ln\left|1-\frac{C^2+\Phi}{(1-\beta_2)\Phi}\right|-T\ln\beta_2\right)$.

*Remark.* We first make a few notes about the statement of Theorem 3.5 below.

First, the definition of $\delta_0$ arises from an application of concentration bounds used in the proof; in particular, it arises in Corollary 1 of (Tang et al., 2023). Essentially, the constraint on $\delta_0$ is that either $\delta_0\geq\max(\mu^*+\ln(1/\frac{\alpha}{2T})2b^*,\frac{(\nu^*)^2}{b^*})$ or $\delta_0\in[\mu^*+\sqrt{\ln(1/\frac{\alpha}{2T})(2(\nu^*)^2)},\frac{(\nu^*)^2}{b^*}]$.

Second, the right hand side is a fairly ugly expression, but we will derive an asymptotic bound when $T\to\infty$. Supposing that $\eta=T^{-a}$ and $\beta_2=1-T^{-b}$, note that $R=\mathcal{O}(\ln T+T^{1-b})=\mathcal{O}(T^{1-b}\ln T)$. Moreover, since $b^*=\mathcal{O}(1)$, $\nu^*=\mathcal{O}(1)$ if $b<1$ and $\nu^*=\mathcal{O}(\sqrt{T})$ otherwise, and $\mu^*=\mathcal{O}(1)$ if $b<1$ and $\mu^*=\mathcal{O}(T)$ otherwise, we find that $\delta_0=\mathcal{O}(\ln T)$ if $b<1$ and $\delta_0=\mathcal{O}(T)$ otherwise. Taking $b<1$, the first inequality in the theorem becomes

$$\frac{1}{T}\sum_{i=0}^{T-1}\mathbb{E}[\|\nabla F(\theta_i)\|^2]\leq\ln T\cdot\mathcal{O}(T^{a-1}$$
$$+(T^{\frac{b}{2}}+T^{b-a})T^{-b}+T^{b/2-a}(T^{\frac{3}{2}-a}+T^{2-2a})T^{-b}).$$

The right hand side is optimized for the choice $b\to1^-$, $a=\frac{2}{3}$, under which we obtain

$$\frac{1}{T}\sum_{i=0}^{T-1}\mathbb{E}[\|\nabla F(\theta_i)\|^2]=\mathcal{O}(T^{-\frac{1}{3}+\zeta}\ln T)\overset{T\to\infty}{\longrightarrow}0,$$

where $\zeta$ is arbitrarily small. An analogous result[4] can be attained for DP-AdamW-BC. Hence, one can interpret these results as saying that under an 'optimal' learning rate regime, the average gradient of the objective function throughout the course of DP-AdamW and DP-AdamW-BC converges to 0 at an inverse-polynomial rate as the time horizon increases. By choosing $\alpha\approx0$, this thus implies that DP-AdamW and DP-AdamW-BC converges with high probability to a local minimum under the given conditions and assumptions.

Lastly, we now make a few notes about the proof of Theorem 3.5. For a broad sketch, we follow the rough outline of (Tang et al., 2023); this suffices to account for all terms other than the weight decay term. To address this term, we bound the expected magnitude of the parameters, $\mathbb{E}\|\theta_i\|^2$, using intermediary inequalities from (Tang et al., 2023). This method can be imitated to show convergence in certain cases of non-private AdamW as well, which may be of independent interest.[5]

We now turn to the general case where $\beta_1$ is not necessarily equal to 0. In this case, we have the following result, attained via techniques similar to those used to prove Theorem 3.5.

**Theorem 3.6.** *Under Assumptions 3.2, 3.3, and 3.4, suppose that* $0<\beta_1<1$, $0<\beta_2<1$, $\alpha\in(0,1)$, *and the learning rate follows* $\eta_t=\eta(1-\beta_1)\sqrt{\frac{1-\beta_2^t}{1-\beta_2}}$. *Let* $\Phi=\left(\frac{\sigma C}{B}\right)^2$ *denote the bias correction term and let* $\mu^*=\frac{\beta_2(1-\beta_2^T)}{1-\beta_2}[(\Phi-\frac{2\Phi}{\pi})+\left(C+\sqrt{\frac{2\Phi}{\pi}}\right)^2]$, $\nu^*=2\beta_2^2\Phi\sqrt{\frac{1-\beta_2^{2T}}{1-\beta_2^2}}$, *and* $b^*=4\beta_2\Phi$ *be constants. Then there exists a constant* $c(\lambda)=c(\beta_1,\beta_2,\lambda,\eta,C_1,L,\theta_0,\epsilon_0,\Phi)$ *such that whenever* $\tilde{T}=T-\frac{\beta_1}{1-\beta_1}>0$ *and*

$$\delta_0\geq\begin{cases}\mu^*+\sqrt{\ln(1/\frac{\alpha}{2T})(2(\nu^*)^2)} & 0\leq\delta_0\leq\frac{(\nu^*)^2}{b^*}\\ \mu^*+\ln(1/\frac{\alpha}{2T})2b^* & \delta_0\geq\frac{(\nu^*)^2}{b^*},\end{cases}$$

*we have with probability at least* $1-\alpha$ *that for DP-AdamW,*

$$\mathbb{E}[\|\nabla F(\theta_\tau)\|^2]\leq\frac{2(\delta_0+C_1)(F(\theta_0)-F_*)}{\eta\tilde{T}}+E\cdot R$$
$$+\frac{1}{2T}\left(C_1^2+\|\theta_0\|^2+c(\lambda)\max_t(\eta_t+\eta_t^2)R\right)$$
$$\left(\lambda\sum_{t=1}^T\eta_t+\frac{L}{2}(\lambda+\lambda^2)\sum_{t=1}^T\eta_t^2\right),$$

*where*

$$E=\frac{\eta dL(1-\beta_1)\delta_0}{(1-\beta_1/\beta_2)(1-\beta_2)}+\frac{2\eta^2dL^2\beta_1}{(1-\beta_1/\beta_2)(1-\beta_2)^{3/2}}$$
$$+\frac{12d\delta_0^2\sqrt{1-\beta_1}}{(1-\beta_1/\beta_2)^{3/2}\sqrt{1-\beta_2}}$$
$$\text{and } R=d\left(\ln\left(1+\frac{\delta_0^2}{\epsilon_0(1-\beta_2)}\right)-T\log\beta_2\right),$$

*and for DP-AdamW-BC,*

$$\mathbb{E}[\|\nabla F(\theta_\tau)\|^2]\leq$$
$$\frac{2\sqrt{(\delta_0+C_1)^2-\Phi}(F(\theta_0)-F_*)}{\eta\tilde{T}}+E_{BC}\cdot R_{BC}$$
$$+\frac{1}{2T}\left(C_1^2+\|\theta_0\|^2+c(\lambda)\max_t(\eta_t+\eta_t^2)R_{BC}\right)$$
$$\left(\lambda\sum_{t=1}^T\eta_t+\frac{L}{2}(\lambda+\lambda^2)\sum_{t=1}^T\eta_t^2\right),$$

*where*

$$E_{BC}=\frac{\eta dL(1-\beta_1)\sqrt{\delta_0^2-\Phi}}{(1-\beta_1/\beta_2)(1-\beta_2)}+\frac{2\eta^2dL^2\beta_1}{(1-\beta_1/\beta_2)(1-\beta_2)^{3/2}}$$
$$+\frac{12d(\delta_0^2-\Phi)\sqrt{1-\beta_1}}{(1-\beta_1/\beta_2)^{3/2}\sqrt{1-\beta_2}}$$
$$\text{and } R_{BC}=d\left(\ln\left|1-\frac{\delta_0^2}{\Phi(1-\beta_2)}\right|-T\log\beta_2\right).$$

[4] In fact, with exactly the same bound.

[5] To our knowledge, there is no stated such result, but one can attain such a result via the same proof method.

*In both above inequalities, the left-hand-side expectations are with respect to sampling $\tau$ from the distribution $\mathbb{P}(\tau = t) \propto 1 - \beta_1^{T-t}$.*

*Remark.* This result is not an exact generalization of Theorem 3.5, but is fairly similar while holding in a more general setting. Moreover, the main steps of the proof are analogous. Lastly, we note that one can attain similar asymptotics for the bounds in this result as in Theorem 3.5.

Proofs of Theorem 3.5 and 3.6 are given in Appendix B.

## 4. Experiments

We evaluate the performance of using the DP-AdamW and DP-AdamW-BC optimizers on image, text, and graph node classification tasks. We then compare our results to those of (Tang et al., 2023). Experimental details are included in Appendix C.

### 4.1. Image Classification

We first replicated the experimental setup of Tang et al. (2023) for DP-Adam and DP-AdamBC on the same CNN architecture and CIFAR-10 task. Our DP-Adam and DP-Adam implementations achieve comparable test accuracies to those reported, validating our baseline implementations. This allows for a direct comparison between the performance of the original DP-Adam variants and our proposed DP-AdamW approaches under corresponding conditions.

We evaluate DP-AdamW and DP-AdamW-BC across the tuned hyperparameters for each $\epsilon \in \{1, 3, 7\}$. Test accuracies on CIFAR-10 are reported in Table 1, along with benchmark results reported in Tang et al. (2023). DP-AdamW outperforms DP-Adam in all privacy budgets. DP-AdamW-BC outperforms DP-AdamBC across privacy budgets except for the large $\epsilon = 7$. The most significant performance increase is in the $\epsilon = 3$ setting, where both DP-AdamW and DP-AdamW-BC outperform DP-Adam and DP-AdamBC by up to 5%.

DP-AdamW achieves state-of-the-art accuracy among the evaluated DP-SGD and DP-Adam variants for CIFAR-10 image classification for tight to moderate privacy constraints ($\epsilon$=1,3). This result is unexpected, as image classification is a known setting where Adam optimizers' performance fall short in comparison to SGD (Loshchilov & Hutter, 2017). Our results confirm the effectiveness of decoupling weight decay, even in the differential privacy setting.

We observe that DP-AdamW consistently outperformed DP-AdamW-BC across all tested privacy budgets, achieving statistically significantly higher final test accuracy. In contrast to findings for DP-AdamBC from Tang et al. (2023), DP-AdamW-BC accuracies show that applying DP bias correction for the second moment estimator degrades test accuracy compared to the uncorrected DP-AdamW. While DP-AdamW-BC offers a theoretically appealing bias correction, in our CIFAR-10 experiments, the DP-AdamW results remain stronger across all privacy budgets. This suggests the decoupling weight decay in the AdamW optimizer induces a greater performance boost, and potentially conflicts with, the effect from the bias correction.

*Table 1.* Image Classification: performance comparison of DP optimizers on CIFAR-10 dataset

| | $\epsilon \approx 1$ | $\epsilon \approx 3$ | $\epsilon \approx 7$ |
|---|---|---|---|
| DP-SGD | 52.37 (0.50) | 57.30 (0.76) | **65.30 (0.33)** |
| DP-Adam | 51.89 (0.69) | 54.08 (0.41) | 62.24 (0.10) |
| DP-AdamBC | 49.75 (0.56) | 54.27 (0.23) | 63.43 (0.43) |
| DP-AdamW | **52.59 (0.44)** | **59.26 (0.30)** | 63.25 (0.53) |
| DP-AdamW-BC | 51.43 (0.52) | 58.16 (0.72) | 62.01 (0.29) |

### 4.2. Text Classification

Table 2 reports the mean test accuracy and standard deviation over five random seeds for each optimizer across privacy budgets. Two trends emerge. First, decoupled weight decay yields a large performance improvement. Switching from DP-Adam to DP-AdamW raises accuracy by over 15 percentage points at every $\epsilon$ which aligns with prior findings in the literature that weight-decay coupling persists and is amplified in the private setting. The accuracies achieved by DP-AdamW without bias correction outperform all previously published DP optimizer fine tuning results on QNLI to our knowledge. Second, adding bias correction hurts on textual tasks. Inspecting all experiments, we observe that DP-AdamW-BC trails DP-AdamW by 2-3 percentage points and hypothesize that the added variance floor $\gamma$ interacts poorly with the small effective batch size and the already low intrinsic gradient noise of the BERT fine-tuning task. For all optimizers tested, test set performance improves as $\epsilon$ increases, which makes sense based on our knowledge of differential privacy because less noise must be added to satisfy looser privacy budgets, resulting in the gradient update possessing more signal comparatively.

*Table 2.* Text Classification: Performance comparison of DP optimizers on QNLI dataset

| | $\epsilon \approx 1$ | $\epsilon \approx 3$ | $\epsilon \approx 7$ |
|---|---|---|---|
| DP-SGD | 57.10 (1.59) | 58.85 (1.20) | 58.29 (0.92) |
| DP-Adam | 58.00 (2.05) | 60.72 (1.12) | 61.23 (1.30) |
| DP-AdamBC | 58.32 (1.90) | 61.42 (0.99) | 62.83 (1.60) |
| DP-AdamW | **77.78 (0.13)** | **79.26 (0.18)** | **80.01 (0.19)** |
| DP-AdamW-BC | 75.56 (0.17) | 76.65 (0.22) | 77.29 (0.08) |

### 4.3. Graph Node Classification

We found that DP-AdamW and DP-AdamW-BC outperform DP-SGD and DP-Adam across all $\epsilon$. Furthermore, both DP-AdamW and DP-AdamW-BC outperform DP-AdamBC for

$\epsilon \approx 12$, while DP-AdamW outperforms DP-AdamBC for $\epsilon \approx 6$. This shows that DP-AdamW and DP-AdamW-BC, like their non-private counterparts, perform better than DP-SGD and DP-Adam, while exceeding the performance of DP-AdamBC under looser privacy budgets.

*Table 3.* Graph Node Classification: Performance comparison of DP optimizers on obgn-arxiv node classification

| | $\epsilon \approx 3$ | $\epsilon \approx 6$ | $\epsilon \approx 12$ |
|---|---|---|---|
| DP-SGD | 45.35 (1.38) | 49.12 (1.90) | 54.20 (0.62) |
| DP-Adam | 46.55 (0.54) | 51.98 (0.48) | 54.02 (0.18) |
| DP-AdamBC | **50.51** (0.56) | 53.40 (0.28) | 53.81 (0.34) |
| DP-AdamW | 50.20 (0.75) | **53.41** (0.37) | **54.78**(0.24) |
| DP-AdamW-BC | 48.99 (0.67) | 52.89 (0.40) | 54.53 (0.43) |

## 5. Discussion

Our findings reveal three key insights into differentially private optimization.

*First, DP-AdamW consistently outperforms DP-Adam* across all tested privacy budgets and tasks, particularly in the moderate privacy setting ($\epsilon = 3$). This suggests that AdamW's core benefit—decoupling weight decay from adaptive gradient updates—translates effectively to the differentially private (DP) setting. By applying weight decay directly to parameters, DP-AdamW likely achieves more stable and effective regularization compared to DP-Adam, mitigating potential negative interactions between DP noise, the adaptive moment estimates, and the regularization term. Decoupled weight decay offers a twofold advantage: regularization strength becomes independent of instantaneous noise levels, and a mild pre-conditioning effect reduces the likelihood of gradients hitting the clipping threshold, thus preserving more signal.

*Second, DP-AdamW demonstrates superior performance over DP-SGD on CIFAR-10 image classification under tighter privacy constraints* ($\epsilon = 1, 3$), a noteworthy result given that non-private Adam variants often underperform SGD on such tasks (Loshchilov & Hutter, 2017). This strong performance indicates that the combination of adaptive learning rates and decoupled weight decay is particularly advantageous when dealing with the noise inherent in DP training. DP training introduces gradient clipping and additive noise; AdamW's coordinate-wise adaptive updates help counteract these distortions by allowing weights in low-variance directions to move appropriately without letting those in high-variance directions explode. This corrective effect is more pronounced with higher noise levels, explaining the significant gains at tighter privacy budgets.

*Third, and most surprisingly, bias correction in DP-AdamW-BC consistently degrades test accuracy compared to DP-AdamW* across all tasks and privacy budgets. This contrasts sharply with findings for Tang et al. (2023), where DP-AdamBC typically improves upon DP-Adam. While bias correction aims to counteract DP noise effects in the second moment estimate, our results suggest a negative interaction with AdamW's decoupled weight decay. The theoretical benefit of bias correction might be less relevant or even detrimental when decoupled weight decay is the primary regularization force. Specifically, the bias correction term can cause the denominator in the update rule (e.g., $\hat{v}_t - \Phi$) to become very small (clamping to $\gamma$), effectively freezing the adaptive schedule. With decoupled decay, this leads to large parameter steps that are not offset within Adam's update, potentially adding optimization noise and reducing accuracy.

Overall, our findings underscore *DP-AdamW's promise as a robust optimizer for differentially private deep learning*, highlighting the benefits of decoupled weight decay under privacy constraints. The unexpected underperformance of DP-AdamW-BC warrants further investigation into the complex interplay between DP noise, adaptive moment estimation, bias correction, and regularization strategies.

## 6. Conclusion

This paper analyzes DP-AdamW and DP-AdamW-BC, differentially private versions of the AdamW optimizer, establishing their theoretical privacy and convergence guarantees which align with existing DP-SGD and DP-Adam literature. Empirical evaluations on image, text, and graph node classification tasks reveal that DP-AdamW generally outperforms standard baselines like DP-SGD and DP-Adam, especially under tighter privacy constraints, indicating the benefits of decoupled weight decay persist in the DP setting. Counterintuitively, the bias-corrected variant, DP-AdamW-BC, consistently led to worse performance than DP-AdamW across experiments, challenging the assumption that such bias correction is universally beneficial when combined with decoupled weight decay. Overall, the work positions DP-AdamW as a promising and effective optimizer for privacy-preserving deep learning, offering improved utility over existing methods on diverse classification tasks.

## Acknowledgements

The authors thank Salil Vadhan for his teachings on differential privacy, which inspired this paper. In addition, the authors thank Salil Vadhan and Zachary Ratliff for their thoughtful discussions and helpful suggestions for this manuscript.

## Impact Statement

This research investigates DP-AdamW and DP-AdamW-BC, more effective optimizers for differentially private deep learning, significantly advancing the protection of sensitive data. The key positive impact is enhanced data privacy, enabling AI in critical sectors like healthcare and finance with stronger safeguards. DP-AdamW's improved performance over existing private optimizers can lead to more useful and accurate AI models that respect user privacy, encouraging broader adoption of privacy-preserving practices and unlocking AI applications on sensitive datasets. However, challenges include the persistent privacy-utility trade-off and the risk of misinterpreting privacy guarantees. Ultimately, this work provides better tools for private AI, contributing to more trustworthy and responsible systems, though continued research and clear communication are vital for realizing its full positive potential.

## References


Abadi, M., Chu, A., Goodfellow, I., 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*, pp. 308–318, 2016.

Balle, B., Cherubin, G., and Hayes, J. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1138–1156. IEEE, 2022.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. A large annotated corpus for learning natural language inference. In Màrquez, L., Callison-Burch, C., and Su, J. (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075/.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners, July 2020. URL http://arxiv.org/abs/2005.14165. arXiv:2005.14165 [cs].

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pp. 1897–1914. IEEE, 2022.

Daigavane, A., Madan, G., Sinha, A., Thakurta, A. G., Aggarwal, G., and Jain, P. Node-Level Differentially Private Graph Neural Networks, August 2022. URL http://arxiv.org/abs/2111.15521. arXiv:2111.15521 [cs].

Dwork, C., McSherry, F., Nissim, K., and Smith, A. 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 3*, pp. 265–284. Springer, 2006.

Défossez, A., Bottou, L., Bach, F., and Usunier, N. A simple convergence proof of adam and adagrad, 2022. URL https://arxiv.org/abs/2003.02395.

Ho, J., Jain, A., and Abbeel, P. Denoising Diffusion Probabilistic Models, December 2020. URL http://arxiv.org/abs/2006.11239. arXiv:2006.11239 [cs].

Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. Open Graph Benchmark: Datasets for Machine Learning on Graphs, February 2021. URL http://arxiv.org/abs/2005.00687. arXiv:2005.00687 [cs].

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Li, T., Zaheer, M., Reddi, S., and Smith, V. Private adaptive optimization with side information. In *International Conference on Machine Learning*, pp. 13086–13105. PMLR, 2022a.

Li, X., Tramèr, F., Liang, P., and Hashimoto, T. Large Language Models Can Be Strong Differentially Private Learners, November 2022b. URL http://arxiv.org/abs/2110.05679. arXiv:2110.05679 [cs].

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Tang, Q., Shpilevskiy, F., and Lécuyer, M. DP-AdamBC: Your DP-Adam Is Actually DP-SGD (Unless You Apply Bias Correction), December 2023. URL http://arxiv.org/abs/2312.14334. arXiv:2312.14334 [cs].

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. GLUE: A Multi-Task Benchmark and Analysis

Platform for Natural Language Understanding, February 2019. URL http://arxiv.org/abs/1804.07461. arXiv:1804.07461 [cs].

Zhou, Y., Chen, X., Hong, M., Wu, Z. S., and Banerjee, A. Private stochastic non-convex optimization: Adaptive algorithms and tighter generalization bounds. *arXiv preprint arXiv:2006.13501*, 2020.

## A. Privacy Guarantees: Proofs

We provide the proof of Theorem 3.1 below.

*Proof of Theorem 3.1.* We exactly follow the proof of the Proposition 5 in (Tang et al., 2023). Let $PrivitizeGradient(\theta, X, y, C, \sigma, B)$ denote the function whose output is the noised gradient $\tilde{g} = \frac{1}{B}\left(\sum_i \text{clip}(g(x_i) + \mathcal{N}(0, \sigma^2 C^2)\right)$. Critically, this is the same function as used in DP-SGD and DP-AdamBC. Suppose that these algorithms are differentially private with privacy analysis given by $\phi(T, \theta_i) = (\epsilon, \delta)$. Then note that the update of both DP-AdamW and DP-AdamW-BC do not involve any private information beyond the noised gradients. By the adaptive postprocessing property of DP, it follows that DP-AdamW and DP-AdamW-BC are also $(\epsilon, \delta)$-DP. Thus the privacy analysis of DP-AdamW and DP-AdamW-BC are also given by $\phi(T, \theta_i)$, completing the proof. □

*Remark.* Recall from Theorem 2 of (Tang et al., 2023) that DP-SGD has privacy guarantee $(\epsilon, \delta)$ whenever the noise satisfies $\sigma \geq \frac{c_2 q \sqrt{T \log \frac{1}{\delta}}}{\epsilon}$. This thus provides a concrete privacy guarantee for DP-AdamW and DP-AdamW-BC as well.

## B. Convergence Guarantees: Proofs

We now provide proofs for the convergence results presented in the main body of the paper.

### B.1. Proof of Theorem 3.5

We prove the convergence result given in Theorem 3.5 below.

*Proof of Theorem 3.5.* We will only prove the result for DP-AdamW; the exact same proof technique will suffice for DP-AdamW-BC. As in (Tang et al., 2023), define the notation $u_t = \frac{\nabla_i f_t^p(\theta_{t-1})}{\sqrt{v_t^p}+\epsilon_0}$ denote the main 'update term' in DP-Adam; that is, $\nabla_i f_t^p(\theta_{t-1}) = \tilde{m}_t$ and $v_t^p = \tilde{v}_t$. We first prove a bound on the expected magnitude of the parameters, then proceed as in (Tang et al., 2023).

We first note that Lemma 3 and Corollary 1 in Appendix F of (Tang et al., 2023) hold, with the proof remaining identical. Moreover, identical arguments as in the proof of Proposition 6 of (Tang et al., 2023) yield the inequalities[6]

$$\sum_{t=0}^{T-1} \mathbb{E}[\|u_t\|^2] \leq d\left(\ln\left(1+\frac{C^2+\Phi}{(1-\beta_2)\epsilon_0}\right) - T\ln\beta_2\right) \tag{1}$$

and moreover

$$\sum_{t=1}^{T} \eta_t \mathbb{E}\left[\nabla F(\theta_{t-1})^\top \frac{\nabla f_t^p(\theta_{t-1})}{\sqrt{v_t^p}+\epsilon_0}\right] \geq -\frac{\eta}{2(\delta_0+C_1)} \sum_{t=1}^{T} \mathbb{E}[\|\nabla F(\theta_{t-1})\|^2] \tag{2}$$

$$+ \frac{2\eta\sqrt{C^2+\Phi}}{\sqrt{1-\beta_2}} \sum_{t=0}^{T-1} \mathbb{E}[\|u_t\|]^2. \tag{3}$$

In particular, the first statement follows from Lemma 5.2 in (Défossez et al., 2022), and the second by using Lemmas 5.1 and 5.2 of the aforementioned work.

We now proceed as follows. First, we bound the magnitude of the parameters $\theta_t$ in expectation. Let

$$R = d\left(\ln\left(1+\frac{C^2+\Phi}{(1-\beta_2)\epsilon_0}\right) - T\ln\beta_2\right).$$

**Claim:** $\mathbb{E}\|\theta_t\|^2 \leq \|\theta_0\|^2 + c(\lambda)R$, where $c(\lambda)$ is a constant depending on $\lambda$ and all other constant parameters. [7]

**Proof:** The main idea is to combine the recursive definition of $\theta_t$ with the AM-GM inequality and 2, Let $x_t = \mathbb{E}\|\theta_t\|^2$ and write

$$\begin{aligned}
x_t = \mathbb{E}\|\theta_t\|^2 &= (1-\eta_t\lambda)^2\mathbb{E}\|\theta_{t-1}\|^2 + \eta_t^2\mathbb{E}\|u_t\|^2 + 2\eta_t(1-\eta_t\lambda)\mathbb{E}[\langle\theta_{t-1}, u_t\rangle] \\
&\leq (1-\eta_t\lambda)^2 x_{t-1} + \eta_t^2\mathbb{E}\|u_t\|^2 + 2\eta_t(1-\eta_t\lambda)\left(\frac{\lambda}{2}\mathbb{E}\|\theta_{t-1}\|^2 + \frac{1}{2\lambda}\mathbb{E}\|u_t\|^2\right) \\
&= \left[\eta_t^2 + \frac{\eta_t(1-\eta_t\lambda)}{\lambda}\right]\mathbb{E}\|u_t\|^2 + (1-\eta_t\lambda)\mathbb{E}\|\theta_{t-1}\|^2 \\
&\leq \left(\eta_t^2 + \frac{\eta_t}{\lambda}\right)\mathbb{E}\|u_t\|^2 + x_{t-1},
\end{aligned}$$

where we have used the fact that $\langle x, y\rangle = \sum x_i y_i \leq \lambda\sum x_i^2 + \frac{1}{\lambda}\sum y_i^2$. It follows by summing the above inequality for all $t$, telescoping, and applying inequality 1 that

$$\begin{aligned}
x_t &\leq \|\theta_0\|^2 + (1+\frac{1}{\lambda})\left(\max_t(\eta_t+\eta_t^2)d\left(\ln\left(1+\frac{C^2+\Phi}{(1-\beta_2)\epsilon_0}\right) - T\ln\beta_2\right)\right) \\
&= \|\theta_0\|^2 + c(\lambda)\max_t(\eta_t+\eta_t^2)R.
\end{aligned}$$

[6] These results are directly shown and/or implicit in the referenced proof; the proof is immediately adaptable.

[7] Here $c(\lambda)$ notation encapsulates all constant parameters, as well as $\theta_0$.

This completes the proof of the claim.

Now we return to the proof of the main theorem. We follow the rough outline used by (Tang et al., 2023). The main idea is to use the Lipschitzness of the gradient to bound the difference $F(\theta_t) - F(\theta_{t-1})$. One can then telescope this difference and combine the claim with inequalities 1 and 2 and some algebraic manipulations.

Let us write $\theta_t = \theta_{t-1} - \eta_t(u_t + \lambda_t\theta_{t-1})$. Using the Lipschitzness assumption of the gradient, we have

$$F(\theta_t) \leq F(\theta_{t-1}) - \eta_t\langle \nabla F(\theta_{t-1}), u_t\rangle - \eta_t\lambda\langle \nabla F(\theta_{t-1}), \theta_{t-1}\rangle + \frac{\eta_t^2 L}{2}\|u_t\|^2 \tag{4}$$

$$+ \frac{\eta_t^2 L}{2}\lambda^2\|\theta_{t-1}\|^2 + \eta_t^2 L\lambda\langle u_t, \theta_{t-1}\rangle. \tag{5}$$

Taking expectations, we find by the above claim that $\mathbb{E}\|\theta_{t-1}\|^2 = \|\theta_0\|^2 + c(\lambda)\max_t(\eta_t + \eta_t^2)R$ and hence

$$\mathbb{E}|\langle \nabla F(\theta_{t-1}), \theta_{t-1}\rangle| \leq \frac{1}{2}(\mathbb{E}\|\nabla F(\theta_{t-1})\|^2 + \mathbb{E}\|\theta_{t-1}\|^2) \leq \frac{1}{2}(C_1^2 + \|\theta_0\|^2 + c(\lambda)\max_t(\eta_t + \eta_t^2)R).$$

Moreover, recall that

$$\mathbb{E}|\langle u_t, \theta_{t-1}\rangle| \leq \frac{1}{2}(\mathbb{E}\|u_t\|^2 + \mathbb{E}\|\theta_{t-1}\|^2) \leq \frac{1}{2}\mathbb{E}\|u_t\|^2 + \frac{1}{2}(\|\theta_0\|^2 + c(\lambda)\max_t(\eta_t + \eta_t^2)R).$$

Therefore, taking expectations of 4, summing over all $t$, combining inequality 2 with the above, and lastly using Corollary 1 of (Tang et al., 2023), we find that

$$\begin{aligned}\mathbb{E}F(\theta_T) \leq F(\theta_0) &- \frac{\eta}{2(\delta_0 + C_1)}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|^2 \\ &+ \left[\frac{2\eta\sqrt{C^2+\Phi}}{\sqrt{1-\beta_2}} + \frac{\eta^2 L(1+\lambda)}{2(1-\beta_2)}\right]\sum_{0}^{T-1}\mathbb{E}\|u_t\|^2 \\ &+ \frac{1}{2}(C_1^2 + \|\theta_0\|^2 + c(\lambda)\max_t(\eta_t + \eta_t^2)R)\left(\lambda\sum_{t=1}^{T}\eta_t + L(\lambda+\lambda^2)\sum_{t=1}^{T}\eta_t^2\right).\end{aligned}$$

Rearranging, noting that $\mathbb{E}F(\theta_T) \geq F_*$, and applying inequality 1, we find that

$$\begin{aligned}\frac{1}{T}\sum_{0}^{T-1}\mathbb{E}\|\nabla F(\theta_{t-1})\|^2 \leq \frac{2(\delta_0 + C_1)}{\eta}&\left(\frac{F(\theta_0) - F_*}{T}\right) \\ &+ \left(\frac{4d(C^2+\Phi)}{\sqrt{1-\beta_2}} + \frac{\eta d L\sqrt{C^2+\Phi}(1+\lambda)}{1-\beta_2}\right)\cdot\frac{R}{T} \\ &+ \frac{1}{2T}\left(C_1^2 + \|\theta_0\|^2 + c(\lambda)\max_t(\eta_t + \eta_t^2)R\right)\left(\lambda\sum_{t=1}^{T}\eta_t + \frac{L}{2}(\lambda+\lambda^2)\sum_{t=1}^{T}\eta_t^2\right).\end{aligned}$$

This is exactly the desired inequality and completes the proof of the theorem. □

*Remark.* We owe a quick explanation of the starting line of the proof; for DP-AdamW-BC, one only needs to replace the inequalities given by 2 and 1 with the corresponding bounds in the proof of Proposition 6 of (Tang et al., 2023); the remainder of the proof is the exact same.

For an interpretation of the asymptotic growth of this result, see the remark following the initial theorem statement.

### B.2. Proof of Theorem 3.6

We prove the convergence result given in Theorem 3.6 below.

*Proof of Theorem 3.6.* We will only prove the result for DP-AdamW; the exact same proof technique will suffice for DP-AdamW-BC. As in (Tang et al., 2023), define the notation $u_t = \frac{\nabla_i f_t^p(\theta_{t-1})}{\sqrt{v_t^p + \epsilon_0}}$ denote the main 'update term' in DP-Adam; that is, $\nabla_i f_t^p(\theta_{t-1}) = \tilde{m}_t$ and $v_t^p = \tilde{v}_t$. We first prove a bound on the expected magnitude of the parameters, then proceed as in (Tang et al., 2023).

We first note that Lemma 3 and Corollary 1 in Appendix F of (Tang et al., 2023) hold, with the proof remaining identical. Moreover, identical arguments as in the proof of Proposition 7 of (Tang et al., 2023) yield the inequalities[8]

$$\sum_{t=0}^{T-1} \mathbb{E}[\|u_t\|^2] \leq d\left(\ln\left(1 + \frac{\delta_0^2}{\epsilon_0(1-\beta_2)}\right) - T\log\beta_2\right) \tag{6}$$

and moreover

$$\mathbb{E}\left[\nabla F(\theta_{t-1})^\top u_t\right] \geq \frac{1}{2}\sum_{i\in[d]}\sum_{k=0}^{t-1}\beta_1^k \mathbb{E}\left[\frac{\nabla_i F(\theta_{t-k-1})}{\sqrt{\tilde{v}_{t,k+1,i} - \Phi}}\right] \tag{7}$$

$$- \frac{\sqrt{1-\beta_1}\eta_t^2 L^2}{4\delta_0}\sum_{l=1}^{t-1}\|u_{t-l}\|^2\sum_{k=1}^{l-1}\beta_1^k\sqrt{k} \tag{8}$$

$$- \frac{3\delta_0}{\sqrt{1-\beta_1}}\sum_{k=0}^{t-1}\left(\frac{\beta_1}{\beta_2}\right)^2\|u_{t-k}\|^2. \tag{9}$$

In particular, the first statement follows from Lemma 5.2 in (Défossez et al., 2022), and the second by using Lemmas 5.1 and 5.2 of the aforementioned work.

We now proceed as follows. First, we bound the magnitude of the parameters $\theta_t$ in expectation. Let

$$R = d\left(\ln\left(1 + \frac{\delta_0^2}{\epsilon_0(1-\beta_2)}\right) - T\log\beta_2\right).$$

**Claim:** $\mathbb{E}\|\theta_t\|^2 \leq \|\theta_0\|^2 + c(\lambda)R$, where $c(\lambda)$ is a constant depending on $\lambda$ and all other constant parameters. [9]

**Proof:** The main idea is to combine the recursive definition of $\theta_t$ with the AM-GM inequality and 2, Let $x_t = \mathbb{E}\|\theta_t\|^2$ and write

$$\begin{aligned}
x_t = \mathbb{E}\|\theta_t\|^2 &= (1-\eta_t\lambda)^2\mathbb{E}\|\theta_{t-1}\|^2 + \eta_t^2\mathbb{E}\|u_t\|^2 + 2\eta_t(1-\eta_t\lambda)\mathbb{E}[\langle\theta_{t-1}, u_t\rangle] \\
&\leq (1-\eta_t\lambda)^2 x_{t-1} + \eta_t^2\mathbb{E}\|u_t\|^2 + 2\eta_t(1-\eta_t\lambda)\left(\frac{\lambda}{2}\mathbb{E}\|\theta_{t-1}\|^2 + \frac{1}{2\lambda}\mathbb{E}\|u_t\|^2\right) \\
&= \left[\eta_t^2 + \frac{\eta_t(1-\eta_t\lambda)}{\lambda}\right]\mathbb{E}\|u_t\|^2 + (1-\eta_t\lambda)\mathbb{E}\|\theta_{t-1}\|^2 \\
&\leq \left(\eta_t^2 + \frac{\eta_t}{\lambda}\right)\mathbb{E}\|u_t\|^2 + x_{t-1},
\end{aligned}$$

where we have used the fact that $\langle x, y\rangle = \sum x_i y_i \leq \lambda\sum x_i^2 + \frac{1}{\lambda}\sum y_i^2$. It follows by summing across all $t$, telescoping, and applying inequality 1 that

$$\begin{aligned}
x_t &\leq \|\theta_0\|^2 + (1+\frac{1}{\lambda})\left(\max_t(\eta_t + \eta_t^2)d\left(\ln\left(1 + \frac{\delta_0^2}{\epsilon_0(1-\beta_2)}\right) - T\log\beta_2\right)\right) \\
&= \|\theta_0\|^2 + c(\lambda)\max_t(\eta_t + \eta_t^2)R.
\end{aligned}$$

[8] These results are directly shown and/or implicit in the referenced proof; the proof is immediately adaptable.

[9] Here $c(\lambda)$ notation encapsulates all constant parameters, as well as $\theta_0$.

This completes the proof of the claim.

Now we return to the proof of the main theorem. We follow the rough outline used by (Tang et al., 2023). The main idea is to use the Lipschitzness of the gradient to bound the difference $F(\theta_t) - F(\theta_{t-1})$. One can then telescope this difference and combine the claim with inequalities 1 and 2 and some algebraic manipulations.

Let us write $\theta_t = \theta_{t-1} - \eta_t(u_t + \lambda_t\theta_{t-1})$. Using the Lipschitzness assumption of the gradient, we have

$$F(\theta_t) \leq F(\theta_{t-1}) - \eta_t\langle \nabla F(\theta_{t-1}), u_t\rangle - \eta_t\lambda\langle \nabla F(\theta_{t-1}), \theta_{t-1}\rangle + \frac{\eta_t^2 L}{2}\|u_t\|^2 \tag{10}$$

$$+ \frac{\eta_t^2 L}{2}\lambda^2\|\theta_{t-1}\|^2 + \eta_t^2 L\lambda\langle u_t, \theta_{t-1}\rangle. \tag{11}$$

Taking expectations, we find by the above claim that $\mathbb{E}\|\theta_{t-1}\|^2 = \|\theta_0\|^2 + c(\lambda)\max_t(\eta_t + \eta_t^2)R$ and hence

$$\mathbb{E}|\langle \nabla F(\theta_{t-1}), \theta_{t-1}\rangle| \leq \frac{1}{2}(\mathbb{E}\|\nabla F(\theta_{t-1})\|^2 + \mathbb{E}\|\theta_{t-1}\|^2) \leq \frac{1}{2}(C_1^2 + \|\theta_0\|^2 + c(\lambda)\max_t(\eta_t + \eta_t^2)R).$$

Moreover, recall that

$$\mathbb{E}|\langle u_t, \theta_{t-1}\rangle| \leq \frac{1}{2}(\mathbb{E}\|u_t\|^2 + \mathbb{E}\|\theta_{t-1}\|^2) \leq \frac{1}{2}\mathbb{E}\|u_t\|^2 + \frac{1}{2}(\|\theta_0\|^2 + c(\lambda)\max_t(\eta_t + \eta_t^2)R).$$

Therefore, taking expectations of 10, summing over all $t$, combining inequality 7 with the above, and lastly using Corollary 1 of (Tang et al., 2023) and the simple inequality $\eta_t \leq \eta_T$, we find that

$$\begin{aligned}\frac{\sum_{t=1}^T \frac{\eta_t}{\Omega_t}\sum_{k=0}^{t-1}\beta_1^k\mathbb{E}\left[\|\nabla F(\theta_{t-k-1})\|^2\right]}{2(\delta_0 + C_1)} &\leq F(\theta_0) - F_* \\ &+ \frac{\eta_T^2 L}{2}\sum_{t=1}^T\mathbb{E}\left[\|u_t\|^2\right] + \frac{\eta_T^3 L^2\sqrt{1-\beta_1}}{4(\delta + C_1)}\sum_{t=1}^T\sum_{l=1}^{t-1}\beta_1^l\sqrt{l} \\ &+ \frac{3\eta_T(\delta + C_1)}{\sqrt{1-\beta_1}}\sum_{t=1}^T\sum_{k=0}^{t-1}\left(\frac{\beta_1}{\beta_2}\right)^k\mathbb{E}\left[\|u_{t-k}\|^2\right].\end{aligned}$$

where $\Omega_t = \sqrt{\sum_{j=0}^{t-1}\beta_2^j}$. Rearranging, noting that $\mathbb{E}F(\theta_T) \geq F_*$, and applying inequality 6, we find that

$$\mathbb{E}[\|\nabla F(\theta_\tau)\|^2] \leq \frac{2(\delta_0 + C_1)(F(\theta_0) - F_*)}{\eta\tilde{T}} + ER,$$

where the left hand side expectation is taken with respect to the sample $\mathbb{P}(\tau = t) \propto 1 - \beta_1^{T-t}$, $\tilde{T} = T - \frac{\beta_1}{1-\beta_1}$, and

$$E = \frac{\eta dL(1-\beta_1)(\delta_0 + C_1)}{(1-\beta_1/\beta_2)(1-\beta_2)} + \frac{2\eta^2 dL^2\beta_1}{(1-\beta_1/\beta_2)(1-\beta_2)^{3/2}} + \frac{12d(\delta_0 + C_1)^2\sqrt{1-\beta_1}}{(1-\beta_1/\beta_2)^{3/2}\sqrt{1-\beta_2}},$$

as in the statement of the theorem. This is exactly the desired inequality and completes the proof of the theorem. □

*Remark.* Similar to the proof of Theorem 3.5, we note that for DP-AdamW-BC, one only needs to replace the inequalities given by 7 and 6 with the corresponding bounds in the proof of Proposition 7 of (Tang et al., 2023); the remainder of the proof is the exact same.

# C. Experimental Details

## C.1. Image Classification

We train a 5-layer CNN model to perform image classification on the CIFAR-10 dataset (Krizhevsky et al., 2009). The CIFAR-10 dataset includes 60,000 32x32 images, with 6,000 images for each of 10 classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck. We first train on the training set containing 50,000 images, then we test on the test set of the 10,000 remaining images. We train the model from scratch with randomized initial model parameters.

To ensure a fair comparison and optimal performance for each optimizer, we conduct extensive hyperparameter sweeps. Following preliminary exploration and to maintain consistency with Tang et al. (2023), we fix the $L_2$-norm clipping bound at $C = 1.0$ and utilized Adam stability parameters of $5 \times 10^{-8}$. Our primary hyperparameter sweep focuses on optimizing learning rate ($\alpha$) and weight decay ($\lambda$) for each target privacy level ($\epsilon \in \{1, 3, 7\}$) and for both DP-AdamW variants (with and without bias correction). We explore a range of learning rates and weight decay values suitable for each privacy setting based on values commonly used in related DP deep learning literature. All experiments are run for 70 epochs using a batch size of 1024, and results are averaged over 5 trials with different random seeds. We select weight decay $\lambda = 1 \times 10^{-5}$ after sampling over $\lambda \in \{0.01, 0.001, 1 \times 10^{-4}, 1 \times 10^{-5}\}$ and optimizing training loss.

Figures 1 and 2 in Appendix D show the training loss and test accuracy over learning rates for $\epsilon = 1$. We select learning rates by examining the values that minimized training loss. We outline learning rates over $\epsilon$ values in Table 4.

## C.2. Text Classification

We study natural-language understanding under differential privacy on the standard question-natural language inference (QNLI) task from the GLUE natural language benchmark (Wang et al., 2019). QNLI contains 115,669 (question, sentence) pairs split into 105K/5.5K/5.5K (train/validation/test) examples and asks whether the sentence answers the corresponding question. Performance is reported as classification accuracy on the held-out test set. We use the DP-Adam and DP-Adam-BC results from Tang et al. (2023) as a baseline comparison, although we independently fine-tune hyperparameters for our implementations of DP-AdamW and DP-AdamW-BC.

Following common practice for private fine-tuning (e.g. Li et al., 2022b; Tang et al., 2023) we start from the `bert-base-cased` checkpoint with 110M parameters and freeze all layers except the last Transformer block, the pooler, and the task-specific classifier head, which yields approximately 7M trainable parameters. The model is trained for 10 epochs with a batch size of $B = 32$. Gaussian noise calculated to match the target privacy budget $\epsilon \in \{1, 3, 7\}$ is added through OPACUS' `PrivacyEngine` and per-sample gradients are clipped using an $\ell_2$-norm threshold before aggregation and noise addition. To ensure compatibility, the same noise parameters are used across all optimizers.

For every combination of ($\epsilon$, optimizer) under consideration and following the approach of (Tang et al., 2023), we perform a grid search over hyperparameters: learning rate $\alpha \in [1 \times 10^{-6}, 5 \times 10^{-3}]$, weight decay $\lambda \in \{0, 10^{-5}, 10^{-4}\}$, and gradient clipping bound $C \in \{0.05, 0.1, 0.2\}$. Additionally, after preliminary exploration, we fix $\gamma = 10^{-8}$ for the minimum variance term in DP-AdamW-BC. For each privacy setting, the learning rates and weight decay values were chosen based on those commonly used in the DP deep learning literature. Moreover, the candidate values for the gradient clipping bound were those examined by (Tang et al., 2023) therefore we consider the same values for experimental consistency purposes. We optimize for standard cross-entropy loss during training. Note that this choice of loss function is correct because the specific problem is a classification task, where cross-entropy is the canonical choice, and it is differentiable meaning that it works nicely with the per-sample gradient computation that OPACUS requires for differential privacy.

## C.3. Graph Node Classification

We evaluated different optimizers on the graph node classification task using the ogbn-arxiv task from the Open Graph Benchmark ((Hu et al., 2021)) and trained a 2-layer differentially private graph convolutional network model ((Daigavane et al., 2022)) without per-layer clipping from scratch. We tested for $\epsilon \in \{3, 6, 12\}$ across the optimizers. The total number of parameters is 173,695 and the specific architecture details can be found in the Appendix at Section E.

For each new optimizer-$\epsilon$ pair, we did a hyperparameter sweep of 10 runs over a validation set using a Bayes-informed sampling over a log-uniform distribution to maximize validation accuracy. The results of this preliminary sweep are reported in Table 6 in the appendix. With this preliminary finding, we then adjusted the range to be swept to encompass our best guess of the best hyperparameters to run another sweep of 30 runs. The final range is $[10^{-1}, 10^{-6}]$ for the learning rate and

$[10^{-1}, 10^{-7}]$ for the weight decay. The best learning rate and weight decay chosen from the 30-run sweep are reported in Table 7 in the Appendix. We noticed that as $\epsilon$ increases, the best learning rate decreases while the weight decay generally, though not always, increases. We found that validation accuracy increases as $\epsilon$ increases, which is expected since models perform better with looser privacy budgets. We attach the plots of the losses in Appendix E for the case of DP-AdamW under $\epsilon = 3$.

## D. Image Classification Experimental Details

*Table 4.* Learning rates for $\epsilon$ values on CIFAR-10

| | $\epsilon = 1$ | $\epsilon = 3$ | $\epsilon = 7$ |
|---|---|---|---|
| DP-AdamW | 0.0015 | 0.0015 | 0.002 |
| DP-AdamW-BC | $1 \times 10^{-4}$ | $2 \times 10^{-4}$ | $4 \times 10^{-4}$ |

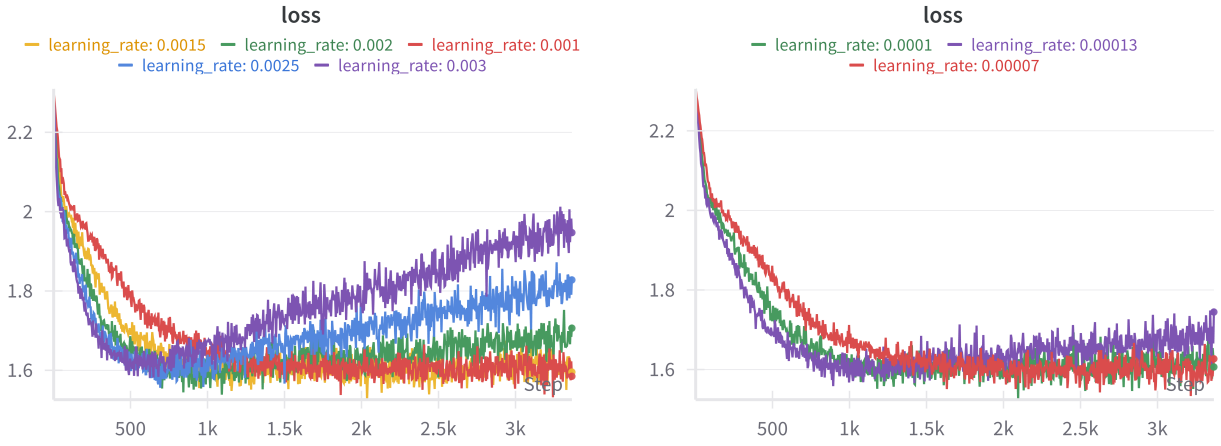


*Figure 1.* Training CIFAR-10 for $\epsilon = 1$ across learning rates for DP-AdamW (left) and DP-AdamW-BC (right), with step on x-axis and training loss on y-axis

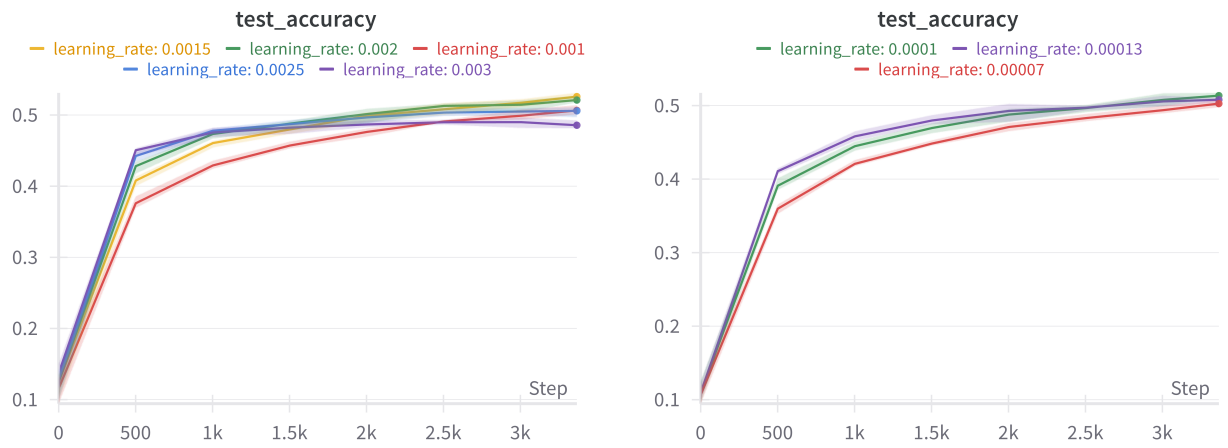


*Figure 2.* Evaluating on CIFAR-10 for $\epsilon = 1$ across learning rates for DP-AdamW (left) and DP-AdamW-BC (right), with step on x-axis and test accuracy (proportion) on y-axis

## E. Graph Node Classification Experimental Details

Below we provide more details on the training process of the graph node classification task on which the optimizers were evaluated. Table 5 shows the architecture of the DP-GCN used in Section 4.3, with a total of 173,695 parameters (694,780 bytes).

*Table 5.* Model parameter names, shapes, and data types

| Parameter | Shape | Dtype |
|---|---|---|
| Core Layer 0: Dense Bias | (255,) | float32 |
| Core Layer 0: Dense Kernel | (255, 255) | float32 |
| Decoder Layer 0: Dense Bias | (255,) | float32 |
| Decoder Layer 0: Dense Kernel | (255, 255) | float32 |
| Decoder Layer 1: Dense Bias | (40,) | float32 |
| Decoder Layer 1: Dense Kernel | (255, 40) | float32 |
| Encoder Layer 0: Dense Bias | (255,) | float32 |
| Encoder Layer 0: Dense Kernel | (128, 255) | float32 |

Table 6 shows the best hyperparameters for the GNN task identified after the initial 10 runs of the hyperparameter sweep.

*Table 6.* Hyperparameters and validation performance for DP-AdamW and DP-AdamW-BC on `obgn-arxiv` node classification (best validation score of 10)

| Optimizer | Epsilon | Learning Rate | Weight Decay | Val. Acc. (%) |
|---|---|---|---|---|
| AdamW | 3 | 0.01494606 | 0.00047264 | 51.32 |
| | 6 | 0.00752970 | 0.00095908 | 54.44 |
| | 12 | 0.00434687 | 0.00051222 | 56.16 |
| AdamW-BC | 3 | 0.00018378 | 0.0000026934 | 50.30 |
| | 6 | 0.00009699 | 0.0000024191 | 54.28 |
| | 12 | 0.00004261 | 0.00013755 | 55.86 |

Table 7 shows the best hyperparameters for the GNN task identified after the follow-up 30 runs of the hyperparameter sweep.

*Table 7.* Hyperparameters and validation performance for DP-AdamW and DP-AdamW-BC on `obgn-arxiv` node classification (best validation score of 30)

| Optimizer | Epsilon | Learning Rate | Weight Decay | Val. Acc. (%) |
|---|---|---|---|---|
| DP-AdamW | $\epsilon \approx 3$ | 0.01246956 | 0.0000004780 | 52.41 |
| | $\epsilon \approx 6$ | 0.00806904 | 0.0000091133 | 55.18 |
| | $\epsilon \approx 12$ | 0.00427425 | 0.0000709134 | 56.27 |
| DP-AdamW-BC | $\epsilon \approx 3$ | 0.00021068 | 0.0000324537 | 50.78 |
| | $\epsilon \approx 6$ | 0.00007306 | 0.0000012885 | 54.66 |
| | $\epsilon \approx 12$ | 0.00004086 | 0.03112375 | 56.15 |

Figure 3 shows DP-AdamW training, validation and test losses across $[10^{-1}, 10^{-6}]$ for learning rate and $[10^{-1}, 10^{-7}]$ for weight decay using a privacy budget under $\epsilon = 3$.

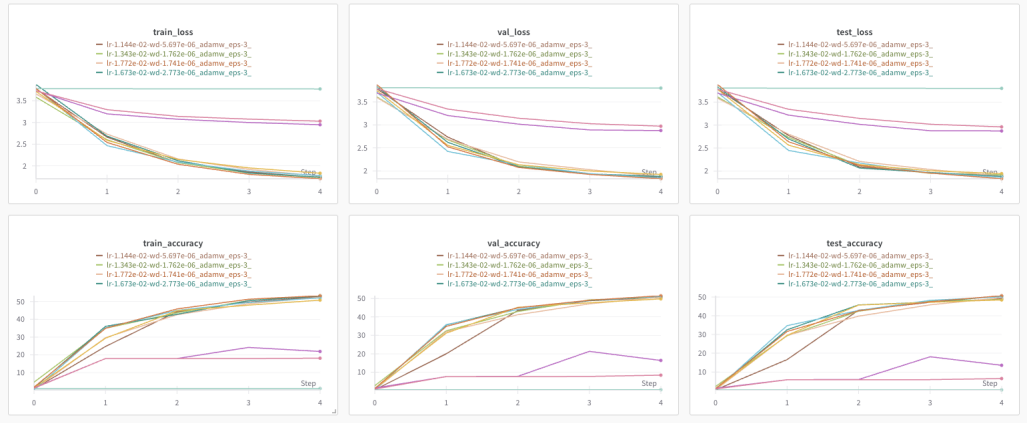


*Figure 3.* Losses when using DP-AdamW under $\epsilon = 3$

# F. Future Work

## F.1. Broader empirical generalization within text classification

Our text classification results in Table 2 show that DP-AdamW both with and without bias correction significantly outperforms other optimizers on the QNLI natural language task. This performance improvement is substantially higher than the corresponding improvements for the image and graph node classification tasks so we would like to investigate further. Potential patterns to examine include persistence across different text subtasks and scaling with dataset size and class balance. One potential research direction is to run our existing QNLI fine-tuning protocol on the SNLI single-sentence entailment, MultiNLI multi-genre entailment, and QQP paraphrase detection tasks described in (Bowman et al., 2015; Wang et al., 2019), containing diverse training dataset sizes (QNLI $\approx$ 105K, SNLI $\approx$ 550K, MultiNLI $\approx$ 393K, QQP $\approx$ 364K) and subtask types.

## F.2. Scaling laws for DP-AdamW and DP-AdamW-BC

In addition to increasing the robustness of our text classification analysis, we would also like to investigate the scaling laws of the proposed optimizers on both image and graph node classification. We hypothesize that the difference in performance on text classification compared to the other two tasks may stem from the fact that text classification operates in the large-model, large-dataset regime, while the other two tasks are in the small model regime. Note that the text classification task uses a model with over 110M parameters (bert-base-uncased), while the image classification task uses a 5-layer CNN model while the node classification task uses a 2-layer GNN model, which are several orders of magnitude smaller than that used for text. Evaluating the performance of DP-AdamW and DP-AdamW-BC on other tasks that use large models and large datasets could uncover a larger performance differential relative to other optimizers as found in the text classification results but unobserved in our image classification and node classification results.

## F.3. Examination of bias correction underperformance

Recall that across our experiments (Table 1, Table 2, Table 3), adding bias correction consistently reduces accuracy on all three classification tasks for DP-AdamW, although the opposite phenomenon seems true for DP-Adam. We suspect that the $\Phi = (\sigma C/B)^2$ interaction term that is subtracted inside the square root in the denominator could become large and force the algorithm to clamp to $\gamma$; under decoupled weight decay, the subsequent large adaptive step is uniquely not offset inside the gradient so the next decay must account for it. We propose to train a lightweight classifier on a small dataset and plot histograms every epoch of measurements such as the clamping rate, which would help determine if the bias term becomes dominated by estimation noise.