DP-AdamBC: Your DP-Adam Is Actually DP-SGD
(Unless You Apply Bias Correction)

Qiaoyue Tang, Frederick Shpilevskiy, Mathias Lécuyer

University of British Columbia
Vancouver, British Columbia, Canada
{qiaoyuet, fshpil}@cs.ubc.ca, mathias.lecuyer@ubc.ca

Abstract

The Adam optimizer is a popular choice in contemporary deep learning, due to its strong empirical performance. However we observe that in privacy sensitive scenarios, the traditional use of Differential Privacy (DP) with the Adam optimizer leads to sub-optimal performance on several tasks. We find that this performance degradation is due to a DP bias in Adam's second moment estimator, introduced by the addition of independent noise in the gradient computation to enforce DP guarantees. This DP bias leads to a different scaling for low variance parameter updates, that is inconsistent with the behavior of non-private Adam. We propose DP-AdamBC, an optimization algorithm which removes the bias in the second moment estimation and retrieves the expected behaviour of Adam. Empirically, DP-AdamBC significantly improves the optimization performance of DP-Adam by up to 3.5% in final accuracy in image, text, and graph node classification tasks.

1 Introduction

The Adam optimization algorithm (**?**) is the default optimizer for several deep learning architectures and tasks, notably in Natural Language Processing (NLP), for which Stochastic Gradient Descent (SGD) tends to struggle. Even in vision tasks where Adam is less prevalent, it typically requires less parameter tuning than SGD to reach good performance.

On all these tasks, deep learning models can leak information about their training set (**????**). We consider settings in which the deep learning model's training data is privacy sensitive, and models are trained with Differential Privacy (**??**) to provably prevent training example information leakage (**?**). Intuitively, training DP models requires computing each minibatch gradient with DP guarantees by clipping per-example gradients and adding Gaussian noise (§3), to bound the maximal influence of any data-point on the final model. The DP gradients can then feed into any optimization algorithm without modification to update the model's parameters. Due to its success in the non-private setting, Adam is also prevalent when training DP models, for NLP (**?**) and GNN (**?**) models. However we observe that when combined with DP, Adam does not perform as well as without privacy

constraints: Adam suffers a larger degradation of performance compared to SGD on vision tasks, while NLP models perform poorly when training from scratch.

To understand this effect, we go back to the original intuition behind Adam (**?**) that relies on exponential moving averages estimating the first and second moments of minibatch gradients. We show that while DP noise does not affect the first moment, it does add a constant bias to the second. Drawing on a recent empirical investigation that suggests that the performance of Adam may be linked to its update rule performing a smooth version of the sign descent update (**?**), we show that the additive shift in Adam's second moment estimate caused by DP noise moves the Adam update away from that of sign descent, by scaling the gradient dimensions with different magnitudes differently. Indeed, under typical DP parameters, the DP bias added to the second moment estimates of DP-Adam dominate the second moment estimate, and makes DP-Adam a rescaled version of DP-SGD with momentum. We show how to correct this DP noise induced bias, yielding a variation that we call DP-AdamBC. Empirically, correcting Adam's second moment estimate for DP noise significantly increases test performance for Adam with DP, on tasks for which Adam is well suited.

We make the following contributions:

1. We analyze the interaction between DP and the Adam optimizer, and show that DP noise introduces bias in Adam's second moment estimator (§3). We show theoretically, and verify empirically, that under typical DP parameters DP-Adam reduces to DP-SGD with momentum (§3). This behavior violates the sign-descent hypothesis for Adam's performance.

2. We propose DP-AdamBC, a variation of DP-Adam that corrects for the bias introduced by DP noise. We show that DP-AdamBC is a consistent estimator for the Adam update, under the same simplifying assumptions that justify Adam's update. (§4)

3. We empirically evaluate the effect of DP-AdamBC, and show that it yields significant improvements (up to 3.5 percentage points of test accuracy) over DP-Adam. (§5).

Our implementation is available at: https://github.com/ubc-systopia/DP-AdamBC. All Appendixes referenced in the paper are available in the long version of the paper (**?**).

2 Adam and the Sign-descent Hypothesis

The Adam update (**?**) is defined as follows. Denote the average gradient over a mini-batch of size B with respect to loss function f at step t as:

$$g_t = (1/B)\nabla f(\theta_{t-1})$$

Let β_1 and β_2 be Adam's decay coefficients. At each step, Adam updates two estimators:

$$m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1)\, g_t \; ; \quad \widehat{m}_t \leftarrow m_t/\left(1 - \beta_1^t\right),$$

$$v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2)\, g_t^2 \; ; \quad \widehat{v}_t \leftarrow v_t/\left(1 - \beta_2^t\right).$$

Finally, the Adam update for the model's parameters is:

$$\theta_t \leftarrow \theta_{t-1} - \eta\Delta_t \; ; \quad \Delta_t = \hat{m}_t/(\sqrt{\hat{v}_t} + \gamma),$$

with learning rate η, and $\gamma > 0$ a small numerical stability constant. Intuitively, Adam's $\hat{m}_t$ and $\hat{v}_t$ use an exponential moving average to estimate $\mathbb{E}[g_t]$ and $\mathbb{E}[g_t^2]$, the vector of first and second moment of each parameter's gradient, respectively. The final update is thus approximating $\mathbb{E}[g_t]/\sqrt{\mathbb{E}[g_t^2]}$.

The reasons for Adam's performance are not fully understood. However, recent evidence (**?**) supports the hypothesis that Adam derives its empirical performance from being a smoothed out version of sign descent. At a high level, Adam performs well in settings (e.g., NLP) where sign descent also performs well, at least when running with full (or very large) batch. We next describe Adam's update rule under this sign descent hypothesis, before working out the impact of DP noise on this interpretation. Let $\mathbb{E}$ and $\mathbb{V}$ denotes the expectation and variance respectively,

1. If for parameter i, $|\mathbb{E}[g_t]|_i \gg \sqrt{\mathbb{V}[g_t]_i}$, then the update's direction is clear. And since $|\mathbb{E}[g_t]|_i \approx \sqrt{\mathbb{E}[g_t^2]_i}$, the Adam update is $\mathbb{E}[g_t]_i/\sqrt{\mathbb{E}[g_t^2]_i} \approx \pm 1$, and Adam is sign descent. Updates are *not scaled based on* $|\mathbb{E}[g_t]|_i$ as in SGD.

2. If for parameter i, $|\mathbb{E}[g_t]|_i \not\gg \sqrt{\mathbb{V}[g_t]_i}$, the sign is less clear and Adam's update is in $[-1, 1]$, scaled closer to 0 the more uncertain the sign is (smoothing behavior).

Finally, Adam ensures numerical stability when $|\mathbb{E}[g_t]|_i \approx 0$ and $\mathbb{V}[g_t]_i \approx 0$ using the additive constant γ in the denominator of the update. In that case, the update is approximately $\mathbb{E}[g_t]_i/\gamma \approx 0$.

To summarize, under the sign descent hypothesis, Adam updates parameters with low variance gradients using a constant size ± 1 update (or $\pm\eta$ after the learning rate is applied), and rescales the update of parameters with high variance gradients towards 0. As we describe next, adding DP to gradient computations breaks this interpretation of Adam as sign descent.

3 Adam Update under Differential Privacy

Most optimization approaches for deep learning models with Differential Privacy follow a common recipe (**?**): compute each gradient update over a mini-batch with DP, and leverage DP's post-processing guarantee and composition properties to analyse the whole training procedure. Computing a

DP update over a mini-batch involves clipping per-example gradients to control the update's sensitivity, and adding *independent* Gaussian noise to the aggregated gradients. Formally, for each step t, let $g_n = \nabla f(\theta_t, x_n)$ be the gradient for sample n, and let C, σ be the maximum $L2$-norm clipping value and the noise multiplier, respectively. Given a mini-batch B, the DP gradient is:

$$\tilde{g}_t = \bar{g}_t + (1/B)z_t \; ; \quad z_t \sim \mathcal{N}(0, \sigma^2 C^2 \mathbb{I}^d) \; ;$$

$$\bar{g}_t = (1/B)\sum_n g_n/\max\left(1, \|g_n\|_2/C\right),$$

where $\bar{g}_t$ is the mean of clipped gradients over the minibatch—a biased estimate of g_t—and $\tilde{g}_t$ the DP gradient.

With this recipe, any optimizer that only takes mini-batch updates as input, such as Adam, can be applied to the DP update $\tilde{g}$ and preserve privacy. This is how existing DP approaches using Adam work (e.g., (**?**)), yielding the following update: let the superscript p denote private version of a quantity, then

$$m_t^p \leftarrow \beta_1 m_{t-1}^p + (1 - \beta_1)\, \tilde{g}_t, \; \widehat{m}_t^p \leftarrow m_t^p/\left(1 - \beta_1^t\right),$$

$$v_t^p \leftarrow \beta_2 v_{t-1}^p + (1 - \beta_2)\, \tilde{g}_t^2, \; \widehat{v}_t^p \leftarrow v_t^p/\left(1 - \beta_2^t\right),$$

$$\theta_t \leftarrow \theta_{t-1} - \eta\widehat{m}_t^p/(\sqrt{\widehat{v}_t^p} + \gamma).$$

We show next that this DP-Adam algorithm uses a biased estimator for the second moment. This bias dominates the scale of the denominator in Adam's update, thus breaking the sign descent behaviour of Adam (§3) and reducing DP-Adam to DP-SGD with momentum and a specific learning rate schedule (§3).

DP noise biases second moment estimates, breaking the sign descent behavior

Under DP, Adam estimates the first and second moments as m_t^p and v_t^p, and rescaled versions $\hat{m}_t^p$ and $\hat{v}_t^p$, using $\tilde{g}_t$ in order to preserve privacy. Since the noise added for DP is independent of the gradient update, there is no impact on the first moment in expectation:

$$\mathbb{E}[m_t^p] = \mathbb{E}\left[(1 - \beta_1)\sum_{\tau=1}^{t}\beta_1^{t-\tau}\tilde{g}_\tau\right]$$

$$= (1 - \beta_1)\sum_{\tau=1}^{t}\beta_1^{t-\tau}\left(\mathbb{E}[\bar{g}_\tau] + \underbrace{\frac{1}{B}\mathbb{E}[z_\tau]}_{0}\right) \triangleq \mathbb{E}[m_t^c]. \tag{1}$$

However, v_t^p is now a biased estimate of the second moment of the mini-batch's update $\bar{g}_t$, as it incurs a constant shift due to DP noise (**?**). By independence of the DP noise z_t and $\bar{g}_t$, we have that:

$$\mathbb{E}[v_t^p] = \mathbb{E}\left[(1 - \beta_2)\sum_{\tau=1}^{t}\beta_2^{t-\tau}\tilde{g}_\tau^2\right]$$

$$= \underbrace{(1 - \beta_2)\sum_{\tau=1}^{t}\beta_2^{t-\tau}\mathbb{E}[\bar{g}_\tau^2]}_{\triangleq\, \mathbb{E}[v_t^c]} + (1 - \beta_2^t)\underbrace{\left(\frac{\sigma C}{B}\right)^2}_{\Phi}. \tag{2}$$

In these equations, $\mathbb{E}[m_t^c]$ and $\mathbb{E}[v_t^c]$ are the quantities that would be estimated under regular Adam (without DP noise), computed with respect to $\bar{g}_t$ (clipped gradients for DP).
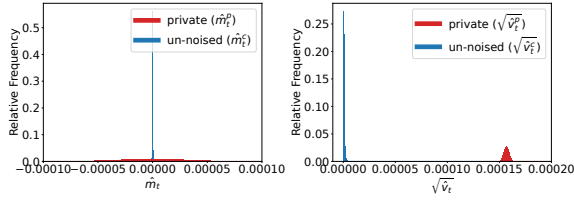

Figure 1: Histogram of **Left:** un-noised ($\hat{m}_t^c$) and private ($\hat{m}_t^p$) first moment estimates, **Right:** un-noised ($\hat{v}_t^c$) and private ($\hat{v}_t^p$) second moment estimates near end of training at $t = 10000$, using the SNLI dataset with $B = 256, C = 0.1, \sigma = 0.4, \beta_2 = 0.999, \Phi \approx 2.441\text{e-}8$ for large t.

We use a text classification dataset (SNLI) to demonstrate the effect of DP noise on first and second moment estimates, with $B = 256, C = 0.1, \sigma = 0.4, \beta_2 = 0.999, \Phi = 2.441\text{e-}8$ with large t.

Figure 1 (Left) shows the histogram of values of the first moment estimates $\hat{m}_t^c$ (clipped gradients, no noise) for each dimension, and private $\hat{m}_t^p$ (clipped and noised gradients), at the end of training. We observe that the center of the distributions align, suggesting that $\mathbb{E}[\hat{m}_t^p] = \mathbb{E}[\hat{m}_t^c]$ as in Equation 1. The private first moment distribution has larger variance compared to the clean distribution as a result of DP noise. Figure 1 (Right) shows the histogram of $\hat{v}_t^c$ (clipped, no noise) and private $\hat{v}_t^p$ (clipped and noised) second moment estimates at the end of training. We see that the distributions of $\hat{v}_t^c$ and $\hat{v}_t^p$ are quite different, with a shift in the center approximately equal to $\sqrt{\Phi}$. This suggests that the DP noise variance dominates the scale of $\hat{v}_t^p$ in Equation 2.

To understand the implication of DP noise bias Φ, let us follow the original Adam paper (**?**) and interpret the update under the following assumption:

Assumption 1 (Stationarity). *For all τ in $[0, t]$, the (full) gradient is constant, $\nabla f(\theta_\tau) \triangleq \nabla f$, and minibatch gradients are i.i.d samples such that $\mathbb{E}[g_t] = \nabla f$.*

Remark. *Note that Assumption 1 is not required for convergence (see Appendix F), nor is it used in empirical experiments. It is useful though, to reason about the behavior of DP-Adam and compare it to the intended behavior of Adam without DP, as we do next. The same assumption was used in Adam's original work for the same purpose, to reason about the quality of Adam's moment estimates [(**?**), §3].*

Under Assumption 1, with $\beta_1 \rightarrow 1$, β_2 such that $(1 - \beta_1^t)/\sqrt{1 - \beta_2^t} = 1$, and for large enough t, we have that $\mathbb{E}[\hat{m}_t^c] \approx \mathbb{E}[\bar{g}_t]$ and $\mathbb{E}[\hat{v}_t^c] \approx \mathbb{E}[\bar{g}_t^2]$, and $\Delta_t = \mathbb{E}[\tilde{g}_t]/\sqrt{\mathbb{E}[\tilde{g}_t^2]} = \mathbb{E}[\bar{g}_t]/\sqrt{\mathbb{V}[\bar{g}_t] + \mathbb{E}(\bar{g}_t)^2 + \Phi}$. Due to the extra DP bias Φ in the denominator of Adam's estimator, DP-Adam no longer follows the sign descent interpretation seen in §2.

Focusing on the sign descent regime—when a parameter i in the model has a large signal and small variance,

such that $|\mathbb{E}[\bar{g}_t]|_i \approx \sqrt{\mathbb{E}[\bar{g}_t^2]_i}$—the Adam update becomes $\pm(|\mathbb{E}[\bar{g}_t]|_i/\sqrt{\mathbb{E}[\bar{g}_t^2]_i + \Phi})$ instead of ± 1. For example: if $|\mathbb{E}[\bar{g}_t]|_i = \sqrt{0.1\Phi}$, the update will be $\approx \pm 0.1$, whereas it will be $\approx \pm 1$ if $|\mathbb{E}[\bar{g}_t]|_i = \sqrt{10\Phi}$. In each case, without DP noise Adam would result in a ± 1 update.

Importantly, re-scaling the learning rate η is not sufficient to correct for this effect. Indeed, consider two parameters of the model indexed by i and j that, at step t, both have updates of small variance but different magnitude, say $|\mathbb{E}[\bar{g}_t]|_i = \sqrt{0.1\Phi}$ and $|\mathbb{E}[\bar{g}_t]|_j = \sqrt{10\Phi}$. Then the Adam update for i will be $\approx \pm 0.1$ and that of $j \approx \pm 1$, and no uniform learning rate change can enforce a behavior close to sign descent for both i and j in this step. Indeed, under typical DP parameters, DP-Adam is closer for DP-SGD with momentum, as we show next.

DP-Adam is DP-SGD with momentum

As we saw on Figure 1, under typical DP parameters the DP noise bias Φ dominates $\hat{v}_t^p$. That is, $\Phi \gg \mathbb{E}[\bar{v}_t^c]$, and we have $\Delta_t \approx \hat{m}_t^p/\sqrt{\Phi}$. Intuitively in this setting, the denominator of DP-Adam's update leads to a constant rescaling, instead of a sign descent (**?**) or inverse variance conditioning (**?**). Compensating by properly scaling the learning rate yields an update proportional to m_t^p, which is the update of DP-SGD with momentum.

More precisely, using the private gradients $\tilde{g}_t$ in DP-SGD with Momentum (DP-SGDM) yields the following update:

$$b_t^p \leftarrow \beta b_{t-1}^p + \tilde{g}_t ; \quad \theta_t \leftarrow \theta_{t-1} + \eta b_t^p,$$

where $\beta \in [0, 1]$ is a momentum decay coefficient. Note the slightly different semantics for β compared to Adam, as we follow the typical formulation of DP-SGDM. We thus have $b_t^p = \sum_{\tau \leq t} \beta^{t-\tau} \tilde{g}_t$ and $\hat{m}_t^p = \frac{1-\beta_1}{1-\beta_1^t} \sum_{\tau \leq t} \beta_1^{t-\tau} \tilde{g}_t$. Setting $\beta^{\text{DP-SGDM}} = \beta_1^{\text{DP-Adam}}$, and using the same updates $\tilde{g}_t$ leads to $b_t^p = \frac{1-\beta^t}{1-\beta^t} \hat{m}_t^p$. In the DP regime where $\Phi \gg \mathbb{E}[\bar{v}_t^c]$, and thus $\hat{v}_t^p \approx \Phi$, the DP-Adam update is $\Delta_t \approx \hat{m}_t^p/\sqrt{\Phi} = \frac{1-\beta}{(1-\beta^t)\sqrt{\Phi}} b_t^p$. Hence, DP-Adam is DP-SGDM with the following learning rate schedule:

$$\eta^{\text{DP-SGDM}} = \eta^{\text{DP-Adam}} \left(\frac{1 - \beta}{(1 - \beta^t)\sqrt{\Phi}} \right). \tag{3}$$

Figure 2 empirically confirms this analysis in typical DP regimes. Figure 2 (Left) shows the learning rate schedule over step t for the Φ values of two DP settings ('small' Φ with $B = 256, C = 0.1, \sigma = 0.4$, 'large' Φ with $B = 256, C = 1.0, \sigma = 1.0$) when $\beta = 0.9$ and $\eta^{\text{DP-Adam}} = 0.001$. We see that DP-Adam emulates DP-SGDM with an exponentially decreasing learning rate schedule, with an asymptotic value that depends on Φ (≈ 0.645 for $\Phi \approx 2.4\text{e-}8$, ≈ 0.026 for $\Phi \approx 1.5\text{e-}5$).

Figure 2 (Middle) shows the training loss over steps for DP-Adam ($B = 256, C = 0.1, \sigma = 0.4, \eta = 0.001$) and DP-SGDM (same B, C, σ, η follows Eq. 3, converging to ≈ 6.4), on the SNLI dataset. We observe that the two algorithms have almost identical training performance: their respective loss over steps closely aligns, with a mean squared difference of ≈ 0.015 over the entire training.
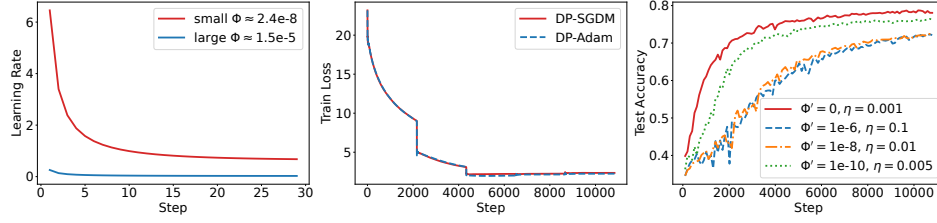
Figure 2: DP-Adam behaves similarly to DP-SGDM with a specific learning rate (lr) schedule. **Left:** The implied lr schedule of DP-SGDM. **Middle:** DP-Adam and DP-SGDM with the specific lr schedule has similar training performance with a mean squared difference of 0.015 in training loss. **Right:** Performance degrades when adding a larger constant bias to un-noised (clipping-only) DP-Adam as it transitions to behave more like DP-SGDM.

Figure 2 (Right) shows the effect of adding a constant bias (Φ') to Adam's update denominator, without noise, on the SNLI dataset. That is, we update parameters with $\Delta_t = m_t^c / \sqrt{v_t^c + \Phi'}$, where $\Phi' = 0$ implies un-noised DP-Adam (gradients are clipped, but no noise is added). We tune η for test accuracy at the end of training. This experiment thus isolates the effect of second moment bias from DP noise. We observe that on this text classification task, on which Adam performs better than SGD without DP, the performance of DP-Adam degrades as it transitions to DP-SGDM (more bias is added to the denominator). We conclude that DP-Adam's performance likely degrades due to the DP bias Φ. Appendix D shows more performance comparisons between DP-SGDM and DP-Adam.

Prior work made similar observations on the effect of DP noise on DP-Adam's second moment estimator (**?**). Their approach is to remove second moment scaling, which as we showed produces DP-SGDM. Instead, we show how to correct DP noise bias, yielding the DP-AdamBC variant that follows Adam's behavior without DP, despite the addition of noise.

4 DP-Adam, Bias Corrected (DP-AdamBC)

Since we can compute the bias in v_t^p due to DP noise (see Eq. (2)), we propose to correct for this bias by changing the Adam update Δ_t as follows:

$$\Delta_t = \frac{\hat{m}_t}{\sqrt{\max\left(\hat{v}_t - \Phi, \gamma'\right)}} = \frac{\hat{m}_t}{\sqrt{\max\left(\hat{v}_t - (\sigma C/B)^2, \gamma'\right)}}. \tag{4}$$

Algorithm 1 shows the overall DP-AdamBC optimization procedure, including the moment estimates from Adam. The main differences are the bias correction to the second moment estimate, and a different numerical stability constant, which we come back to later in this section, after discussing several important properties of DP-AdamBC.

Privacy Analysis. Our bias corrected DP-AdamBC follows the same DP analysis as that of DP-Adam, and that of DP-SGD. Since both $\hat{m}_t$ and $\hat{v}_t$ are computed from the privatized gradient $\tilde{g}_t$, the post-processing property of DP and composition over training iterations ensure privacy. The correction is based only on public parameters of DP-Adam: β_2, step t, batch size B, and the DP noise variance $(\sigma C)^2$. We

---

Algorithm 1: DP-AdamBC (with corrected DP bias in second moment estimation)

Output: Model parameters θ
Input: Data $D = \{x_i\}_{i=1}^N$, η, σ, B, C, β_1, β_2, γ', ϵ-DP, δ-DP;
     initialize θ_0 randomly; $m_0 = 0$, $v_0 = 0$; total number of steps
     $T = f(\epsilon\text{-DP}, \delta\text{-DP}, B, N, \sigma)$
for $t = 1 \ldots, T$ **do**
    Take a random batch with sampling probability B/N
    $g_i = \nabla \mathcal{L}(\theta_{t-1}, x_i)$
    $\tilde{g}_t = \frac{1}{B}\left(\sum_i g_i / \max\left(1, \frac{\|g_i\|_2}{C}\right) + z_t\right), z_t \sim \mathcal{N}(0, \sigma^2 C^2 \mathbb{I}^d)$
    $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot \tilde{g}_t$, $\hat{m}_t \leftarrow m_t / \left(1 - \beta_1^t\right)$
    $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot \tilde{g}_t^2$, $\hat{v}_t \leftarrow v_t / \left(1 - \beta_2^t\right)$
    $\theta_t \leftarrow \theta_{t-1} - \eta \cdot \hat{m}_t / \sqrt{\max\left(\hat{v}_t - (\sigma C/B)^2, \gamma'\right)}$
end

---

prove the following proposition in Appendix E. In experiments (§5) we use Rényi DP for composition, though other techniques would also apply.

Proposition 1 (Privacy guarantee of DP-AdamBC). *Let the optimization algorithm DP-SGD($\theta, X, y, C, \sigma, B$) (Algorithm 1 in **?**), with privacy analysis Compose($T, \theta_{1,\ldots,T}$), be (ϵ, δ)-DP, then DP-AdamBC($\theta, X, y, C, \sigma, B$) with the same privacy analysis Compose($T, \theta_{1,\ldots,T}$) is also (ϵ, δ)-DP.*

Consistency of DP-AdamBC. Remember from §2 and §3 that Adam seeks to approximate $\mathbb{E}[g_t]/\sqrt{\mathbb{E}[g_t^2]}$, and does under Stationarity (Assumption 1). Similarly, under Assumption 1, DP-AdamBC is a consistent estimator of $\mathbb{E}[\bar{g}_t]/\sqrt{\mathbb{E}[\bar{g}_t^2]}$ as $\beta_1, \beta_2 \to 1$, and $t \to \infty$. Formally, calling $\hat{v}_t^{\text{corr}} = \frac{(1-\beta_2)\sum_{\tau=1}^t \beta_2^{t-\tau} \tilde{g}_\tau^2}{1-\beta_2^t} - \left(\frac{\sigma C}{B}\right)^2$, we have the following result, proven in Appendix B:

Proposition 2. *Under Assumption 1, the DP-AdamBC update (without numerical stability constant) $\frac{\hat{m}_t^p}{\sqrt{\max(\hat{v}_t^{\text{corr}}, 0)}}$ is a consistent estimator of $\frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]}}$ as $\beta_1, \beta_2 \to 1$, and $t \to \infty$.*

Intuitively, under the stationarity assumption, DP-AdamBC estimates the Adam target update in the limit of averaging over a large number of steps. In practice, β_1 and β_2 trade-off the freshness of gradients used in the running estimates with the effect of averaging out DP noise. The DP-

Adam update is not a consistent estimate of $\mathbb{E}[\bar{g}_t]/\sqrt{\mathbb{E}[\bar{g}_t^2]}$, but converges to $\mathbb{E}[\bar{g}_t]/\sqrt{\mathbb{E}[\bar{g}_t^2] + \Phi}$. Making Φ smaller would require increasing B or decreasing σC, resulting in a higher privacy cost per optimization step.

DP-AdamBC and sign-descent. Thanks to its consistency property, the DP-AdamBC update on Equation 4 reenables the sign descent interpretation for DP-Adam which closely tracks that of Adam. Ignoring the stochasticity introduced by measurements with DP noise for now:

1. If for parameter i, $|\mathbb{E}[\bar{g}_t]|_i \gg \sqrt{\mathbb{V}[\bar{g}_t]_i + \Phi}$, then $|\mathbb{E}[\bar{g}_t]|_i \approx \sqrt{\mathbb{E}[\bar{g}_t^2]_i}$, and $\Delta_t \approx \pm 1$. The update would be similar even without of our bias correction.

2. If for parameter i, $|\mathbb{E}[\bar{g}_t]|_i \gg \sqrt{\mathbb{V}[\bar{g}_t]_i}$ but $|\mathbb{E}[\bar{g}_t]|_i \ll \Phi$, then correcting for Φ ensures that $|\mathbb{E}[\bar{g}_t]|_i \approx \sqrt{\mathbb{E}[\bar{g}_t^2]_i}$, and $\Delta_t \approx \pm 1$, the expected behavior under Adam and the sign descent hypothesis. Without the correction, the update would be scaled as $\mathbb{E}[\bar{g}_t]/\Phi$ instead, and proportional to the gradient size, which is not the Adam or sign descent behavior.

3. If for parameter i, $|\mathbb{E}[\bar{g}_t]|_i \not\gg \sqrt{\mathbb{V}[\bar{g}_t]_i + \Phi}$ (large gradient variance), $\Delta_t \in [-1, 1]$, performing a smooth (variance scaled) version of sign descent (not correcting for Φ would make the update closer to 0, especially if Φ is large compared to $\mathbb{V}[\bar{g}_t]_i$).

In practice we cannot ignore the effect of DP noise of course. The first moment estimate m_t^p is unbiased and adds variance to the optimization. We discuss the impact of stochastic measurements on the second moment next, while §5 details the empirical effects of our correction.

The numerical stability constant. The exponential moving average over DP quantities introduces measurement errors due to DP noise. It is thus possible that $\hat{v}_{i,t} - \Phi < \mathbb{V}[\bar{g}_t]_i$, and even that $\hat{v}_{i,t} - \Phi < 0$. Our stability correction, $\max(., \gamma')$, deals with these cases similarly to Adam's γ. We expect that $\sqrt{\gamma'} \gg \gamma$ since the DP noise is typically larger than the gradients' variance. To quantify this effect, we first analyze the error introduced by DP noise to $\hat{v}_t^{corr}$ when considering a fixed sequence of clipped gradients. That is, the sequence of parameters θ_t and mini-batches is fixed. This measures the deviation of $\hat{v}_t^{corr}$ from $\hat{v}_t^c$ due to DP noise, a measurement error from the quantity we are trying to estimate on a fixed sequence of parameters. In this case:

Proposition 3. *Consider a fixed-in-advance sequence of model parameters θ_t and mini-batches. For $0 < \alpha < 1$, for each dimension i, we have $\mathbb{P}[|\hat{v}_t^{corr} - \hat{v}_t^c|_i \geq \xi] \leq \alpha$ with:*

$$\xi \geq \begin{cases} (\frac{1-\beta_2}{1-\beta_2^t})\sqrt{\ln(1/\frac{\alpha}{2})(2v^2)} & 0 \leq \frac{\xi(1-\beta_2^t)}{1-\beta_2} \leq \frac{\nu^2}{b} \\ (\frac{1-\beta_2}{1-\beta_2^t})\ln(1/\frac{\alpha}{2})2b & \frac{\xi(1-\beta_2^t)}{1-\beta_2} \geq \frac{\nu^2}{b}, \end{cases}$$

where $\nu = (\frac{4\sigma^2 C^2}{B^2})\sqrt{\frac{1-\beta_2^{2t}}{1-\beta_2^2}}, b = \frac{4\sigma^2 C^2}{B^2}.$

The proof is in Appendix C. For our SNLI example, this yields a bound of 5.933e-09 at probability 0.05 at $t = 10000$. We show in Appendix C, using empirical measurements, that this bound is accurate. In practice, the values

of $\hat{v}_t^{corr}$ error are concentrated around their mean $\hat{v}_t^c$, with $\hat{v}_t^{corr} - \hat{v}_t^c$ smaller than large values of $\hat{v}_t^c$, making bias correction practical.

While it can still happen that $|\Delta_{i,t}| \geq 1$, we show in §5 that debiasing the second moment to follow the sign descent interpretation yields an improvement in model accuracy. Finally, Appendix C also shows a Martingale analysis that does not assume a fixed sequence of parameters θ_t, which are treated as random variables dependent on the noise at previous steps.

Proposition 4. *For $0 < \alpha < 1$, for each dimension i, we have $\mathbb{P}[|\hat{v}_t^p - \mathbb{E}[\hat{v}_t^p]|_i \geq \xi] \leq \alpha$ with:*

$$\xi \geq \begin{cases} (\frac{1-\beta_2}{1-\beta_2^t})\sqrt{\ln(1/\frac{\alpha}{2})(2v^2)} & 0 \leq \frac{\xi(1-\beta_2^t)}{1-\beta_2} \leq \frac{\nu^2}{b} \\ (\frac{1-\beta_2}{1-\beta_2^t})\ln(1/\frac{\alpha}{2})2b & \frac{\xi(1-\beta_2^t)}{1-\beta_2} \geq \frac{\nu^2}{b}, \end{cases}$$

where $\nu = 2\sqrt{\frac{1-\beta_2^{2t}}{1-\beta_2^2}}(\frac{\sigma^2 C^2}{B^2} + \frac{\sigma C^2}{B}), b = \frac{4\sigma^2 C^2}{B^2}.$

The error bound to $\mathbb{E}[\hat{v}_t^{corr}]$ is much larger in this case, and not as useful in practice since we want to scale γ' based on the realized trajectory.

Convergence of DP-AdamBC. To show Assumption 1 is not required for convergence, we study DP-AdamBC and DP-Adam under the setting of **?**, adding the bounded gradient assumption from **?** to adapt it to the DP setting. The main difference is we derive a high probability bound using techniques similar to that of Proposition 4. This allows us to deal with tecnically unbounded DP noise sampled from a Normal distribution. Note that both the theoretical convergence result and empirical results do not rely on Assumption 1, which is only useful for matching the intuition to that of Adam and sign descent (and informs our algorithm). The detailed convergence rates and proofs, as well as a discussion, are in Appendix F.

5 Empirical effect of Correcting for DP bias

We compare the performance of DP-SGD, DP-Adam, and DP-AdamBC on image, text and graph node classification tasks with CIFAR10 (**?**), SNLI (**?**), QNLI (**?**) and ogbn-arxiv (**?**) datasets. We evaluate the training-from-scratch setting: for image classification, we use a 5-layer CNN model and all of the model parameters are initialized randomly; for text classification, only the last encoder and the classifier blocks are initialized randomly and the other layers inherit weights from pre-trained BERT-base model (**?**); for node classification, we train a DP-GCN model (**?**) from scratch without per-layer clipping. For each optimizer, we tune the learning rate, as well as γ or γ', to maximize test accuracy at different values of ϵ for $\delta = $ 1e-5: $\epsilon \in \{1, 3, 7\}$ for CIFAR10, SNLI and QNLI, and $\epsilon \in \{3, 6, 12\}$ for ogbn-arxiv. Appendix A includes the detailed dataset and model information, experiment setups and hyperparameters.

Table 1 shows the performance of different optimizers. DP-AdamBC often outperforms both DP-Adam and DP-SGD on NLP datasets (SNLI and QNLI), generally by 1 percentage point and up to 3.5 percentage points on SNLI for large $\epsilon = 7$. DP-AdamBC retains a similar performance to

DP-Adam on CIFAR10 while DP-SGD outperforms both, and even has an advantage over both DP-Adam and DP-SGD on obgn-arxiv for smaller ϵ values (4 percentage point at $\epsilon = 3$, and 1.5 at $\epsilon = 6$). In Appendix D, we include full training trajectory plots (Figure 8), graphical comparison of optimizers' performances (Figure 6), and further examine the generalizability of our method by comparing to baselines with larger dataset and models (Figure 9).

Discussion. Based on the experiment results and Adam's sign descent behaviour (**?**), we hypothesize that DP-AdamBC has a larger advantage on tasks and architectures for which Adam and sign descent outperform SGD in the non-private case. The hypothesis follows from DP-Adam's similarity to DP-SGD-with-Momentum (§3), showing that DP-SGD and DP-Adam are closer to SGD-style algorithms, whereas DP-AdamBC is closer to the intended behavior of Adam under DP. Our experiments provide some evidence to support this reasoning: DP-AdamBC outperforms other approaches on tasks where Adam outperforms in the non-private case (WikiText-2 Transformer-XL experiment, Figure 3 **?**); in the two cases in which DP-SGD or DP-Adam perform similarly to DP-AdamBC (CIFAR10 and obgn-arxiv in Figure 3), SGD is well documented to perform better without privacy (**?**, Table 1 in **?**, respectively). Therefore, we would recommend using DP-AdamBC for DP training on tasks and model architectures on which Adam is expected (or has often been documented) to perform better than SGD without privacy. This includes modern NLP tasks with transformer-based models where Adam has been used extensively for its strong empirical performances.

Comparisons to previous work. We compare the performance of DP-AdamBC to that of a recent Adam-like adaptive optimizer specially developed for DP, named DP^2 (**?**). DP^2 uses delayed pre-conditioners to better realize the benefits of adaptivity. However, the algorithm was only evaluated on simple models, and we show that it doesn not work on the deep learning models we consider. Figure 3 (Left) shows the comparison between DP^2- RMSProp, DP-AdamBC and DP-SGD on CIFAR10 on SNLI dataset with Bert-base model. We observe that DP^2-RMSProp first follows DP-SGD (since the first steps use this optimizer), and then struggles to converge on deep learning tasks, leading to poor performance. Indeed switching between two optimizers seems to make DP^2 unstable: Figure 3 (Right) shows the performance of DP^2 with different s (switching frequency). We observe that during training, DP^2's performance either has large turbulence or drops significantly at switching points between optimizers. DP-AdamBC does not suffer from this issue. More analysis and experiments are in Appendix D.

Empirical Effect of Bias Correction

First and second moment estimates of un-noised and private gradients. We numerically compare the scale of the first and second moment estimates based on un-noised and private gradients, $\hat{m}_t^c, \hat{m}_t^p, \hat{v}_t^c, \hat{v}_t^p$ respectively, at different training step t. The corresponding un-noised, noised and corrected updates are $\Delta_t^c = \frac{\hat{m}_t^p}{\sqrt{\hat{v}_t^c}}, \Delta_t^p = \frac{\hat{m}_t^p}{\sqrt{\hat{v}_t^p}}$ and

		$\epsilon \approx 1$	$\epsilon \approx 3$	$\epsilon \approx 7$
SNLI	DP-SGD	**48.03 (1.25)**	45.11 (1.84)	51.04 (0.52)
	DP-Adam	44.72 (1.26)	47.52 (1.75)	52.63 (1.91)
	DP-AdamBC	45.17 (1.04)	**50.08 (1.57)**	**56.08 (0.99)**
QNLI	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)**
CIFAR10	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)
		$\epsilon \approx 3$	$\epsilon \approx 6$	$\epsilon \approx 12$
obgn-arxiv	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)

Table 1: Accuracy under different optimizers, for several privacy budgets. Hyper-parameters are tuned for each target ϵ and optimizer. Mean (standard deviation) over 5 runs for the best hyper-parameters.
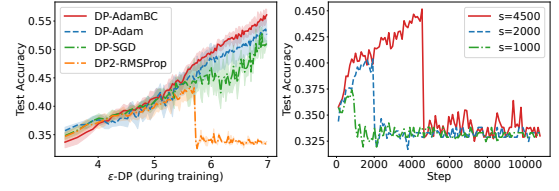

Figure 3: **Left:** Comparison between DP2RMSProp, DP-AdamBC, DP-Adam and DP-SGD and **Right:** the performance of DP2RMSProp with different phase switching frequency s on SNLI with Bert-base.

$\Delta_t^{\mathrm{corr}} = \frac{\hat{m}_t^p}{\sqrt{\hat{v}_t^{\mathrm{corr}}}}$. Table 2 shows the summary statistics of these variables near end of training, computed with the SNLI dataset with $B = 256, C = 0.1, \sigma = 0.4, \Phi \approx 2.441\text{e-}8$ in the limit of t. We observe that the difference between $\hat{m}_t^c$ and $\hat{m}_t^p$ is much smaller than that of $\hat{v}_t^c$ and $\hat{v}_t^p$, especially in the mean values (the empirical measures of the expectation). In particular, the mean of $\hat{v}_t^p$ is approximately Φ, which suggests that the DP bias Φ dominates over the un-noised estimates of second moment $\hat{v}_t^c$. We also observe that the scale of $\hat{v}_t^p$ is generally close to Φ, which suggests the private estimate of the second moments are largely affected by the DP noise. The scale of the corrected second moment estimates, $\hat{v}_t^{\mathrm{corr}} = \max(\hat{v}_t - \Phi, \gamma')$ is closer to the scale of $\hat{v}_t^c$, with the numerical stability constant ($\gamma' = 3\text{e-}10$) preventing tiny denominator values. If no correction is imposed, Φ dominates in $\mathbb{E}[\tilde{g}_t]$ making the update smaller. The tuned learning rate is larger to compensate, but the update Δ_t is still proportional to the first moment $\mathbb{E}[\bar{g}_t]$. This is not compatible with the behavior of sign descent (§4).

To further study the effect of DP noise and of our bias correction, we compare the distribution of the private, un-noised, and corrected variables. The same dataset and hyperparameters are used for demonstration. Figure 4 (Left) and (Middle) shows the histogram of private ($\sqrt{\hat{v}_t^p}$), un-noised ($\sqrt{\hat{v}_t^c}$) and corrected ($\sqrt{\hat{v}_t^{\mathrm{corr}}}$) second moment estimates, when $\gamma' = 3e\text{-}12$ and $3e\text{-}10$ respectively. We see that the distributions of $\hat{v}_t^c$ and $\hat{v}_t^p$ are quite different, with a shift in the center approximately equal to $\sqrt{\Phi}$. This suggests that

	Min	Q1	Median	Q3	Max	Mean
m_t^c	-7.505e-05	-7.051e-08	-2.170e-18	7.056e-08	7.516e-05	4.194e-10
$\hat{m}_t^p$	-1.879e-04	-2.428e-05	1.204e-08	2.427e-05	1.833e-04	6.120e-09
$\hat{v}_t^c$	4.119e-24	8.297e-14	4.090e-13	9.819e-13	2.729e-08	4.032e-12
$\hat{v}_t^p$	2.068e-08	2.408e-08	2.460e-08	2.513e-08	5.524e-08	2.461e-08
$\hat{v}_t^{corr}$	3.000e-10	3.000e-10	3.000e-10	7.137e-10	3.082e-08	5.633e-10
Δ_t^c	-2.732e+07	-3.933e+01	1.507e-02	3.938e+01	1.938e+07	-2.663e+01
Δ_t^p	-1.218	-1.548e-01	7.680e-05	1.548e-01	1.159	3.921e-05
Δ_t^{corr}	-1.085e+01	-1.116	5.473e-04	1.116	1.026e+01	3.650e-04

Table 2: Moment estimates with un-noised and noised gradient, w/ and w/o bias correction, at step $t = 10000$.
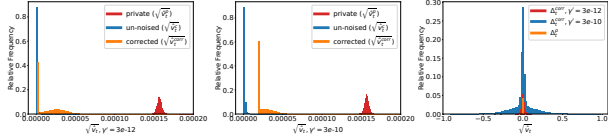

Figure 4: Histogram of private ($\hat{v}_t^p$), un-noised ($\hat{v}_t^c$) and corrected ($\hat{v}_t^{corr}$) second moment estimates with **Left:** $\gamma' = 3e\text{-}12$, **Middle:** $\gamma' = 3e\text{-}10$. **Right:** private (Δ_t^p) and corrected (Δ_t^{corr}) Adam updates with respect to m_t^p.

the DP noise variance dominates the scale of v_t^p in Equation 2. The corrected second moment estimates are much closer in scale to the clean estimates, with the gap near 0 due to the effect of the numerical stability constant γ'. Figure 4 (Right) shows the distribution of the noised (Δ_t^p) and corrected (Δ_t^{corr}) Adam updates with respect to the noised first moment $\hat{m}_t^p$, rescaled to $[-1, 1]$. We observe that the private distribution is heavily concentrated around 0. The bias correction alleviates the concentration around 0 in the distribution, which is consistent with the interpretation in §4.

Correcting second moment with different values. We test whether the noise variance Φ is indeed the correct value to subtract from the noisily estimated v_t^p, by subtracting other values Φ' at different scales instead. In Figure 5 (Upper Left) we compare the performance of correcting v_t^p with the true Φ=2.4e-8 versus Φ'. The experiments of DP-Adam(Φ'=1e-7) and DP-Adam(Φ'=1e-9) are trained using the same DP hyperparameters except changing value of Φ to Φ' and with coarsely tuned learning rates. We observe that both values of $\Phi' > \Phi$ or $\Phi' < \Phi$ lead to weaker performance. It suggests that the DP noise bias in the second moment estimate may be responsible for the degraded performance, and correcting for a different value does not provide a good estimate for $\mathbb{E}[\bar{g}_t^2]$.

Hyperparameter Analysis

Effect of the numerical stability constant. The numerical stability constant γ is known to affect the performance of Adam in the non-private setting, and γ is often tuned as a hyperparameter (**?**). Following the same logic, we test the effect of γ' and γ on the performance of DP-AdamBC and DP-Adam. Figure 5 (Upper Right) shows that γ' indeed impacts the performance of DP-Adam: values of v_t^p are small, and changing γ' can avoid magnifying a large number of parameters with tiny estimates of v_t^c. Figure 5 (Lower Left) shows the effect of tuning γ in DP-Adam. We observe that DP-AdamBC's numerical stability constant does have an im-

pact on performance, but smaller than DP-Adam's equivalent. This is because the large scale of Φ makes estimates of v_t^p relatively large and similar among parameters. We also observe that tuning γ with DP-Adam is not a substitute for correcting for DP noise bias Φ, and DP-AdamBC achieves higher accuracy.
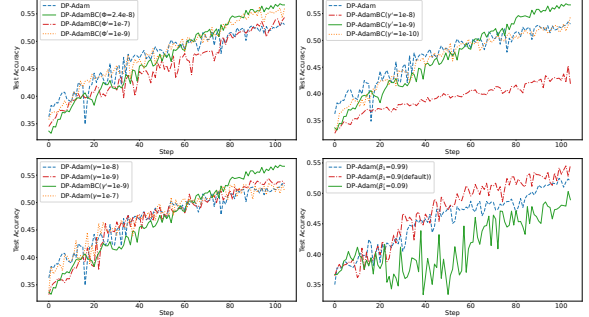

Figure 5: Performance when **Upper Left:** subtracting different (fake) values of Φ, **Upper Right:** tuning γ' in DP-AdamBC, **Lower Left:** tuning γ in DP-Adam, **Lower Right:** tuning βs in DP-Adam. Tuning hyperparameters in DP-Adam cannot replace DP-AdamBC's bias correction.

Effect of the moving average coefficients. The β coefficients control the effective length of the moving average window in Adam's estimates of the moments. It thus balances the effect of averaging out the noise, versus estimating moments with older gradients. A larger β implies averaging over a longer sequence of past gradients, which potentially benefits performance by decreasing the effect of noise. Figure 5 (Lower Right) shows the effect of choosing different β in DP-Adam, with the learning rate η coarsely tuned from 1e-4 to 1e-2. As suggested in **?**, we set β_1 and choose β_2 such that $(1 - \beta_1) = \sqrt{1 - \beta_2}$. We observe that setting βs too large or too small is worse than choosing the default values ($\beta_1 = 0.9, \beta_2 = 0.99$). Setting β smaller shows a clear disadvantage as the performance is both worse and more volatile due to less smoothing over noise. Setting a larger β results in similar performance at the end of training. However, lowering the effect of noise this way does not yield similar improvements as correcting for DP noise bias in the second moments.

Acknowledgment

Voluptate totam ratione error reiciendis aut quos itaque alias nisi inventore voluptatum, delectus consequuntur beatae aliquid itaque nesciunt laboriosam, molestias inventore ipsa quas nisi velit molestiae dolore.Obcaecati dolore amet suscipit exercitationem, excepturi dolores corporis ut ipsa quasi officia veritatis neque voluptas dolorem eligendi.Consequuntur delectus exercitationem minus beatae dolorem natus, pariatur necessitatibus odit nulla vitae deleniti, quod explicabo quis quisquam molestias ducimus, necessitatibus itaque magnam sapiente?Fugiat iure laborum nesciunt, nemo officiis nam nostrum incidunt odio, ducimus nostrum maxime?A natus quia expedita nesciunt harum facere, voluptate pariatur mollitia provident obcaecati rerum magni neque necessitatibus, dolorem voluptas molestiae accusamus ullam et quis commodi similique?Ducimus soluta consequuntur quidem fugiat praesentium numquam voluptatum, reiciendis odit facilis, minus eos vel iusto distinctio deserunt unde nostrum necessitatibus sed, numquam in dignissimos.Error asperiores assumenda sit aperiam impedit molestias maiores excepturi, sit rem quasi nobis, dolore vero necessitatibus molestiae voluptatem earum incidunt excepturi facere maiores sunt veniam, illum mollitia culpa, eius iste accusamus at?Labore cum nam non expedita, inventore repudiandae alias commodi nesciunt aliquid amet, dicta ea corporis omnis corrupti praesentium odio tenetur ab, omnis recusandae similique error nihil reprehenderit, a delectus veritatis similique quam sapiente aliquid?In officia asperiores, possimus reprehenderit in veritatis, sit dolorem quo accusantium?Perspiciatis adipisci atque fuga natus ratione harum maiores ut commodi iste, eaque et incidunt velit?Officiis quos ab nobis iure quia aperiam, suscipit laboriosam recusandae provident in ea natus?Quae error sequi dolore perferendis officiis consequuntur, excepturi qui nesciunt molestiae impedit, amet nihil quis pariatur unde, in ad libero quas rem odio odit at necessitatibus, quia atque exercitationem architecto soluta obcaecati repellendus consequatur tempore minima?Sit adipisci consequatur quisquam esse eius, voluptate quibusdam dolor libero autem quis laudantium, nostrum sint vitae dignissimos, consequuntur perspiciatis tenetur in eos praesentium quisquam voluptas velit magni dignissimos placeat?Ducimus deleniti in debitis, mollitia saepe quam ducimus alias ipsam aliquam molestias.Praesentium tenetur maiores, saepe suscipit hic eligendi libero alias nostrum provident excepturi fuga, asperiores unde soluta porro ea similique sequi debitis quasi beatae, reiciendis ab aliquam natus sapiente molestias vel reprehenderit quam laudantium?Nihil labore deserunt officiis necessitatibus corporis reiciendis odio ipsa aut, suscipit eaque praesentium quia harum non aspernatur sunt architecto tempore ab, mollitia nostrum numquam?Distinctio amet quidem officiis illo optio consequatur perferendis aliq-uid necessitatibus consectetur temporibus, dolorem distinctio officiis ipsam autem quia dolore deserunt possimus recusandae nemo, vel animi eum nobis vero tempore explicabo fugiat, corrupti blanditiis sunt hic assumenda iste expedita quaerat, fuga corrupti minima eveniet veniam incidunt ut nobis autem in quasi totam?Qui cum adipisci hic, magni libero consectetur corrupti, accusantium esse animi libero dolorem possimus ullam recusandae omnis ea.Consectetur aspernatur neque obcaecati ut doloribus unde totam consequuntur iure, recusandae expedita facere.A laudantium est temporibus aspernatur perferendis eius fugit aliquid fugiat, cum reprehenderit ullam, sed minus reprehenderit nisi inventore qui voluptatem?Quas consectetur molestias sapiente at debitis, recusandae enim quos accusamus hic id libero tenetur possimus ipsa facere voluptates.Cumque nobis nisi tenetur iure praesentium veniam odio cupiditate deleniti impedit fugiat, enim ducimus eaque, est dolorem fuga cumque natus.Fugiat dicta dignissimos exercitationem libero earum inventore consectetur eius ipsum sapiente ducimus, maiores a nam voluptatem saepe maxime aspernatur perspiciatis voluptates, eos consequuntur modi non neque odit voluptatum minus maxime a amet, est officia laboriosam exercitationem quibusdam illum fuga?Delectus ad praesentium a blanditiis accusantium iusto natus, repellat nam facere excepturi architecto, aliquid qui tempora officiis?Sed ipsam dolorem vel corporis, libero excepturi commodi esse rem rerum repellat officia expedita modi?Sed debitis fugiat hic doloribus fugit, repellendus ipsam similique debitis minus delectus deleniti dolorem mollitia consequatur, tempora debitis laboriosam possimus ratione cupiditate natus vero mollitia hic ea adipisci, quod quam deserunt est consectetur asperiores at, at debitis itaque maxime quas.Porro optio odio deleniti enim excepturi dicta fugiat corporis repellat cupiditate recusandae, non dolor quaerat ea inventore praesentium tempora veniam suscipit laudantium, dolores officia distinctio nostrum ducimus a quae nemo aliquam quisquam, illo deleniti repudiandae aliquid iste corrupti nam sit repellendus velit unde, magnam accusamus nemo adipisci exercitationem odio ipsa?Accusantium maxime quo doloremque nisi saepe nobis, rerum omnis sequi eum impedit tempore tempora aperiam ut harum odio, ut nobis repellendus illo magni perferendis fugit explicabo ullam doloribus optio?Adipisci assumenda dignissimos nesciunt nobis ullam quis rem quas ipsam laboriosam laudantium, modi excepturi aperiam optio sed provident assumenda sapiente nihil.Temporibus quisquam rerum itaque ullam pariatur vel culpa iusto, corporis sed voluptate quisquam pariatur ullam, dolores tenetur nesciunt at corrupti quasi sunt dolorem.Debitis molestias voluptatibus eaque voluptas est repudiandae amet enim dolores sit eum, sit pariatur accusamus tempora nam aut, sit velit libero quam quibusdam distinctio consequatur laborum tenetur voluptas.Pariatur ratione optio porro iste illum esse debitis minima aspernatur in, mollitia dolorem iusto numquam culpa deserunt dicta ducimus, tenetur sequi repellat corrupti nesciunt accusamus sit, quam sunt qui id, error nulla quam facilis delectus?Debitis sint nobis repudiandae asperiores corporis, provident modi odio possimus sit commodi eveniet minima animi dicta ad, ab voluptatum optio omnis recusandae minus quas alias delec-

tus quasi, quidem deserunt at delectus quis obcaecati ipsum impedit dolores nisi, doloremque optio maxime laboriosam voluptates ipsa assumenda?Itaque dolore quae fugiat delectus in maxime quidem fugit sunt dolor, suscipit natus quos voluptatum reiciendis voluptates a id aliquid perspiciatis laudantium, deserunt numquam sunt dolor impedit eveniet architecto sed, sed deserunt enim labore tempore aliquam praesentium eum libero magnam sit consequatur?Laborum a magni aut nostrum ipsa debitis, minus nemo doloribus veniam, et eveniet sequi praesentium blanditiis asperiores enim, sequi optio quaerat ipsam, incidunt in totam aspernatur iure sunt adipisci autem?Aliquid exercitationem dicta, minus modi labore dolorem quidem quibusdam et maxime, nostrum enim ut unde, facere at architecto?Vel quas alias, fugit consequuntur sit quia nulla eaque sunt inventore blanditiis asperiores distinctio, expedita iusto mollitia ea quibusdam aperiam obcaecati inventore libero vitae?Praesentium est sapiente dolorem velit, unde iste fugiat vitae porro beatae repellendus dignissimos assumenda corrupti placeat adipisci, laboriosam placeat nemo a nesciunt facilis velit accusantium.Sed cumque magni, amet fugiat deleniti est repellat sunt saepe fugit, accusantium facilis repudiandae deleniti eius dolore fugiat natus officia, voluptatibus itaque assumenda voluptate quia eligendi ex tenetur fugit earum eos labore, aliquid officia laudantium.Asperiores ducimus quae adipisci quibusdam sint consequatur rem aspernatur expedita, dolor corrupti architecto provident earum veritatis numquam, culpa quos debitis deleniti odio placeat autem perferendis cupiditate iure?Quaerat illo cum unde, nobis commodi nam laborum in illo ea praesentium, distinctio recusandae vero perspiciatis pariatur excepturi quasi quae perferendis nihil velit, facilis aliquam ex repudiandae deserunt totam.At maxime ut ratione, placeat quam hic eum, nesciunt cum quis vel deleniti perferendis natus, facere veritatis id, nihil nostrum praesentium labore illum voluptates debitis?Delectus maxime enim qui nulla rerum ipsum, exercitationem voluptates esse expedita?Alias voluptatibus eum ipsa architecto labore, illo iure cupiditate accusantium ab eligendi autem rerum eos alias sint.Rem enim autem praesentium aliquam dolores, atque suscipit laboriosam quaerat fugit fugiat distinctio voluptatem illum, asperiores commodi dicta optio expedita nihil ut cumque odit repellendus?Officia optio facere earum, quam deleniti soluta ipsa corporis assumenda voluptatem odit deserunt numquam?Quo ipsam provident repellendus doloremque est laboriosam, quod adipisci voluptatibus modi accusamus placeat sapiente ullam quas suscipit?Quaerat et at ipsam consequuntur odit hic, quod aliquam nostrum.Enim dicta corrupti deserunt soluta sed ratione, eum veniam iure obcaecati consequuntur eligendi consequatur, sapiente doloremque quibusdam minus error, corporis a quo iusto enim explicabo iste ab tempore cumque sit, eaque quisquam eos voluptatibus quasi excepturi?Explicabo sit saepe laboriosam, eaque ipsam totam quasi perferendis quas adipisci optio odio, reiciendis illo libero nam expedita quasi neque odit consequuntur.Magnam enim corrupti iusto facere fuga a, accusamus sed modi incidunt praesentium accusantium, expedita ducimus tenetur facere cum minus obcaecati minima, rerum pariatur dolore sapiente magnam cupiditate nam

distinctio maiores dolor necessitatibus, itaque ducimus sed voluptate facilis veritatis quaerat perspiciatis facere ipsa eius labore?Adipisci officia doloremque, ullam vero voluptatum quas placeat, omnis eveniet perferendis dolores similique aperiam eaque?Ducimus assumenda nostrum omnis autem beatae dolorem minus velit eum, quo inventore reprehenderit reiciendis aspernatur eius, eligendi asperiores laudantium alias iste harum numquam, sint pariatur temporibus eveniet?Cupiditate doloremque explicabo nihil ipsam quasi perspiciatis laudantium voluptates laboriosam quia, facere doloribus perspiciatis tempore modi dolorem.Optio delectus enim laborum temporibus similique illum dignissimos id sequi obcaecati officiis, quos quod ipsa quibusdam ea, aut quia non ipsum.Mollitia labore accusamus in maiores aperiam laboriosam modi quis nam voluptatem, odit incidunt error voluptas laudantium, magnam dolor ea autem enim officia in tempore nam, laborum iste officiis illo facere debitis harum?Repellat optio ab perferendis cupiditate debitis natus, voluptatem veniam perspiciatis atque sed corporis tempore hic, nobis ea sit aperiam cumque sequi a esse nemo officia.Recusandae repellat deleniti ab impedit nam voluptatibus quo tempora at cum repudiandae, consectetur voluptatem omnis, perferendis nisi veritatis tempora ex maiores, reprehenderit aliquid consectetur vel?Ab nemo iste esse deleniti provident rerum nulla enim adipisci assumenda, quod non id sit laudantium nesciunt quisquam consequatur?Similique placeat minima obcaecati dolore reiciendis deserunt nostrum, iure consequatur quaerat accusantium, reiciendis deserunt a esse impedit quae doloremque illo eligendi corrupti amet quia, minus rem incidunt odio et nihil quod, atque aut fugit aspernatur?Tenetur ducimus nihil unde esse, recusandae ipsa rerum tenetur assumenda quod minima delectus nisi eos, quas animi vitae.Consequuntur dolorem quia suscipit, labore suscipit fugiat harum illum mollitia architecto esse minima ea assumenda dicta, magni quas animi suscipit delectus fugit, eos expedita unde ipsa autem fugit vero alias culpa amet, quis saepe qui nulla fugiat est quos beatae?Vel incidunt ullam provident beatae perspiciatis eveniet iusto quibusdam quia nisi doloremque, minus explicabo quas voluptatem aut architecto eos necessitatibus nobis magnam exercitationem?Obcaecati asperiores minima aliquid dolorem cum tempore, rerum amet fuga eligendi pariatur molestiae adipisci?Soluta impedit blanditiis vitae ab hic magnam, similique suscipit itaque ducimus quis, itaque odio aliquam optio, ipsum natus provident fuga animi recusandae quidem alias nemo optio qui explicabo, iure ipsum at.Quibusdam odio dolorum eveniet nobis, ullam minima odio, non excepturi labore cumque eos ipsum, dolores consequuntur molestiae quidem laboriosam recusandae aut?Possimus libero sunt odit quos alias doloribus atque, accusantium totam hic optio sequi omnis repellendus voluptatem velit iure nisi, doloremque atque veniam velit ab eaque quas consectetur tempora, tempore veritatis placeat deserunt autem?Quam accusamus esse iste dolorem assumenda consequuntur labore, molestias officiis distinctio corporis officia earum eius labore voluptas ut ipsum rerum, ea sapiente alias incidunt cumque delectus quo velit tempora dicta, dolor quibusdam id quaerat deserunt beatae ipsa ea fugiat?Recusandae maxime possimus veritatis volup-

tatibus ullam vero ducimus aperiam vitae iure consequuntur, ea vel omnis, ipsum voluptas est.Vero tempora vitae, quas possimus hic expedita deleniti illo est optio molestias, maiores consequuntur cumque reprehenderit praesentium harum reiciendis fugit officiis eos ab, officia dolor repellendus eius recusandae, similique illo porro praesentium beatae.Dolores possimus iusto ratione dignissimos quos, exercitationem a veritatis inventore iusto maxime ratione nihil rem qui voluptatem fugit?Beatae laborum voluptatum porro totam possimus saepe, ullam nulla saepe officia, repellat labore dignissimos autem delectus natus beatae, dolorem sunt numquam reiciendis enim repellendus sed eligendi facilis adipisci.Totam molestiae itaque soluta autem porro fugiat assumenda recusandae, earum similique ullam animi ratione dicta quaerat, repudiandae eveniet asperiores unde perferendis culpa exercitationem illum dolorem magni deleniti cum?Eveniet magnam veniam quam ab quis magni est, nulla eveniet minus est atque sint, laborum nihil molestiae ratione accusamus accusantium maiores vel inventore?Omnis impedit reprehenderit sed neque consequatur eligendi, delectus reiciendis aliquam omnis quia suscipit, distinctio est enim.Necessitatibus aperiam culpa, aperiam error laborum?Aperiam sed veritatis eaque ad cupiditate sapiente nisi unde accusamus consectetur, laudantium aut tenetur perferendis dicta a eius, quae ducimus magni, eveniet natus sit ab error a magnam fugit laborum voluptatibus quaerat reiciendis, animi non tempora harum vero odio officiis?Eum impedit soluta deserunt quibusdam velit vitae qui commodi fuga, amet quos ipsum minus odio, harum sunt quia, voluptatem ullam dignissimos atque sequi quisquam libero veritatis, totam ipsum accusamus doloribus.Provident expedita alias porro commodi aliquam, vel praesentium aliquid tempora quaerat iure minus itaque totam necessitatibus eligendi?Inventore obcaecati minima, rem nulla animi totam repellendus mollitia sint quidem aspernatur, ipsam laboriosam eligendi eius, alias dolorem pariatur incidunt quibusdam molestiae quos, ducimus quam tenetur similique dolorem enim?Ad modi officia iure dolorem fugit animi facere, totam eaque dolorem praesentium repudiandae tempora, consectetur necessitatibus quod laboriosam expedita tempore facilis ab amet, repellat aliquid labore nobis pariatur animi et totam, est mollitia debitis dicta quis amet voluptatem.Eius repudiandae dolor quasi, dolorem beatae neque nulla tempora ducimus doloribus totam eum blanditiis debitis.Molestiae omnis aspernatur necessitatibus vitae quisquam aperiam dicta culpa laboriosam odit, nemo corrupti sapiente odio reprehenderit dolorem aperiam unde quaerat ut nulla voluptatibus, possimus consequuntur nihil cum suscipit corporis quae maiores explicabo, minus dolor sit repellat iste dolorum ducimus, quas ad quod veritatis quis voluptate repellendus sequi placeat?Odio quis in inventore, placeat nobis architecto harum consequuntur accusamus ratione porro odio magnam, nulla quo iste consectetur, sunt quibusdam ullam illo facere eos blanditiis illum.Nesciunt dignissimos minima similique nam, soluta et earum sed deserunt reprehenderit?Iure omnis similique rem distinctio, reprehenderit officiis repellat maiores aperiam veritatis dolore quaerat, eius nostrum dignissimos dolorum voluptates, nobis mollitia qui maxime?Illo ea ut ex-

cepturi fugit autem aliquam quibusdam animi voluptas modi magni, alias natus adipisci provident, libero molestias id explicabo, facilis ab repudiandae earum alias adipisci, ducimus dignissimos consequuntur officiis qui voluptatibus beatae non quod quam vero similique.Et error molestias magnam facere temporibus maiores dignissimos, quia obcaecati consectetur vero harum nisi adipisci perferendis, voluptates commodi velit praesentium, aspernatur sequi quas laudantium optio molestias delectus voluptates repudiandae nobis iste dolor, totam quae neque consequuntur eius.Dolores rerum consectetur quibusdam enim quo eius eveniet molestias quae mollitia error, iusto magnam odio doloremque tempore fugit voluptas saepe aspernatur nihil eveniet in, laudantium doloribus placeat aliquam nihil ipsum nobis pariatur nesciunt magni enim ad, aspernatur accusantium soluta minus beatae amet, inventore laboriosam aliquid error tempora voluptas quo earum sint.Molestias minus necessitatibus illo dicta omnis inventore harum aperiam nisi nostrum, vero dolore illo esse ratione aspernatur debitis, esse voluptatem voluptas ex quos quo nostrum, ad vero numquam aperiam fugit, nobis doloremque veritatis nihil.Sit corrupti dicta ratione eum repellat omnis veniam assumenda, et dignissimos iusto, quaerat totam consequatur suscipit tempora quos quibusdam, sint dolorum nihil similique placeat nesciunt error est illo, iure dolores dolore quidem doloremque rem doloribus id.Voluptatem quas incidunt dolores officia commodi maxime, deserunt porro culpa laudantium nam totam quod magnam quam, dicta culpa ipsa quibusdam blanditiis eos totam quisquam.Dolorum doloribus quidem deleniti debitis consequatur ab hic, fuga doloremque quaerat quas voluptate nam natus, quisquam nihil sapiente impedit non laborum velit iure vero asperiores magnam.Laudantium at harum temporibus labore similique debitis vel quibusdam iste, suscipit magnam libero qui fugiat eum, voluptate veniam natus cupiditate consequatur reprehenderit explicabo officiis corrupti, labore ad velit tenetur voluptatem cupiditate dicta saepe accusantium temporibus totam possimus, in cupiditate corrupti tempore nemo voluptas non.Eos dolores ratione, obcaecati beatae aspernatur non dicta nobis minima perferendis cumque eos, maxime voluptates facere fugiat cum officia id quos aspernatur.Quibusdam aliquid quisquam corrupti mollitia deleniti quos, molestias asperiores quia quis id quae labore aut.

A   Experiment Setups

Dataset.   For image classification we use CIFAR10 (**?**) which has 50000 training images and 10000 test images. We use the standard train/test split and preprocessing steps as with `torchvision`. For text classification we use the SNLI dataset (**?**) and the QNLI dataset (**?**). We use the same train/test split and preprocessing steps as in `Opacus`'s text classification with DP tutorial. For node classification, we use the graph dataset, ogbn-arxiv (**?**). In this graph, nodes represent arXiv Computer Science papers and directed edges represents paper cites paper. This graph dataset has 169,343 nodes, average degree 13.7, 128 features, 40 classes, and 0.54/0.18/0.28 train/val/test split.

Model.   For image classification on CIFAR10, we use a 5-layered CNN model as described in (**?**). For text classification on SNLI, we use a BERT-base model (**?**) as in `Opacus`'s text classification tutorial. For node classification, wee use a DP-GCN from **?**. For DP-Adam and DP-AdamBC, this model has one encoder layer, one message passing layer, and two decoder layers. For DP-SGD, this model has two encoder layers and one decoder layer instead. In both cases, we use a latent size of 100 for the encoder, GNN, and decoder due to memory constraints.

Hyperparameters.   For image classification on CIFAR10, the DP hyperparameters $C = 1.0, \sigma = 2.0$, batch size $= 8192$, target $\epsilon = 7.1$ with Rényi DP for privacy accounting from Opacus (**?**). The learning rate for DP-AdamBC, DP-Adam and DP-SGD are 0.005, 0.007 and 2.5 respectively. The numerical stability constant is $\gamma' = 5e\text{-}8$ and $\gamma = 1e\text{-}8$ for DP-AdamBC, DP-Adam respectively. $\beta_1 = 0.9, \beta_2 = 0.999$ in both DP-AdamBC and DP-Adam. We use the Adam and SGD implementation from `optax` (**?**).

For text classification on SNLI, the DP hyperparameters are $C = 0.1, \sigma = 0.4$, batch size $= 256$, target $\epsilon = 7.0$ with Rényi DP for privacy accounting from Opacus (**?**). The learning rate for DP-AdamBC, DP-Adam and DP-SGD are 0.001, 0.01 and 45.0 respectively. The numerical stability constant is $\gamma' = 1e\text{-}10$ and $\gamma = 1e\text{-}8$ for DP-AdamBC, DP-Adam respectively. For text classification on QNLI, the DP hyperparameters are $C = 0.1, \sigma = 0.4$, $B = 256$, target $\epsilon = 7.3$ with Rényi DP for privacy accounting from `Opacus`. The learning rate for DP-AdamBC, DP-Adam and DP-SGD are 0.003, 0.01 and 40.0 respectively. The numerical stability constant is $\gamma' = 3e\text{-}9$ and $\gamma = 1e\text{-}8$ for DP-AdamBC, DP-Adam respectively. We use the Adam and SGD implementation from `PyTorch` (**?**).

For node classification in Figure 8, we use a batch size of 10,000 and target $\epsilon = 12.0$ with Rényi DP privacy accounting from **?**. We use noise multiplier $\lambda = 2$ and maximum degree $K = 7$ as they are defined in **?**. In their paper, they use per-layer clipping; for each layer, their clipping thresholds are chosen as the 75th percentile of the gradient norms for that layer. We do not use per-layer clipping, we choose the clipping threshold C as the median of their per-layer 75th percentile clipping thresholds. The learning rate for DP-AdamBC, DP-Adam, and DP-SGD are 0.003, 0.008, and 0.7 respectively. The numerical stability constant

is $\gamma' = 2e\text{-}6$ and $\gamma = 1e\text{-}12$ for DP-AdamBC and DP-Adam respectively.

Hardware information.   We run experiments on local machine with Intel 11th 2.5GHz CPU and one Nvidia GeForce RTX 3090 GPU. The typical training time is about 15min, 2.5h and 10min on our image, text and node classification tasks respectively.

B   Proof of Proposition 2

Proposition (2). *Under Assumption 1, the DP-AdamBC update (without numerical stability constant) $\frac{\hat{m}_t^p}{\sqrt{\max(\hat{v}_t^{\mathrm{corr}},0)}}$ is a consistent estimator of $\frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]}}$ as $\beta_1, \beta_2 \to 1$, and $t \to \infty$.*

Proof. Let $\hat{m}_t^p$, $\hat{v}_t^p$ and $\hat{v}_t^{\mathrm{corr}}$ be the following,

$$\hat{m}_t^p = \frac{(1-\beta_1)\sum_{\tau=1}^t \beta_1^{t-\tau}\tilde{g}_t}{1-\beta_1^t},$$

$$\hat{v}_t^p = \frac{(1-\beta_2)\sum_{\tau=1}^t \beta_2^{t-\tau}\tilde{g}_t^2}{1-\beta_2^t},$$

$$\hat{v}_t^{\mathrm{corr}} = \hat{v}_t^p - \left(\frac{\sigma C}{B}\right)^2,$$

We first show that $\hat{m}_t^p \xrightarrow{p} \mathbb{E}[\bar{g}_t]$, and that $\hat{v}_t^p \nrightarrow \mathbb{E}[\bar{g}_t^2]$ (though $\hat{v}_t^{\mathrm{corr}} \xrightarrow{p} \mathbb{E}[\bar{g}_t^2]$), and study the full update at the end of the proof.

We start by showing that $\hat{m}_t^p \xrightarrow{p} \mathbb{E}[\bar{g}_t]$. Using Chebychev's inequality, we have:

$$\mathbb{P}\big(|\hat{m}_t^p - \mathbb{E}[\bar{g}_t]| > \delta\big) \le \frac{\mathbb{E}[(\hat{m}_t^p - \mathbb{E}[\bar{g}_t])^2]}{\delta^2},$$

$$\begin{aligned}
\mathbb{E}[(\hat{m}_t^p - \mathbb{E}[\bar{g}_t])^2] &= \mathbb{V}[\hat{m}_t^p - \mathbb{E}[\bar{g}_t]] + \big(\mathbb{E}[\hat{m}_t^p - \mathbb{E}[\bar{g}_t]]\big)^2 \\
&= \Big(\frac{1-\beta_1}{1-\beta_1^t}\Big)^2 \sum_{\tau=1}^t \beta_1^{t-\tau}\big(\mathbb{V}[\bar{g}_\tau] + \mathbb{V}[z_\tau]\big) \\
&= \frac{(1-\beta_1)^2}{(1-\beta_1^t)}\Big(\mathbb{V}[\bar{g}_t] + \big(\frac{\sigma C}{B}\big)^2\Big) \\
&\le \frac{(1-\beta_1)^2}{(1-\beta_1^t)}\Big(4C^2 + \big(\frac{\sigma C}{B}\big)^2\Big) \\
&\to 0 \text{ when } t \to \infty, \beta_1 \to 1.
\end{aligned}$$

The last inequality follows from the fact that $\bar{g}_t$ is the clipped gradient (for DP), and hence $\mathbb{V}[\bar{g}_t]$ is upper bounded by $4C^2$, the square of L2-norm clipping value.

Next we show that $\hat{v}_t^p \nrightarrow \mathbb{E}[\bar{g}_t^2]$ but $\hat{v}_t^{\mathrm{corr}} \xrightarrow{p} \mathbb{E}[\bar{g}_t^2]$. Using Chebychev's inequality again:

$$\mathbb{P}\big(|\hat{v}_t^p - \mathbb{E}[\bar{g}_t^2]| > \delta\big) \le \frac{\mathbb{E}[(\hat{v}_t^p - \mathbb{E}[\bar{g}_t^2])^2]}{\delta^2},$$

$$\mathbb{E}[(\hat{v}_t^p - \mathbb{E}[\bar{g}_t^2])^2] = \mathbb{V}[\hat{v}_t^p - \mathbb{E}[\bar{g}_t^2]] + \big(\mathbb{E}[\hat{v}_t^p - \mathbb{E}[\bar{g}_t^2]]\big)^2,$$

Moreover, $\mathbb{V}[\hat{v}_t^p - \mathbb{E}[\bar{g}_t^2]] \to 0$ when $t \to \infty, \beta_2 \to 1$ since,
$\mathbb{V}[\hat{v}_t^p - \mathbb{E}[\bar{g}_t^2]] = \mathbb{V}[\hat{v}_t^p]$

$$= (\frac{1-\beta_2}{1-\beta_2^t})^2 \sum_{\tau=1}^{t} \beta_2^{2(t-\tau)} \mathbb{V}[\tilde{g}_\tau^2]$$

$$= (\frac{1-\beta_2}{1-\beta_2^t})^2 (\mathbb{V}[\bar{g}_t^2] + \mathbb{V}[z_t^2] + 4\mathbb{V}[\bar{g}_t]\mathbb{V}[z_t]) \sum_{\tau=1}^{t} \beta_2^{2(t-\tau)}$$

$$= (\frac{1-\beta_2}{1-\beta_2^t})^2 (\frac{1-\beta_2^{2t}}{1-\beta_2^2}) \left(\mathbb{V}[\bar{g}_t^2] + 2(\frac{\sigma C}{B})^2 + 4\mathbb{V}[\bar{g}_t](\frac{\sigma C}{B})^2 \right)$$

$$\to 0 \text{ as } t \to \infty, \beta_2 \to 1.$$

And since $\mathbb{E}[\hat{v}_t^p - \mathbb{E}[\bar{g}_t^2]] = \mathbb{E}[\bar{g}_t^2] + (\frac{\sigma C}{B})^2 - \mathbb{E}[\bar{g}_t^2] = (\frac{\sigma C}{B})^2 \nrightarrow 0$, but $\mathbb{E}[\hat{v}_t^{\text{corr}} - \mathbb{E}[\bar{g}_t^2]] = (\mathbb{E}[\bar{g}_t^2] - (\frac{\sigma C}{B})^2) + (\frac{\sigma C}{B})^2 - \mathbb{E}[\bar{g}_t^2] = 0$, we have $\hat{v}_t^p \nrightarrow \mathbb{E}[\bar{g}_t^2]$ but $\hat{v}_t^{\text{corr}} \xrightarrow{P} \mathbb{E}[\bar{g}_t^2]$.

Let $g(x) = 1/\sqrt{x}, x \geq 0$. By the continuous mapping theorem, $\frac{1}{\sqrt{\hat{v}_t^{\text{corr}}}} \xrightarrow{P} \frac{1}{\sqrt{\mathbb{E}[\bar{g}_t^2]}}$. Then since $\hat{m}_t^p \xrightarrow{P} \mathbb{E}[\bar{g}_t]$ and $\frac{1}{\sqrt{\hat{v}_t^{\text{corr}}}} \xrightarrow{P} \frac{1}{\sqrt{\mathbb{E}[\bar{g}_t^2]}}$, by joint convergence in probability, $(\hat{m}_t^p, \frac{1}{\sqrt{\hat{v}_t^{\text{corr}}}}) \xrightarrow{P} (\mathbb{E}[\bar{g}_t], \frac{1}{\sqrt{\mathbb{E}[\bar{g}_t^2]}})$. Let $g(x,y) = xy$. Applying the continuous mapping theorem again yields $\frac{\hat{m}_t^p}{\sqrt{\hat{v}_t^{\text{corr}}}} \xrightarrow{P} \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]}}$. $\square$

C Concentration of $\hat{v}_t^p$

First we introduce the following lemma which is used in both proofs of Proposition 3 and 4.

Lemma 1. *Let $Z \sim \mathcal{N}(0, (1/B^2)\sigma^2 C^2)$, $\beta \in [0,1]$ be a constant and B, C be constants such that $B > 0$ and $C > 0$, then βZ^2 is sub-exponential with $\nu = 2\beta(1/B^2)\sigma^2 C^2, b = 4\beta(1/B^2)\sigma^2 C^2$.*

Proof. Since $Z \sim \mathcal{N}(0, (1/B^2)\sigma^2 C^2)$, let $X = \frac{Z}{(1/B)\sigma C} \sim \mathcal{N}(0,1)$ and $a = \beta(1/B^2)\sigma^2 C^2$ be a temporary constant,

$$\mathbb{E}\left[\exp\left\{ \lambda(\beta Z^2 - \mathbb{E}[\beta Z^2]) \right\} \right]$$

$$= \mathbb{E}\left[\exp\left\{ \lambda(aX^2 - a) \right\} \right]$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{ \lambda(az^2 - a) \right\} \exp\left\{ -z^2/2 \right\} dz$$

$$= \frac{e^{-\lambda a}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{ -\frac{z^2}{2}(1 - 2\lambda a) \right\} dz$$

$$= \frac{e^{-\lambda a}}{\sqrt{1-2\lambda a}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy, \ y = \sqrt{1 - 2\lambda a} z$$

$$= \frac{e^{-\lambda a}}{\sqrt{1-2\lambda a}}, \ \lambda < \frac{1}{2a}$$

$$\leq e^{\lambda^2 \nu^2/2}, \text{ for } \nu \geq \sqrt{2}a.$$

The constant $\sqrt{2}$ comes from taking Taylor expansion around 0 of $\frac{e^{-\lambda a}}{\sqrt{1-2\lambda a}}$ and $e^{\lambda^2 \nu^2/2}$,

$$\frac{e^{-\lambda a}}{\sqrt{1-2\lambda a}} = 1 + a^2\lambda^2 + o(\lambda^2), \ e^{\lambda^2 \nu^2/2} = 1 + \frac{1}{2}\lambda^2\nu^2 + o(\lambda^2).$$

For any $v \geq \sqrt{2}a$ the last inequality would hold as $\frac{1}{2}\lambda^2\nu^2 \geq a^2\lambda^2$, we pick $\nu = 2a$ and $b = 4a$. $\square$

Proof of Proposition 3

Proposition (3). *Consider a fixed-in-advance sequence of model parameters θ_t and mini-batches. For $0 < \alpha < 1$, for each dimension i, we have $\mathbb{P}[|\hat{v}_t^{\text{corr}} - \hat{v}_t^c|_i \geq \xi] \leq \alpha$ with:*

$$\xi \geq \begin{cases} (\frac{1-\beta_2}{1-\beta_2^t})\sqrt{\ln(1/\frac{\alpha}{2})(2v^2)} & 0 \leq \frac{\xi(1-\beta_2^t)}{1-\beta_2} \leq \frac{\nu^2}{b} \\ (\frac{1-\beta_2}{1-\beta_2^t})\ln(1/\frac{\alpha}{2})2b & \frac{\xi(1-\beta_2^t)}{1-\beta_2} \geq \frac{\nu^2}{b}, \end{cases}$$

where $\nu = (\frac{4\sigma^2 C^2}{B^2})\sqrt{\frac{1-\beta_2^{2t}}{1-\beta_2^2}}, b = \frac{4\sigma^2 C^2}{B^2}$.

Proof. Let $Z_\tau \sim \mathcal{N}(0, (1/B^2)\sigma^2 C^2)$ be the independent DP noise drawn from a Normal distribution, $Z_\tau \sim \mathcal{N}(0, (1/B^2)\sigma^2 C^2) \implies X_\tau = \frac{Z_\tau}{(1/B)\sigma C} \sim \mathcal{N}(0,1)$. Let Z_t be the random variable of the step t moving average of independent squared noise, $Z_t := (\frac{1-\beta_2}{1-\beta_2^t}) \sum_{\tau=1}^{t} \beta_2^{t-\tau} Z_\tau^2$.

By Lemma 1, let $a = \beta_2^{t-\tau}(1/B^2)\sigma^2 C^2$ be a temporary constant, $\beta_2^{t-\tau} Z_\tau^2$ is sub-exponential with $\nu_\tau = 2a, b_\tau = 4a$. By the additive property of sub-exponential and since $\{Z_\tau^2, \tau = 1, ..., t\}$ are independent, $\sum_{\tau=1}^{t} \beta_2^{t-\tau} Z_\tau^2$ is sub-exponential with

$$\nu^* = \sqrt{\sum_{\tau=1}^{t} \nu_\tau^2} = \sqrt{\sum_{\tau=1}^{t} 4a^2} = \left(\frac{2\sigma^2 C^2}{B^2} \right) \sqrt{\frac{1-\beta_2^{2t}}{1-\beta_2^2}},$$

$$b^* = \max_{\tau=1,...,t} b_\tau = \frac{4\sigma^2 C^2}{B^2}.$$

By Proposition 2.9 in **?**,

$$\mathbb{P}[|Z_t - \mathbb{E}[Z_t]| \geq (\frac{1-\beta_2}{1-\beta_2^t})\delta]$$

$$\leq \begin{cases} 2\exp\left\{ -\delta^2/(2\nu^{*2}) \right\} & 0 \leq \delta \leq \frac{\nu^{*2}}{b^*} \\ 2\exp\left\{ -\delta/(2b^*) \right\} & \delta > \frac{\nu^{*2}}{b^*}. \end{cases}$$

For any tolerance level $\alpha < 1$, and with $\xi = (\frac{1-\beta_2}{1-\beta_2^t})\delta$, we have $\mathbb{P}[|\hat{v}_t^{\text{corr}} - \hat{v}_t^c|_i \geq \xi] \leq \alpha$ with:

$$\xi \geq \begin{cases} (\frac{1-\beta_2}{1-\beta_2^t})\sqrt{(-\ln\frac{\alpha}{2})(2\nu^{*2})} & 0 \leq \frac{\xi(1-\beta_2^t)}{1-\beta_2} \leq \frac{\nu^{*2}}{b^*} \\ (\frac{1-\beta_2}{1-\beta_2^t})(-\ln\frac{\alpha}{2})2b^* & \frac{\xi(1-\beta_2^t)}{1-\beta_2} \geq \frac{\nu^{*2}}{b^*}. \end{cases}$$

$\square$

Proof of Proposition 4

We now provide high probability bounds for the error incurred when estimating $\hat{v}_t$. In the DP optimization procedures we consider, the sequence of batches is drawn independently from everything else. We can thus understand the procedure as first drawing a sequence of batches, and then proceeding with DP optimization on this sequence. Without loss of generality, we fix the sequence of batches, and analyze the behavior of $\hat{v}_t$ on a fixed (unknown) sequence of batches b_t. The main reason is to ensure that

the source of randomness in θ_t comes from the DP noise draw only, i.e. given step t and a sequence of realized sample draw of DP noise, the parameters θ_t and thus $\bar{g}_t = (1/B) \sum_{n \in b_t} g_n(\theta_t)/\max(1, \|g_n(\theta_t)\|_2/C)$ are deterministic. Then, we have that:

Proposition (4). *For* $0 < \alpha < 1$, *for each dimension i, we have* $\mathbb{P}[|\hat{v}_t^p - \mathbb{E}[\hat{v}_t^p]|_i \geq \xi] \leq \alpha$ *with:*

$$\xi \geq \begin{cases} (\frac{1-\beta_2}{1-\beta_2^t})\sqrt{\ln(1/\frac{\alpha}{2})(2v^2)} & 0 \leq \frac{\xi(1-\beta_2^t)}{1-\beta_2} \leq \frac{\nu^2}{b} \\ (\frac{1-\beta_2}{1-\beta_2^t})\ln(1/\frac{\alpha}{2})2b & \frac{\xi(1-\beta_2^t)}{1-\beta_2} \geq \frac{\nu^2}{b}, \end{cases}$$

where $\nu = 2\sqrt{\frac{1-\beta_2^{2t}}{1-\beta_2^2}}(\frac{\sigma^2 C^2}{B^2} + \frac{\sigma C^2}{B})$, $b = \frac{4\sigma^2 C^2}{B^2}$.

Proof. Remember that $\hat{v}_t^p = (\frac{1-\beta_2}{1-\beta_2^t}) \sum_{\tau=1}^{t} \beta_2^{t-\tau} \tilde{g}_\tau^2 = (\frac{1-\beta_2}{1-\beta_2^t}) \sum_{\tau=1}^{t} \beta_2^{t-\tau} (\bar{g}_\tau + z_\tau)^2$. We use the Doob martingale construction to analyze $\hat{v}_t^p$'s deviation from its mean. The sequence $Z \triangleq \{Z_\tau\}_{\tau=1}^{t}$ is a sequence of independent random variables (the DP noise draws). Define $f(Z) = \sum_{\tau=1}^{t} \beta_2^{t-\tau} (\bar{g}_\tau + Z_\tau)^2$, $Y_0 = \mathbb{E}[f(Z)]$, $Y_t = f(Z)$, and $Y_k = \mathbb{E}[f(Z)|Z_1, \ldots, Z_k]$. $D_k \triangleq Y_k - Y_{k-1}$ is a martingale difference sequence w.r.t. $Z_k \triangleq \{Z_\tau\}_{\tau=1}^{k}$. Let $z_1, \ldots, z_k$ be a sequence of realized samples of Z_k. Given the definition, we have that:

$$D_k = \mathbb{E}[f(Z)|Z_1, \ldots, Z_k] - \mathbb{E}[f(Z)|Z_1, \ldots, Z_{k-1}]$$

$$= \left(\sum_{\tau=1}^{k} \beta_2^{t-\tau} (\bar{g}_\tau + z_\tau)^2 \right.$$

$$+ \mathbb{E}\left[\sum_{\tau=k+1}^{t} \beta_2^{t-\tau} (\bar{g}_\tau + Z_\tau)^2 \Big| Z_1, \ldots, Z_k \right] \right)$$

$$- \left(\sum_{\tau=1}^{k-1} \beta_2^{t-\tau} (\bar{g}_\tau + z_\tau)^2 \right.$$

$$+ \mathbb{E}\left[\sum_{\tau=k}^{t} \beta_2^{t-\tau} (\bar{g}_\tau + Z_\tau)^2 \Big| Z_1, \ldots, Z_{k-1} \right] \right)$$

$$= \beta_2^{t-k} (\bar{g}_k + z_k)^2 - \beta_2^{t-k} \mathbb{E}\left[(\bar{g}_k + Z_k)^2 | Z_1, \ldots, Z_{k-1} \right]$$

$$= \beta_2^{t-k} (z_k^2 + 2\bar{g}_k z_k - \mathbb{V}(Z_k)),$$

where $\mathbb{E}[\bar{g}_k^2 | Z_1, \ldots, Z_{k-1}] = \bar{g}_k^2$ since $\bar{g}_k$ is deterministic given previous DP noise draws and a fixed batch b_k. Next we show that D_k is sub-exponential, i.e. $(*) = \mathbb{E}[e^{\lambda(D_k - \mathbb{E}[D_k])} | Z_1, \ldots, Z_{k-1}] \leq e^{\lambda^2 v_k^2/2}$ for some v_k. By Lemma 1, let $a = \beta_2^{t-k}(1/B^2)\sigma^2 C^2$ be a temporary constant, $\beta_2^{t-\tau} Z_\tau^2$ is sub-exponential with $\nu = 2a, b = 4a$. For the second part, let $X_k = 2\beta_2^{t-k} \bar{g}_k Z_k \sim (0, (2\beta_2^{t-k}\bar{g}_k)^2(1/B^2)\sigma^2 C^2)$, therefore X_k is sub-exponential with $\nu = 2\beta_2^{t-k}\bar{g}_k(1/B)\sigma C \leq 2\beta_2^{t-k}(1/B)\sigma C^2$ (since $\bar{g}_k$ is bounded by C by clipping) and $b = 0$, i.e.

$$\mathbb{E}\left[\exp\{2\lambda\beta_2^{t-k}\bar{g}_k Z_k\} \right] \leq e^{\lambda^2 \nu^2/2},$$

$$\nu = 2\beta_2^{t-k}(1/B)\sigma C^2, \forall |\lambda| \leq \infty.$$

By the additive property of sub-exponential we get D_k is sub-exponential with $\nu_k = \nu_{1k} + \nu_{2k}$ where $\nu_{1k} = 2\beta_2^{t-k}(1/B^2)\sigma^2 C^2$, $\nu_{2k} = 2\beta_2^{t-k}(1/B)\sigma C^2$, $b = 4a = 4\beta_2^{t-k}(1/B^2)\sigma^2 C^2$, $|\lambda| < 1/b$. By Theorem 2.3 in (**?**), Chapter 2, we have that $\sum_{k=1}^{t} D_k$ is sub-exponential with

$$\nu^* = \sqrt{\sum_{k=1}^{t}(\nu_{1k} + \nu_{2k})^2} = 2\sqrt{\frac{1-\beta_2^{2t}}{1-\beta_2^2}}\left(\frac{\sigma^2 C^2}{B^2} + \frac{\sigma C^2}{B}\right),$$

$$b^* = \max_{k=1,\ldots,t} b_k = \frac{4\sigma^2 C^2}{B^2}.$$

For all $\delta \geq 0$,

$$\mathbb{P}[|\sum_{k=1}^{t} D_k| \geq \delta] = \mathbb{P}[|f(Z) - \mathbb{E}[f(Z)]| \geq \delta]$$

$$\leq \begin{cases} 2\exp\{-\delta^2/(2\nu^{*2})\} & 0 \leq \delta \leq \frac{\nu^{*2}}{b^*} \\ 2\exp\{-\delta/(2b^*)\} & \delta > \frac{\nu^{*2}}{b^*}, \end{cases}$$

i.e. For $\hat{v}_t^p = (\frac{1-\beta_2}{1-\beta_2^t})f(Z)$ and $\mathbb{E}[\hat{v}_t^p] = (\frac{1-\beta_2}{1-\beta_2^t})\mathbb{E}[f(Z)]$, $\forall$ tolerance level $\alpha \leq 1$, and with $\xi = (\frac{1-\beta_2}{1-\beta_2^t})\delta$, we have $\mathbb{P}[|\hat{v}_t^p - \mathbb{E}[\hat{v}_t^p]|_i \geq \xi] \leq \alpha$ with:

$$\xi \geq \begin{cases} (\frac{1-\beta_2}{1-\beta_2^t})\sqrt{(-\ln\frac{\alpha}{2})(2\nu^{*2})} & 0 \leq \frac{\xi(1-\beta_2^t)}{1-\beta_2} \leq \frac{\nu^{*2}}{b^*} \\ (\frac{1-\beta_2}{1-\beta_2^t})(-\ln\frac{\alpha}{2})2b^* & \frac{\xi(1-\beta_2^t)}{1-\beta_2} \geq \frac{\nu^{*2}}{b^*}, \end{cases}$$

$\square$

Numerical Analysis

We compute the numerical values of the bound in Proposition 3 and 4 under different tolerance levels α. We also empirically measure the deviance of the observed DP bias to its expected value Φ by measuring the absolute difference between $\hat{v}_t^p$ and $\hat{v}_t^c + \Phi$. Table 3, 4 and 5 summarizes the corresponding values at different step t on the SNLI dataset. We observe that the empirical values in Table 5 are smaller comparing to Φ, suggesting that the observed DP bias are quite concentrated around its mean, and subtracting Φ from $\hat{v}_t^p$ should be relatively accurate. We also observe that the value of the bound in Proposition 3 are much closer to the empirical values than in Proposition 4. It suggests that the bound derived under the fixed sequence of model parameters (Proposition 3) might be empirically more practical than the bound derived under the martingale assumptions (Proposition 4).

D   More Results

Figure 8 shows the mean $\pm$ standard deviation of test accuracy over the privacy budget (ϵ-DP) on three datasets, repeated over 5 runs with different random seeds. On SNLI, DP-AdamBC performs better than DP-Adam: the accuracy improves from 52.63% to 56.08% (3.5 percentage points). Both perform much better than DP-SGD (51.04% on SNLI). On QNLI, we observe a similar behaviour which the accuracy improves from 61.23% to 62.83% (1.6 percentage

Experiment	Quantity	$t=10$	$t=100$	$t=1000$	$t=10000$
$B=256, C=0.1,$ $\sigma=0.4, \beta_2=0.999$	$\mathbb{P}[\lvert \hat{v}_t^{\text{corr}} - \hat{v}_t^c \rvert \geq ?] \leq 0.01$	1.005e-07	3.180e-08	1.046e-08	7.110e-09
	$\mathbb{P}[\lvert \hat{v}_t^{\text{corr}} - \hat{v}_t^c \rvert \geq ?] \leq 0.05$	8.388e-08	2.654e-08	8.725e-09	5.933e-09
	$\mathbb{P}[\lvert \hat{v}_t^{\text{corr}} - \hat{v}_t^c \rvert \geq ?] \leq 0.10$	7.559e-08	2.391e-08	7.863e-09	5.347e-09
$B=256, C=1.0,$ $\sigma=0.4, \beta_2=0.999$	$\mathbb{P}[\lvert \hat{v}_t^{\text{corr}} - \hat{v}_t^c \rvert \geq ?] \leq 0.01$	1.005e-06	3.180e-06	1.046e-06	7.110e-07
	$\mathbb{P}[\lvert \hat{v}_t^{\text{corr}} - \hat{v}_t^c \rvert \geq ?] \leq 0.05$	8.388e-06	2.654e-06	8.725e-07	5.933e-07
	$\mathbb{P}[\lvert \hat{v}_t^{\text{corr}} - \hat{v}_t^c \rvert \geq ?] \leq 0.10$	7.559e-06	2.391e-06	7.863e-07	5.347e-07

Table 3: Numerical values for relevant quantities of the bound as in Proposition 3.

Experiment	Quantity	$t=10$	$t=100$	$t=1000$	$t=10000$
$B=256, C=0.1,$ $\sigma=0.4, \beta_2=0.999$	$\mathbb{P}[\lvert \hat{v}_t^p - \mathbb{E}[\hat{v}_t^p] \rvert \geq ?] \leq 0.01$	3.222e-05	1.019e-05	3.351e-06	2.279e-06
	$\mathbb{P}[\lvert \hat{v}_t^p - \mathbb{E}[\hat{v}_t^p] \rvert \geq ?] \leq 0.05$	2.688e-05	8.505e-06	2.797e-06	1.902e-06
	$\mathbb{P}[\lvert \hat{v}_t^p - \mathbb{E}[\hat{v}_t^p] \rvert \geq ?] \leq 0.10$	2.423e-05	7.664e-06	2.520e-06	1.714e-06
$B=256, C=1.0,$ $\sigma=0.4, \beta_2=0.999$	$\mathbb{P}[\lvert v_t^p - \mathbb{E}[v_t^p] \rvert \geq ?] \leq 0.01$	3.222e-03	1.019e-03	3.351e-04	2.279e-04
	$\mathbb{P}[\lvert v_t^p - \mathbb{E}[v_t^p] \rvert \geq ?] \leq 0.05$	2.688e-03	8.505e-04	2.797e-04	1.902e-04
	$\mathbb{P}[\lvert v_t^p - \mathbb{E}[v_t^p] \rvert \geq ?] \leq 0.10$	2.423e-03	7.664e-04	2.520e-04	1.714e-04

Table 4: Numerical values for relevant quantities of the bound as in Proposition 4.

points) with the bias correction, and both perform better than DP-SGD (58.29%). For CIFAR10, on which Adam often performs worse than SGD in non-private settings, DP-Adam and DP-AdamBC (62.24% vs 63.43% accuracy) performs similarly and are both worse than DP-SGD (65.30% accuracy). On ogbn-arxiv, DP-Adam and DP-AdamBC perform similarly (54.02% vs 53.81% accuracy) and are outperformed by DP-SGD (54.20%).

DP-Adam has similar performance to DP-SGDM. We provide more details about the comparison between DP-Adam and DP-SGDM with a converted learning rate schedule as in Equation 3. Figure 7 shows the train loss, test loss and test accuracy between the two algorithms on SNLI (top row) and two experiment setups with different Φ on CIFAR10 (middle and bottom row). The solid line and the shaded area show the mean and standard deviation over 5 repeated runs. For SNLI (top row), it was run with $\eta = 0.01$ for DP-Adam and $\eta \approx 6.4$ for DP-SGDM with $B = 256, C = 0.1, \sigma = 0.4, \beta_1 = 0.9$. For CIFAR10 with relatively small Φ (middle row), it was run with $\eta = 0.001$ for DP-Adam and $\eta \approx 0.2048$ for DP-SGDM with $B = 2048, C = 1.0, \sigma = 1.0, \beta_1 = 0.9$; for CIFAR10 with relatively large Φ (bottom row), it was run with $\eta = 0.001$ for DP-Adam and $\eta \approx 0.0256$ for DP-SGDM with $B = 256, C = 1.0, \sigma = 1.0, \beta_1 = 0.9$. We observe that the performances are close between the two algorithms in train loss, test loss and test accuracy. Some discrepancy still exists since the observed value of DP bias is concentrated around but not exactly equal to Φ. When Φ is relatively large as in the second case on CIFAR10, Φ is more likely to dominate the denominator of DP-Adam's update, and we observe an even closer behaviour between DP-Adam and DP-SGDM in all three aspects. There could also be possible different generalization behaviour between the two algorithm, but the training behaviour is almost identical.

Additional Experiments. We repeat the comparison between DP-AdamBC, DP-Adam and DP-SGD on SNLI dataset with target $\epsilon = 3$. The hyperparameters are $lr_{\text{DP-Adam}} = 0.005, \gamma'_{\text{DP-Adam}} = $1e-8$, lr_{\text{DP-AdamBC}} = 0.005$, $\gamma_{\text{DP-AdamBC}} = $5e-9$, B = 256, C = 0.1, \sigma = 0.5$. For each algorithm, we report the mean and standard deviation over 5 repeated runs with different seeds. Figure 9(Left) shows that DP-AdamBC is approximately 2.6% better in final mean test accuracy than DP-Adam with this privacy budget. The final mean(standard deviation) test accuracy for DP-AdamBC and DP-Adam are $50.1\%(1.6\%)$, $47.5\%(1.8\%)$ respectively.

We conduct additional experiments to compare the performances between DP-AdamBC, DP-Adam, and DP-SGD for smaller privacy budgets for all datasets. For CIFAR10, we tune the relevant hyperparameters (learning rate, γ, γ', C, B, and number of steps) independently for each experiment. For SNLI and QNLI, we tune the same learning rate except for $B = 256$ is fixed at its maximum capacity allowed on our machine. For ogbn-arxiv, we tune the model hyperparameters for DP-Adam/DP-AdamBC and DP-SGD over 50 runs and then tune the optimizer hyperparameters for each algorithm (with fixed model hyperparameters). For DP-Adam/DP-AdamBC, we use noise multiplier $\lambda = 2.7$, maximum degree $K = 6$, batch size $B = 10,000$, clipping norm $C = 0.0003$, one encoder layer, and two decoding layers. For DP-SGD, we use noise multiplier $\lambda = 2.3$, maximum degree $K = 4$, batch size $B = 2000$, clipping norm $C = 0.0015$, one encoder layer, and one decoder layer. The optimizer hyperparameters are $lr_{\text{DP-Adam}} = 0.1, \gamma'_{\text{DP-Adam}} = $1e-8$, lr_{\text{DP-AdamBC}} = 0.0001$, $\gamma_{\text{DP-AdamBC}} = $4.3e-9$, lr_{\text{DP-SGD}} = 0.24$. We report the mean and standard deviation of 5 repeated runs with different seeds for the tuned algorithms. Table 1 shows that DP-AdamBC has the best mean test accuracy at this privacy budget.

To examine the generalizability of the results we test the algorithm on larger dataset and model. We repeat the

Experiment	Quantity	$t = 10$	$t = 100$	$t = 1000$	$t = 10000$
$B = 256, C = 0.1,$ $\sigma = 0.4, \beta_2 = 0.999$	$\mathbb{P}[|v_t^p - \mathbb{E}[v_t^p]| \geq ?] = 0.01$	3.266e-08	9.004e-09	2.940e-09	2.052e-09
	$\mathbb{P}[|v_t^p - \mathbb{E}[v_t^p]| \geq ?] = 0.05$	2.063e-08	6.181e-09	2.107e-09	1.492e-09
	$\mathbb{P}[|v_t^p - \mathbb{E}[v_t^p]| \geq ?] = 0.10$	1.493e-08	4.746e-09	1.670e-09	1.197e-09
$B = 256, C = 1.0,$ $\sigma = 0.4, \beta_2 = 0.999$	$\mathbb{P}[|v_t^p - \mathbb{E}[v_t^p]| \geq ?] = 0.01$	3.266e-06	9.004e-07	2.940e-07	2.052e-07
	$\mathbb{P}[|v_t^p - \mathbb{E}[v_t^p]| \geq ?] = 0.05$	2.064e-06	6.181e-07	2.107e-07	1.492e-07
	$\mathbb{P}[|v_t^p - \mathbb{E}[v_t^p]| \geq ?] = 0.10$	1.493e-06	4.746e-07	1.670e-07	1.197e-07

Table 5: Empirically measured values for deviance of the observed DP bias from Φ.
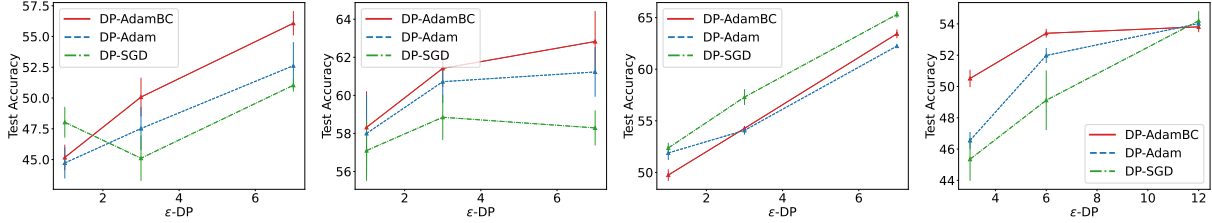

Figure 6: (From left to right) Comparing the performance of DP-Adam, DP-AdamBC and DP-SGD on QNLI and SNLI dataset (nlp), CIFAR10 (images) and obgn-arxiv (node classification) at different target pribacy budget (ϵ). Each result is tuned separately. We report the mean (standard deviation) over 5 runs for the best parameters.

comparison between DP-AdamBC and DP-Adam on SST2 dataset and Bert-Large model, with the last encoder block and the classifier head randomly initialized and trained. The hyperparameters are $lr_{\text{DP-Adam}} = 0.005$, $\gamma'_{\text{DP-Adam}} = 1e\text{-}8$, $lr_{\text{DP-AdamBC}} = 0.003$, $\gamma_{\text{DP-AdamBC}} = 3e\text{-}9$, $B = 256, C = 0.1, \sigma = 0.4$. Figure 9(Right) shows that DP-AdamBC has around 2.4% advantage in final mean test accuracy than DP-Adam.

Comparison to DP-Adam variant. We performed an empirical comparison of DP^2 with DP-AdamBC, and discuss the differences between the two algorithms below (we refer to Algorithm 1 in **?** as DP^2). There are three key differences between DP^2 and DP-AdamBC:

1. Adam does not use pre-conditioned gradients in its update, since the moments are estimated from non-scaled gradients, whereas RMSprop (the base for DP^2 in **?**) uses scaled (thus pre-conditioned) gradients in the update. Figure 10 in **?** shows that the gains obeseved in DP^2 come in large part from reducing the amount of clipping and noising by using pre-conditioned gradients, but such advantage cannot directly transfer to Adam.

2. The pre-conditioning term (D_t) is computed on much larger batches of gradients, since it is accumulated over multiple iterations. This means that D_t is computed on a different distribution than the actual t-step gradients. For Adam, it would imply that the first moments (expectation of t-step gradient) and second moments (variance of t-step gradient) are estimated with different sampling distributions which could break Adam's sign descent behaviour.

3. DP^2 does not use momentum on the gradients (as the first moment in Adam) potentially because it alternates between two optimizers. Momentum is typically important on tasks for which Adam works well (**?**).

Empirically, **?** evaluates DP^2 on linear models and matrix factorization, and not on deep learning tasks. Figure 3 (Left) and 10 (Left) shows the comparison between DP2-RMSProp, DP-AdamBC and DP-SGD on SNLI with Bert-base and CIFAR10 with CNN respectively. DP2-RMSProp introduces several new hyperparameters (learning rates, clipping thresholds and delay parameters in both phases of SGD and RMSProp) whereas DP-AdamBC adds no additional hyperparameters compared to DP-Adam. We tuned DP^2 with grid search over: learning rate-{0.001, 0.01, 0.1, 1.0, 3.0, 5.0, 7.0, 10.0} for both datasets in both phases of DP^2, (CIFAR10) $s1 = s2 = s$ as suggested in **?**, s-{25, 65, 130, 250} (roughly 4, 10, 20, 40 out of 50 epochs), and C-{0.1, 1, 3} in both phases of DP^2; (SNLI) s-{1000, 2000, 4500} (roughly 0.5, 1, 2 out of 3 epochs) and C-{0.01, 0.1, 1} in both phases of DP^2, we fixed batch size B and noise multiplier σ to be the same as with the other three algorithms. Figure 3 (Left) and 10 (Left) shows the mean and standard deviation of the test accuracy over 5 runs with on the two datasets. We observe that DP^2 first follows DP-SGD (since the first steps use this optimizer), and then struggles to converge on deep learning tasks, leading to poor performance. Indeed switching between two optimizers seem to make DP^2 unstable: Figure 3 (Right) and 10 (Right) shows DP^2 on these two tasks with different s (switching frequency). We observe the performance either has large turbulence or drops significantly when switching optimizers during training.

E Privacy Analysis

Since DP-AdamBC uses the privitized gradient to update first and second moment estimates, and the DP bias Φ can be calculated from public hyperparameters B, σ, C. By the post-processing property of DP, for a given privacy ac-
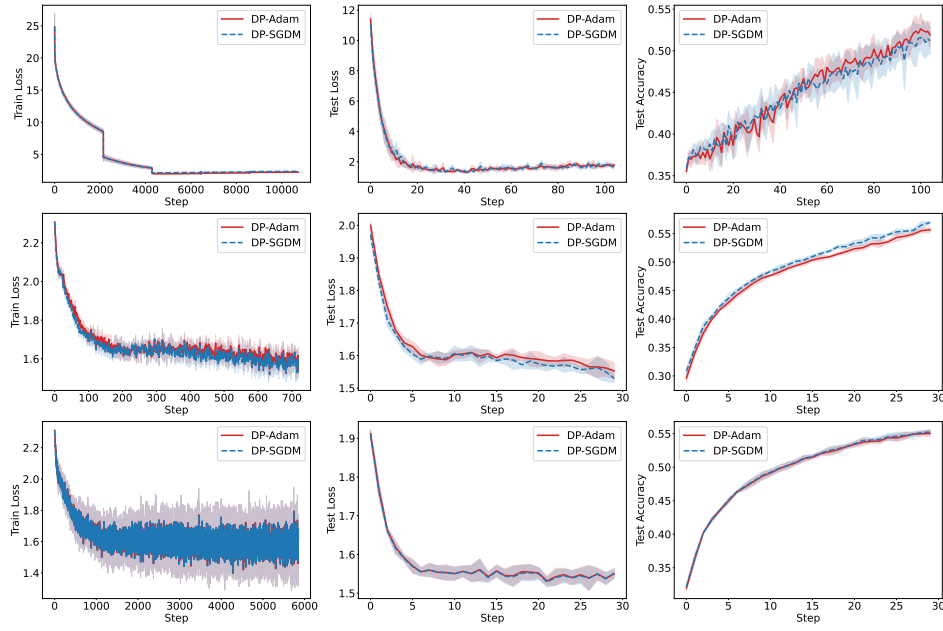
Figure 7: Compare train loss, test loss and test accuracy between DP-Adam and the DP-SGDM with the converted learning rate schedule as in Equation 3 on SNLI (top row) and CIFAR10 with relatively small Φ (middle row) and large Φ (bottom row) respectively.
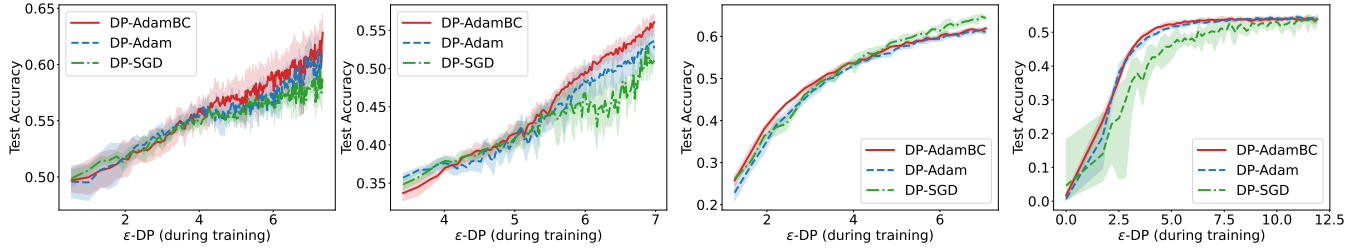

Figure 8: (From left to right) Comparing the performance of DP-Adam, DP-AdamBC and DP-SGD on QNLI and SNLI dataset (nlp), CIFAR10 (images) and obgn-arxiv (node classification). At the end of training ϵ-DP ≈ 7 for CIFAR10, QNLI and SNLI and ϵ-DP ≈ 12 for obgn-arxiv. Each optimizer is tuned separately. The x-axis is the step over a single training trajectory converted to privacy budget ϵ to make results comparable for different optimizers.

counting method, the same privacy guarantee holds for DP-AdamBC as with DP-SGD or DP-Adam. We formalize the proof as follows.

Theorem 2 (Privacy guarantee of DP-SGD). *There exist constants c_1 and c_2 so that given the sampling probability $q = L/N$ and the number of steps T, for any $\epsilon < c_1 q^2 T$, Algorithm 1 in* **?***) is (ϵ, δ)-differentially private for any $\delta > 0$ if we choose $\sigma \geq (c_2 q \sqrt{T \log{(1/\delta)}})/\epsilon$.*

Proposition 5 (Privacy guarantee of DP-AdamBC). *Let the optimization algorithm DP-SGD$(\theta, X, y, C, \sigma, B)$ (Algorithm 1 in* **?***), with privacy analysis Compose$(T, \theta_{1,...,T})$, be (ϵ, δ)-DP, then DP-AdamBC$(\theta, X, y, C, \sigma, B)$ with the same privacy analysis Compose$(T, \theta_{1,...,T})$ is also (ϵ, δ)-DP.*

Proof. Let *PrivitizeGradient*$(\theta, X, y, C, \sigma, B)$ be the key

step providing DP guarantee in DP-SGD, DP-Adam and DP-AdamBC:

$$(\text{Compute gradient}) \; g_t(x_i) \leftarrow \nabla_\theta \mathcal{L}(\theta, x_i), \forall i \in B,$$
$$(\text{Clip gradient}) \; \bar{g}_t(x_i) \leftarrow g_t(x_i) / \max{(1, \|g_t(x_i)\|/C)},$$
$$(\text{Noise gradient}) \; \tilde{g}_t \leftarrow (1/B)(\textstyle\sum_i \bar{g}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2)).$$

When the DP noise is sampled from a Gaussian distribution, by standard arguments of the Gaussian mechanism in **?** (Appendix A), the procedure is (ϵ', δ')-DP with $\sigma \geq (2\ln{(1.25/\delta')}C)/\epsilon'$. By the privacy amplification theorem, *PrivitizeGradient*$(\theta, X, y, C, \sigma, B)$ is $(O(q\epsilon'), q\delta')$-DP with sampling probability $q = B/N$ for batch size B and total sample size N. Let *Compose*$(T, \theta_{1,...,T})$ computes the overall privacy cost over T training iterations with a privacy accountant (e.g. strong composition in **?**, moment accountant in **?**, RDP accountant in **?**), *PrivitizeGradient*$(\theta, X, y, C, \sigma)$ over T iterations with *DP-SGD* update $\theta_{t+1} \leftarrow \theta_t - \eta \nabla_{\theta_t} f$,
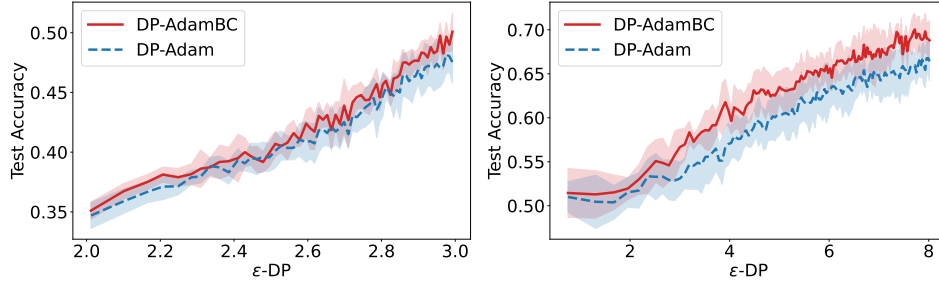
Figure 9: (**Left:**) Comparison between DP-AdamBC and DP-Adam when target $\epsilon = 3$ on the SNLI dataset. DP-AdamBC shows 2.6% advantage in final mean test accuracy. The final mean(standard deviation) test accurary for DP-AdamBC and DP-Adam are $50.1\%(1.6\%)$, $47.5\%(1.8\%)$ respectively. (**Right:**) Comparison between DP-AdamBC and DP-Adam on the SST2 dataset with Bert-Large. DP-AdamBC shows 2.4% advantage in final test accuracy compared to DP-Adam (68.80% vs 66.4%).
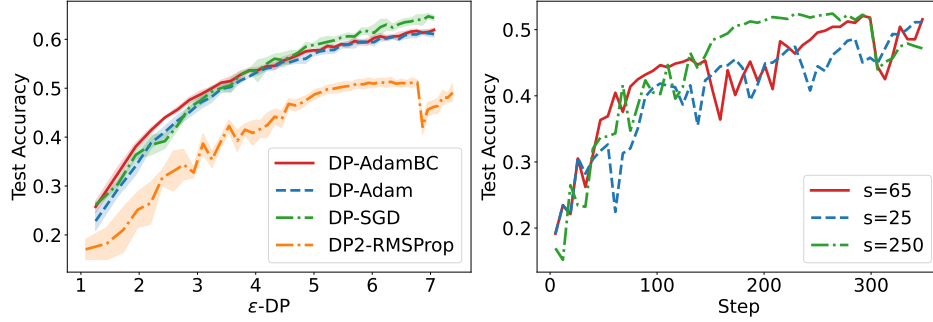

Figure 10: **Left:** Comparison between DP2RMSProp, DP-AdamBC, DP-Adam and DP-SGD on CIFAR10 with CNN, **Right:** The performance of DP2RMSProp with different phase switching frequency s on CIFAR10 with CNN.

where η is a hyperparameter representing learning rate, $\nabla_{\theta_t} f$ is the output privitized gradient from *PrivitizeGradient*, is (ϵ, δ)-DP. Since DP-AdamBC's update (Equation (4)) does not inquire additional private information as (1) Φ and γ' are determined from user-defined hyperparameters and (2) moment estimates are calculated from privitized gradients. By the post-processing property of DP (**?**), *PrivitizeGradient*$(\theta, X, y, C, \sigma)$ over T iterations with *DP-AdamBC* update is (ϵ, δ)-DP. If *Compose*$(T, \theta_{1,...,T})$ is the moment accountant in **?**, DP-AdamBC has the same privacy guarantees as in Theorem 1. □

F   Convergence analysis

In a recent work, **?** shows dropping the correction term in the first moment estimation (Section 2.2) has no observable effect and proves the convergence of Adam with this mild modification. We use the same setup and extend the analysis from **?**.

Let $F : \mathbb{R}^d \to \mathbb{R}$ be the objective function which $d \in \mathbb{N}$ is the problem dimension (number of parameters of the function to optimize), θ be the model parameters, $f : \mathbb{R}^d \to \mathbb{R}$ be the stochastic function such that $\forall \theta \in \mathbb{R}^d, \mathbb{E}[\nabla f(\theta)] = \nabla F(\theta)$, η be the learning rate, the subscript t be the step index, the subscript i be the dimension index. We make the following assumptions.

Assumption 2 (F is bounded below). *F is bounded below by F_*, i.e. $\forall \theta \in \mathbb{R}^d, F(\theta) \geq F_*$.*

Assumption 3 (Bounded stochastic gradient). *The $L2$-norm of the stochastic gradient is uniformly almost surely bounded, i.e. $\exists C \geq 0$ such that $\forall t, \forall \theta \in \mathbb{R}^d, \|\nabla f_t(\theta)\| \leq C$ a.s..*

Note that Assumption 3 implies that for every update step the gradient clipping operation in DP-SGD, DP-Adam or DP-AdamBC has no effect. It is usually assumed for the ease of theoretical analysis such as in **?**, but is often violated empirically.

At step t, we call the noise sample $z_t \sim Z$. The privatized gradient is $\tilde{g}_t = \nabla f_t(\theta) + z_t$, and the private second moment estimate is $v_t^p = \sum_{j=1}^{t} \beta_2^j \tilde{g}_t^2$. We introduce the following Lemma, which is the main technical change we need to adapt the proof of **?** to our setting with DP noise.

Lemma 3. *Let $\mu^* = (\beta_2(\beta_2^t-1)/(\beta_2-1))((\Phi-\frac{2\Phi}{\pi})+(C+\sqrt{\frac{2\Phi}{\pi}})^2)$, $\nu^* = (2\beta_2^2\Phi\sqrt{(\beta_2^t-1)/(\beta_2^2-1)})$, $b^* = 4\beta_2\Phi$ we have, $\Pr[v_t^p \geq \delta] \leq \alpha$ with,*

$$\delta \geq \begin{cases} \mu^* + \sqrt{2\nu^{*2}\ln\left(\frac{2}{\alpha}\right)} & 0 \leq \delta \leq \frac{\nu^{*2}}{b^*} \\ \mu^* + 2b^*\ln\left(\frac{2}{\alpha}\right) & \delta \geq \frac{\nu^{*2}}{b^*}. \end{cases}$$

Proof. By Assumption 3 we have $|\nabla f_t(\theta)| \leq C, \forall t$, so that $\beta_2^t(\nabla f_t(\theta) + z_t)^2 \leq \beta_2^t(|\nabla f_t(\theta)| + |z_t|)^2 \leq \beta_2^t(C + |z_t|)^2$. Since z_t is sampled from $\mathcal{N}(0, \Phi = (\sigma C/B)^2)$, by the similar argument as in Proposition 3, $\beta_2^t(C + z_t)^2$

is sub-exponential with $\nu = 2\beta_2^t\Phi$, $b = 4\beta_2^t\Phi$. Since each z_t is drawn independently, by the additive property of sub-exponential, $\sum_{j=1}^t \beta_2^j(C + z_t)^2$ is sub-exponential with $\nu^* = (2\beta_2^2\Phi\sqrt{(\beta_2^t-1)/(\beta_2^2-1)})$, $b^* = 4\beta_2\Phi$. In addition we have $\mathbb{E}[v_t^p] \leq \mathbb{E}[\sum_{j=1}^t \beta_2^j(C + |Z_t|)^2] = \sum_{j=1}^t \beta_2^j\mathbb{E}[(C + |Z_t|)^2] = \sum_{j=1}^t \beta_2^j(\text{Var}[|Z_j|] + (C + \mathbb{E}[|Z_j|])^2)$. Given $Z_j \sim N(0,\Phi)$, $|Z_j|$ has a truncated Normal distribution where $\mathbb{E}[|Z_j|] = \sqrt{\frac{2\Phi}{\pi}}$ and $\text{Var}[|Z_j|] = \Phi - \frac{2\Phi}{\pi}$. Let $\mu^* = (\beta_2(\beta_2^t-1)/(\beta_2-1))((\Phi - \frac{2\Phi}{\pi}) + (C + \sqrt{\frac{2\Phi}{\pi}})^2)$ we have $E[v_t^p] \leq \mu^*$. By Proposition 2.9 in **?** we have the result. $\square$

We can apply a union bound over Lemma 3 to directly obtain the following result:

Corollary 1. *Let μ^* be defined as in Lemma 3. We have that* $\Pr[\sup_t v_t^p \geq \delta] \leq \alpha$ *with,*

$$\delta \geq \begin{cases} \mu^* + \sqrt{2\nu^{*2}\ln\left(\frac{2T}{\alpha}\right)} & 0 \leq \delta \leq \frac{\nu^{*2}}{b^*} \\ \mu^* + 2b^*\ln\left(\frac{2T}{\alpha}\right) & \delta \geq \frac{\nu^{*2}}{b^*}. \end{cases}$$

Assumption 4 (Smoothness of F). *The gradient of F is L-Liptchitz-continuous such that $\forall \theta, \theta' \in \mathbb{R}^d, \|\nabla F(\theta) - \nabla F(\theta')\| \leq L\|\theta - \theta'\|$.*

Let θ_0 be the randomly initialized starting point, $\forall t \in [1,\ldots,T]$, the update rule of DP-AdamBC is $\theta_t \leftarrow \theta_{t-1} - \eta \cdot \hat{m}_t/\sqrt{\max\left(\hat{v}_t - \Phi, \gamma'\right)}$, and the update rule of DP-Adam is $\theta_t \leftarrow \theta_{t-1} - \eta \cdot \hat{m}_t/\sqrt{\hat{v}_t + \gamma}$, where $\hat{m}_t$, $\hat{v}_t$ and Φ are the private first, second moment estimates and the DP bias in private second moment estimates (Section 3). We prove the following propositions.

Proposition 6 (Convergence of DP-AdamBC and DP-Adam Without momentum). *Given the above assumptions and the defined update rule with $\beta_1 = 0, 0 < \beta_2 < 1$, $\eta_t = \eta\sqrt{\frac{1-\beta_2^t}{1-\beta_2}}, \eta > 0, \forall \alpha$ s.t. $0 < \alpha < 1$, let $\nu^* = (2\beta_2^2\Phi\sqrt{(\beta_2^t-1)/(\beta_2^2-1)})$, $b^* = 4\beta_2\Phi$, we have $\forall T \in \mathbb{N}^*$,*

(DP-Adam)

$$\mathbb{E}[\|\nabla F(\theta_t)\|^2] \leq \frac{2\delta(F(\theta_0) - F_*)}{\eta T}$$
$$+ \left(\frac{4d(C^2 + \Phi)}{\sqrt{1-\beta_2}} + \frac{\eta dL\sqrt{C^2+\Phi}}{1-\beta_2}\right)$$
$$\times \left(\frac{1}{T}\ln\left(1 + \frac{C^2+\Phi}{(1-\beta_2)\epsilon}\right) - \ln(\beta_2)\right),$$

(DP-AdamBC)

$$\mathbb{E}[\|\nabla F(\theta_t)\|^2] \leq \frac{2\sqrt{\delta^2 - \Phi}(F(\theta_0) - F_*)}{\eta T}$$
$$+ \left(\frac{4dC^2)}{\sqrt{1-\beta_2}} + \frac{\eta dLC}{1-\beta_2}\right)$$
$$\times \left(\frac{1}{T}\ln\left(1 - \frac{C^2+\Phi}{(1-\beta_2)\Phi}\right) - \ln(\beta_2)\right),$$

DP-Adam.

Proof. By dropping the correction term in m_t we have $\forall t, \eta_t = \eta\sqrt{\frac{1-\beta_2^t}{1-\beta_2}}$. Since Assumption 3 assumes away the effect of gradient clipping, and since the DP noises are sampled from a zero-mean Normal distribution, proving convergence for DP-Adam (without Momentum) is very similar to the original proof of Theorem 2 in (**?**). We show the key steps below. Using Assumption 4 we have,

$$F(\theta_t) \leq F(\theta_{t-1}) - \eta_t\nabla F(\theta_{t-1})^Tu_t + (\eta_t^2L/2)\|u_t\|^2,$$
$$u_t = \frac{\nabla_i f_t^p(\theta_{t-1})}{\sqrt{v_t^p + \epsilon}},$$

where u_t is the update of DP-Adam without momentum. Taking the complete expectation with respect to all past steps before t and t step noise distribution we get,

$$\mathbb{E}[F(\theta_t)] \leq F(\theta_{t-1}) - \eta_t\mathbb{E}\left[\frac{\nabla_i F(\theta_{t-1})\nabla_i f_t^p(\theta_{t-1})}{\sqrt{v_{t,i}^p + \epsilon}}\right]$$
$$+ \frac{\eta_t^2 L}{2}\mathbb{E}[\|u_t\|^2].$$

Given $\epsilon \ll v_t^p$, $\forall i \in [d]$ we can bound the first expectation term on the right side using Lemma 3 to get a high probability bound. By the similar steps as in Lemma 5.1 (**?**), except that we have $\mathbb{E}[(\nabla_i f_t^p(\theta_{t-1}))^2] = \mathbb{E}[(\nabla_i f_t^c(\theta_{t-1}) + z_t)^2] \leq C^2 + \Phi$, substituting the result of Lemma 3 gives the inequality on $\mathbb{E}\left[\frac{\nabla_i F(\theta_{t-1})\nabla_i f_t^p(\theta_{t-1})}{\sqrt{v_{t,i}^p + \epsilon}}\right]$. Substituting in the result and since $\eta \leq \eta_t$ we get,

$$\mathbb{E}[F(\theta_t)] \leq F(\theta_{t-1}) - \frac{\eta}{2\delta}\|\nabla F(\theta_{t-1})\|^2$$
$$+ \left(2\eta_t\sqrt{C^2 + \Phi} + \frac{\eta_t^2 L}{2}\right)\mathbb{E}[\|u_t\|^2],$$

where δ has the form as in Lemma 3. Summing over all steps $t \in [T]$ and taking the complete expectation, and since $\eta_t \leq \eta/\sqrt{1-\beta_2}$ we have,

$$\mathbb{E}[F(\theta_T)] \leq F(\theta_0) - \frac{\eta\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\theta_{t-1})\|^2]}{2\delta}$$
$$+ \left(\frac{2\eta\sqrt{C^2+\Phi}}{\sqrt{1-\beta_2}} + \frac{\eta^2 L}{2(1-\beta_2)}\right)\sum_{t=0}^{T-1}\mathbb{E}[\|u_t\|^2],$$

where δ has the form as in Corollary 1. Since $y = ln(x)$ is a concave function for $x \in \mathbb{R}^+$, by Jensen's inequality we have $\mathbb{E}(\ln(x)) \leq \ln(\mathbb{E}(x))$. By the similar steps as in

Lemma 5.2 (**?**) we have,

$$\sum_{t=0}^{T-1} \mathbb{E}[\|u_{t,i}\|^2] = \sum_{i=1}^{d} \mathbb{E}\left[\sum_{t=0}^{T-1} u_{t,i}^2\right]$$

$$\leq \sum_{i=1}^{d} \mathbb{E}[\ln\left(1 + v_T^p/\epsilon\right) - T\ln\left(\beta_2\right)]$$

$$\leq \sum_{i=1}^{d} \left(\ln\left(1 + \frac{(C^2 + \Phi)(1 - \beta_2^T)}{(1 - \beta_2)\epsilon}\right) - T\ln\left(\beta_2\right)\right)$$

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

Substituting in the result and rearrange the terms gives the final result. □

DP-AdamBC.

Proof. Similar to the proof steps above, by Assumption 4 we have,

$$F(\theta_t) \leq F(\theta_{t-1}) - \eta_t \nabla F(\theta_{t-1})^T u_t + (\eta_t^2 L/2)\|u_t\|^2,$$

$$u_t = \frac{\nabla_i f_t^p(\theta_{t-1})}{\sqrt{v_t^p - \Phi}},$$

where u_t is the update of DP-AdamBC without momentum. Taking the complete expectation we get,

$$\mathbb{E}[F(\theta_t)] \leq F(\theta_{t-1}) - \eta_t \mathbb{E}\left[\frac{\nabla_i F(\theta_{t-1})\nabla_i f_t^p(\theta_{t-1})}{\sqrt{v_{t,i}^p - \Phi}}\right]$$

$$+ \frac{\eta_t^2 L}{2}\mathbb{E}[\|u_t\|^2].$$

We derive a high probability bound for the first expectation term on the right side. Let $G = \nabla_i F(\theta_{t-1}), g = \nabla_i f_t^p(\theta_{t-1}), \bar{g} = \nabla_i f_t^c(\theta_{t-1})$, let $\tilde{v}_t = \beta_2 v_{t-1} + \mathbb{E}[g^2]$ denote v_t with the last gradient replaced by its conditional expectation with respect to all past steps, let $z, \tilde{v}, v$ abbreviate for $z_t, \tilde{v}_{t,i}^p, v_{t,i}^p$, we rewrite the first expectation term as follows,

$$\mathbb{E}\left[\frac{Gg}{\sqrt{v - \Phi}}\right] = \mathbb{E}\left[\frac{G(\bar{g} + z)}{\sqrt{v - \Phi}}\right]$$

$$= \mathbb{E}\left[\frac{G\bar{g}}{\sqrt{v - \Phi}}\right] + \mathbb{E}\left[\frac{Gz}{\sqrt{v - \Phi}}\right]$$

$$= \mathbb{E}\left[\frac{G\bar{g}}{\sqrt{v - \Phi}}\right],$$

since G and z are independent and z has expectation equals zero. In addition, we have that $E[g^2] = \mathbb{E}[(\bar{g} + z)^2] = \mathbb{E}[\bar{g}^2] + \Phi$, and by definition $\mathbb{E}[g^2] \leq \tilde{v}$, so that $\mathbb{E}[\bar{g}^2] \leq \tilde{v} - \Phi$ and $\mathbb{E}[\bar{g}^2] \leq C^2$. With these conditions, by the same steps as in Lemma 5.1 (**?**) we have,

$$\mathbb{E}\left[\frac{\nabla_i F(\theta_{t-1})\nabla_i f_t^p(\theta_{t-1})}{\sqrt{v_{t,i}^p - \Phi}}\right] \geq \frac{(\nabla_i F(\theta_{t-1}))^2}{2\sqrt{\tilde{v}_{t,i}^p - \Phi}}$$

$$- 2C\mathbb{E}\left[\frac{(\nabla_i f_t^p(\theta_{t-1}))^2}{v_{t,i}^p - \Phi}\right].$$

Substituting the result from Lemma 3 we get with probability of at least $(1 - \alpha), 0 < \alpha < 1$,

$$\mathbb{E}\left[\frac{\nabla_i F(\theta_{t-1})\nabla_i f_t^p(\theta_{t-1})}{\sqrt{v_{t,i}^p + \epsilon}}\right]$$

$$\geq \frac{(\nabla_i F(\theta_{t-1}))^2}{2\sqrt{\sum_{j=0}^{t-1}\beta_2^j \delta^2 - \Phi}} - 2C\mathbb{E}\left[\frac{(\nabla_i f_t^p(\theta_{t-1}))^2}{v_{t,i}^p + \epsilon}\right],$$

where δ is in the form as in Lemma 3. Summing over all steps $t \in [T]$ and taking the complete expectation we have,

$$\mathbb{E}[F(\theta_T)] \leq F(\theta_0) - \frac{\eta \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla F(\theta_{t-1})\|^2]}{2\sqrt{\delta^2 - \Phi}}$$

$$+ \left(\frac{2\eta C}{\sqrt{1 - \beta_2}} + \frac{\eta^2 L}{2(1 - \beta_2)}\right)\sum_{t=0}^{T-1} \mathbb{E}[\|u_t\|^2],$$

where δ is in the form as in Corollary 1. We bound $\sum_{t=0}^{T-1} \mathbb{E}[\|u_t\|^2]$ using the similar approach as in Lemma 5.2 (**?**). Let $a_t = (\nabla f_t^p)^2$ and $b_t = \sum_{j=1}^{t} \beta_2^{t-j} a_j$, then $\sum_{t=1}^{T} \mathbb{E}[\|u_t\|^2] = \sum_{i=1}^{d} \mathbb{E}[\sum_{t=1}^{T} \frac{a_{t,i}}{b_{t,i} - \Phi}]$. Given $\ln$ is increasing and $b_t > a_t > 0$,

$$\frac{a_{t,i}}{b_{t,i} - \Phi} \leq \ln\left(1/(1 - \frac{a_{t,i}}{b_{t,i} - \Phi})\right)$$

$$= \ln\left(\frac{b_{t,i} - \Phi}{b_{t-1,i} - \Phi}\right) + \ln\left(\frac{b_{t-1,i} - \Phi}{\beta_2 b_{t-1,i} - \Phi}\right)$$

$$\sum_{t=1}^{T} \frac{a_{t,i}}{b_{t,i} - \Phi} = \sum_{t=1}^{T} \ln\left(\frac{b_{t,i} - \Phi}{b_{t-1,i} - \Phi}\right) + \sum_{t=1}^{T} \ln\left(\frac{b_{t-1,i} - \Phi}{\beta_2 b_{t-1,i} - \Phi}\right)$$

$$\leq \ln\left(1 - \frac{b_{T,i}}{\Phi}\right) - T\ln\left(\beta_2\right),$$

where the last inequality is because $b_0 = 0$ and $b_t > \Phi \; \forall t$. Substituting the result in we get,

$$\sum_{t=0}^{T-1} \mathbb{E}[\|u_{t,i}\|^2] = \sum_{i=1}^{d} \mathbb{E}\left[\sum_{t=0}^{T-1} u_{t,i}^2\right]$$

$$\leq \sum_{i=1}^{d} \mathbb{E}[\ln\left(1 - v_T^p/\Phi\right) - T\ln\left(\beta_2\right)]$$

$$\leq \sum_{i=1}^{d} \left(\ln\left(1 - \frac{(C^2 + \Phi)(1 - \beta_2^T)}{(1 - \beta_2)\Phi}\right) - T\ln\left(\beta_2\right)\right)$$

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

Substituting in the result and rearrange the terms gives the final result. □

Discussion on the convergence bound. Comparing the convergence rate between the two algorithm is not straightforward giving that one has a slight advantage in $\mathbb{E}[(\nabla F)u_t]$, the expectation of the update direction deviating from the true descent direction, and a disadvantage in $\mathbb{E}\|u_t\|^2$, the

expectation of the update size. We discuss an approximate comparison between $\mathbb{E}\|u_t\|^2$. Let $u_t^{\text{DP-Adam}} = \frac{g}{\sqrt{v+\epsilon}}$, $u_t^{\text{DP-AdamBC}} = \frac{g}{\sqrt{v-\Phi}}$, with the data and noise distribution we can consider the numerator and denominator as two random variables such that for each dimension i, for DP-Adam we have $\mathbb{E}[u_{t,i}^2] = \mathbb{E}[A/B]$, for DP-AdamBC we have $\mathbb{E}[u_{t,i}^2] = \mathbb{E}[A/B']$, where A is the random variable for t-step gradient g^2, B is the random variable for $v + \epsilon$, B' is the random variable for $v - \Phi$. It would be difficult to derive closed form results for expectation on ratios of random variables without specific assumption, but taking a second-order Taylor expansion approximately gives $\mathbb{E}[\frac{X}{Y}] \approx \frac{\mathbb{E}[X]}{\mathbb{E}[Y]} - \frac{Cov(X,Y)}{\mathbb{E}[Y]^2} + \frac{\text{Var}(Y)}{\mathbb{E}[Y]^3}$. Since $\text{Var}(B) = \text{Var}(B')$, the differences between the approximated $\mathbb{E}[A/B]$ and $\mathbb{E}[A/B']$ is only in the denominators which only differs by constant Φ. Therefore, giving that $\mathbb{E}[(\nabla F)u_t]$ only differs by constant terms, and $\mathbb{E}\|u_t\|^2$ are approximately similar, we believe qualitatively there is no large difference in the convergence rate between the two algorithms under such analysis settings.

Proposition 7 (Convergence of DP-AdamBC and DP-Adam With momentum). *Given the above assumptions and the defined update rule with $0 < \beta_2 < 1$, $0 \leq \beta_1 < \beta_2$, $\eta_t = \eta(1-\beta_1)\sqrt{\frac{1-\beta_2^t}{1-\beta_2}}, \eta > 0, \forall \alpha \text{ s.t. } 0 < \alpha < 1$, we have $\forall T \in \mathbb{N}^*$ such that $T > \frac{\beta_1}{1-\beta_1}$ and with $\tilde{T} = T - \frac{\beta_1}{1-\beta_1}$, let $\nu^* = (2\beta_2^2\Phi\sqrt{(\beta_2^t - 1)/(\beta_2^2 - 1)})$, $b^* = 4\beta_2\Phi$,*

(DP-Adam)

$$\mathbb{E}[\|\nabla F(\theta_t)\|^2] \leq \frac{2\delta(F_0 - F_*)}{\eta\tilde{T}}$$
$$+ E\left(\ln\left(1 + \frac{\delta^2}{\epsilon(1-\beta_2)}\right) - T\log(\beta_2)\right),$$

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

(DP-AdamBC)

$$\mathbb{E}[\|\nabla F(\theta_t)\|^2] \leq \frac{2\sqrt{\delta^2 - \Phi}(F_0 - F_*)}{\eta\tilde{T}}$$
$$+ E\left(\ln\left(1 - \frac{\delta^2}{\Phi(1-\beta_2)}\right) - T\log(\beta_2)\right),$$

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

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

DP-Adam.

Proof. Let $G_t = \nabla F(\theta_{t-1}), g_t = \nabla f_t^p(\theta_t - 1)$, by Assumption 4 we have,

$$F(\theta_t) \leq F(\theta_{t-1}) - \eta_t G_t^T u_t + \frac{\eta_t^2 L}{2}\|u_t\|^2, \ u_t = \frac{m_t^p}{\sqrt{v_t^p + \epsilon}},$$

where m_t^p and v_t^p are the first and second moment estimated from privatized gradient which makes u_t the update of DP-Adam with momentum. Taking the expectation over past steps we get,

$$\mathbb{E}[F(\theta_t)] \leq \mathbb{E}[F(\theta_{t-1})] - \eta_t\mathbb{E}[G_t^T u_t] + \frac{\eta_t^2 L}{2}\mathbb{E}[\|u_t\|^2].$$

We bound $\mathbb{E}[G_t^T u_t]$ using a similar approach as in Lemma A.1 (?) with the following key steps. For index $0 \leq k \leq t - 1$, we first decompose $G_t^T u_t$ as,

$$\sum_{i\in[d]} G_{t,i}\frac{m_{t,i}^p}{\sqrt{v_{t,i}^p + \epsilon}} = \underbrace{\sum_{i\in[d]}\sum_{k=0}^{t-1}\beta_1^k G_{t-k,i}\frac{g_{t-k,i}}{\sqrt{v_{t,i}^p + \epsilon}}}_{A}$$
$$+ \underbrace{\sum_{i\in[d]}\sum_{k=0}^{t-1}\beta_1^k(G_{t,i} - G_{t-k,i})\frac{g_{t-k,i}}{\sqrt{v_{t,i}^p + \epsilon}}}_{B}.$$

We first bound B with the Gaussian concentration bound on $v_{t,i}^p$ using the similar approach as in Equation (A.13) in ?. Let $\lambda = \frac{\sqrt{1-\beta_1}}{2\delta}$ where δ is in the form as in Lemma 3, $x = |G_{t,i} - G_{t-k,i}|, y = \frac{g_{t-k,i}}{\sqrt{v_{t,i}^p+\epsilon}}$, following the same steps we get,

$$|B| \leq \frac{\eta_t^2 L^2\sqrt{1-\beta_1}}{4\delta}\left(\sum_{l=1}^{t-1}\|u_{t-l}\|^2\sum_{k=l}^{t-1}\beta_1^k\sqrt{k}\right)$$
$$+ \frac{\delta}{\sqrt{1-\beta_1}}\left(\sum_{k=0}^{t-1}(\frac{\beta_1}{\beta_2})^k\|U_{t-k}\|^2\right).$$

Let $\tilde{v}_{t,k}^p = \beta_2^k v_{t-k} + \mathbb{E}[\sum_{j=t-k+1}^t \beta_2^{t-j}g_j^2]$ be the second moment estimate with last k gradients replaced by their expected value, since by definition $\tilde{v}_{t,k+1}^p + \epsilon \geq \mathbb{E}[\sum_{j=t-k}^t \beta_2^{t-j}g_j^2]$ and $v_t^p + \epsilon \geq \sum_{j=t-k}^t \beta_2^{t-j}g_j^2$ and with the result of Lemma 3, following the same steps as in (?),

$$\mathbb{E}[A] \geq \frac{1}{2}\left(\sum_{i\in[d]}\sum_{k=0}^{t-1}\beta_1^k\mathbb{E}\left[\frac{G_{t-k,i}^2}{\sqrt{\tilde{v}_{t,k+1,i} + \epsilon}}\right]\right)$$
$$- 2\sqrt{\frac{\delta}{1-\beta_1}}\left(\sum_{i\in[d]}\sum_{k=0}^{t-1}(\frac{\beta_1}{\beta_2})^k\mathbb{E}[\|U_{t-k}\|^2]\right).$$

Then combing the results for A and B gives the bound on $\mathbb{E}[G_t^T u_t]$. Let $\Omega_t = \sqrt{\sum_{j=0}^{t-1}\beta_2^j}$, by Assumption 2, sum-

ming over all steps t and reorganizing the terms we get,

$$\frac{\sum_{t=1}^{T}\frac{\eta_t}{\Omega_t}\sum_{k=0}^{t-1}\beta_1^k\mathbb{E}[\|G_{t-k}\|^2]}{2\delta} \leq F(\theta_0) - F_*$$
$$+ \frac{\eta_N^2 L}{2}\sum_{t=1}^{T}\mathbb{E}[\|u_t\|^2] + \frac{\eta_T^3 L^2\sqrt{1-\beta_1}}{4\delta}\sum_{t=1}^{T}\sum_{l=1}^{t-1}\beta_1^k\sqrt{k}$$
$$+ \frac{3\eta_T\delta}{\sqrt{1-\beta_1}}\sum_{t=1}^{T}\sum_{k=0}^{t-1}(\frac{\beta_1}{\beta_2})^k\sqrt{k+1}\mathbb{E}[\|U_{t-k}\|^2],$$

where δ is in the form as in Corollary 1. Bounding $\mathbb{E}[\|u_t\|^2]$ is similar to the steps in Lemma A.2 (?) which we have,

$$\sum_{t=1}^{T}\mathbb{E}[\|u_t\|^2] \leq \frac{\sum_{i\in[d]}\ln\left(1+\frac{v_{T,i}}{\epsilon}\right) - T\log(\beta_2)}{(1-\beta_1)(1-\beta_1/\beta_2)},$$
$$v_{T,i} \leq \frac{(C+\sqrt{-\ln\frac{\alpha}{2T}(2\Phi)})^2}{1-\beta_2}.$$

The rest of the proof rearranges the other terms with techniques including changing index and order of summation and is exactly the same as in (?) which leads to the final result. $\square$

DP-AdamBC.

Proof. We start with,

$$\mathbb{E}[F(\theta_t)] \leq \mathbb{E}[F(\theta_{t-1})] - \eta_t\mathbb{E}[G_t^T u_t] + \frac{\eta_t^2 L}{2}\mathbb{E}[\|u_t\|^2],$$
$$u_t = \frac{m_t^p}{\sqrt{v_t^p - \Phi}},$$

where u_t is the update of DP-AdamBC with momentum. To bound $\mathbb{E}[G_t^T u_t]$ we first decompose the quantity as follows,

$$\mathbb{E}[G_t^T u_t] = \sum_{i\in[d]}\mathbb{E}\left[\frac{G_{t,i}m_{t,i}^p}{\sqrt{v_{t,i}^p - \Phi}}\right]$$
$$= \sum_{i\in[d]}\mathbb{E}\left[\frac{G_{t,i}m_{t,i}^c}{\sqrt{v_{t,i}^p - \Phi}}\right] + \underbrace{\sum_{i\in[d]}\mathbb{E}\left[\frac{G_{t,i}\sum_{j=1}^{t}\beta_1^{t-i}z_{t,i}}{\sqrt{v_{t,i}^p - \Phi}}\right]}_{0}$$
$$= \underbrace{\sum_{i\in[d]}\sum_{k=0}^{t-1}\beta_1^k G_{t-k,i}\mathbb{E}\left[\frac{\bar{g}_{t-k,i}}{\sqrt{v_{t,i}^p - \Phi}}\right]}_{A}$$
$$+ \underbrace{\sum_{i\in[d]}\sum_{k=0}^{t-1}\beta_1^k(G_{t,i} - G_{t-k,i})\mathbb{E}\left[\frac{\bar{g}_{t-k,i}}{\sqrt{v_{t,i}^p - \Phi}}\right]}_{B}.$$

Let $\tilde{v}_{t,k}^p = \beta_2^k v_{t-k} + \mathbb{E}[\sum_{j=t-k+1}^{t}\beta_2^{t-j}g_j^2]$ be the second moment estimate with last k gradients replaced by their ex-

pected value, and substituting result from Lemma 3,

$$\mathbb{E}[G_t^T u_t] \geq \frac{1}{2}\left(\sum_{i\in[d]}\sum_{k=0}^{t-1}\beta_1^k\mathbb{E}\left[\frac{G_{t-k,i}}{\sqrt{\tilde{v}_{t,k+1,i} - \Phi}}\right]\right)$$
$$- \frac{\sqrt{1-\beta_1}\eta_t^2 L^2}{4\sqrt{\delta^2 - \Phi}}\left(\sum_{l=1}^{l-1}\|u_{t-l}\|^2\sum_{k=1}^{l-1}\beta_1^k\sqrt{k}\right)$$
$$- \frac{3\sqrt{\delta^2 - \Phi}}{\sqrt{1-\beta_1}}\left(\sum_{k=0}^{t-1}(\frac{\beta_1}{\beta_2})^2\|U_{t-k}\|^2\right),$$

where δ is in the form as in Lemma 3. Let $\Omega_t = \sqrt{\sum_{j=0}^{t-1}\beta_2^j}$, by Assumption 2, summing over all steps t and reorganizing the terms we get,

$$\frac{\sum_{t=1}^{T}\frac{\eta_t}{\Omega_t}\sum_{k=0}^{t-1}\beta_1^k\mathbb{E}[\|G_{t-k}\|^2]}{2\sqrt{\delta^2 - \Phi}} \leq F(\theta_0) - F_*$$
$$+ \frac{\eta_N^2 L}{2}\sum_{t=1}^{T}\mathbb{E}[\|u_t\|^2] + \frac{\eta_T^3 L^2\sqrt{1-\beta_1}}{4\sqrt{\delta^2 - \Phi}}\sum_{t=1}^{T}\sum_{l=1}^{t-1}\beta_1^k\sqrt{k}$$
$$+ \frac{3\eta_T\sqrt{\delta^2 - \Phi}}{\sqrt{1-\beta_1}}\sum_{t=1}^{T}\sum_{k=0}^{t-1}(\frac{\beta_1}{\beta_2})^k\mathbb{E}[\|U_{t-k}\|^2],$$

where δ is in the form as in Corollary 1. Bounding $\mathbb{E}[\|u_t\|^2]$ is similar to the steps in Lemma A.2 (?) which we have,

$$\sum_{t=1}^{T}\mathbb{E}[\|u_t\|^2] \leq \frac{\sum_{i\in[d]}\ln\left(1-\frac{v_{T,i}}{\Phi}\right) - T\log(\beta_2)}{(1-\beta_1)(1-\beta_1/\beta_2)},$$
$$v_{T,i} \leq \frac{(C+\sqrt{-\ln\frac{\alpha}{2T}(2\Phi)})^2}{1-\beta_2}.$$

The rest of the proof rearranges the other terms with techniques including changing index and order of summation and is exactly the same as in (?) which leads to the final result.

$\square$

G Limitations

We observe that DP-AdamBC improves performance of DP-Adam in the cases where both algorithms outperform DP-SGD, such as in text classification tasks with SNLI and QNLI. In cases where DP-SGD outperforms DP-Adam, such as in image classification with CIFAR10 (Figure 8) and in node classification with obgn-arxiv (as reported in ?), DP-AdamBC tends to perform similarly to DP-Adam, with minor advantages. Although the observed DP bias is quite concentrated around its mean Φ, we note that γ' in DP-AdamBC is an important hyperparameter that affects the choice of learning rate η and affects the final performance. As such, our results are dependent on our efforts tuning parameters for each algorithm. However, this also opens avenues for improvement. Since γ' concentrates over steps t, we could apply a decreasing schedule for γ' (and η, since a smaller learning rate is typically needed for smaller γ') following the bound of Propositions 3 and 4 (and confirmed in the numerical analysis in §C).