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 percentages in final accuracy in image, text, and graph node classification tasks.

1 Introduction

The Adam optimization algorithm (Kingma and Ba 2014) 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 (Carlini et al. 2019, 2021, 2022; Balle, Cherubin, and Hayes 2022). We consider settings in which the deep learning model's training data is privacy sensitive, and models are trained with Differential Privacy (Dwork et al. 2006; Abadi et al. 2016) to provably prevent training example information leakage (Wasserman and Zhou 2010). 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 (Li et al. 2021) and GNN (Daigavane et al. 2021) models. However we observe that when combined with DP, Adam does not perform as well as without privacy constraints.

To understand this effect, we go back to the original intuition behind Adam (Kingma and Ba 2014) that relies on exponential moving averages estimating the first and second moments of mini-batch 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 (Kunstner et al. 2023), 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 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 (Tang, Shpilevskiy, and Lécuyer 2023).

2 Adam and the Sign-Descent Hypothesis

The Adam update (Kingma and Ba 2014) 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 $g_t = (1/B)\nabla f(\theta_{t-1})$, and 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 (Kunstner et al. 2023) 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 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 (Abadi et al. 2016): 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:

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

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

where $\overline{g}_t$ 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., (Li et al. 2021)), 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:

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

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 (Tang and Lécuyer 2023). By independence of the DP noise z_t and $\bar{g}_t$, we have that:

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

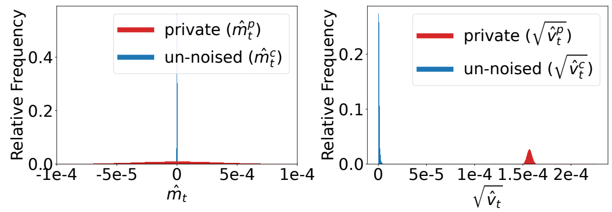

Figure 1: Histogram of (Left:) un-noised ($\hat{m}_t^c$) and private ($\hat{m}_t^p$) first moment estimates, (Right:) un-noised ($\hat{v}_t^c$) and private ($\hat{v}_t^p$) second moment estimates near end of training at $t = 10000$, using the SNLI dataset with $B = 256, C = 0.1, \sigma = 0.4, \beta_2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

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).

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

Figure 1 (Left) shows the histogram 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 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 (Kingma and Ba 2014) 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 [(Kingma and Ba 2014), §3].*

Under Assumption 1, with $\beta_1 \to 1$, 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 (Kunstner et al. 2023) or inverse variance conditioning (Balles and Hennig 2020). 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 $b_t^p = \frac{1-\beta}{1-\beta^t} \hat{m}_t^p$. 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$2.4e-8, ≈ 0.026 for $\Phi \approx$1.5e-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.

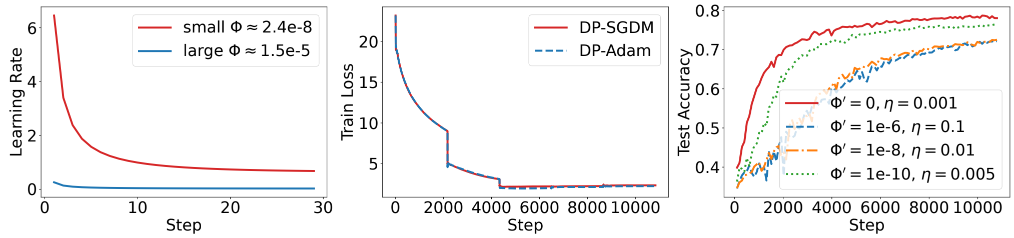

Figure 2: DP-Adam behaves similarly to DP-SGDM with a specific learning rate (lr) schedule. (Left:) The implied lr schedule of DP-SGDM. (Middle:) DP-Adam and DP-SGDM with the specific lr schedule has similar training performance with a mean squared difference of 0.015 in training loss. (Right:) Performance 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 (Mohapatra et al. 2021). 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:

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

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

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
  $g_i = \nabla\mathcal{L}(\theta_{t-1}, x_i)$, $z_t \sim \mathcal{N}\left(0, \sigma^2 C^2 \mathbb{I}^d\right)$
  $\tilde{g}_t = \frac{1}{B}\left(\sum_i g_i / \max\left(1, \frac{\|g_i\|_2}{C}\right) + z_t\right)$
  $m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1)\,\tilde{g}_t$, $\widehat{m}_t \leftarrow m_t / \left(1 - \beta_1^t\right)$
  $v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2)\,\tilde{g}_t^2$, $\widehat{v}_t \leftarrow v_t / \left(1 - \beta_2^t\right)$
  $\theta_t \leftarrow \theta_{t-1} - \eta \cdot \hat{m}_t / \sqrt{\max\left(\hat{v}_t - (\sigma C/B)^2, \gamma'\right)}$
end

Proposition 1 (Privacy guarantee of DP-AdamBC). *Let the optimization algorithm DP-SGD($\theta, X, y, C, \sigma, B$) (Algorithm 1 in Abadi et al. (2016)), 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 $\hat{v}_t^{\text{corr}} = \frac{(1-\beta_2)\sum_{\tau=1}^{t} \beta_2^{t-\tau} \tilde{g}_t^2}{1-\beta_2^t} - \left(\frac{\sigma C}{B}\right)^2$, 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 re-enables 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 $\hat{v}_t^{\text{corr}}$ 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 $\hat{v}_t^{\text{corr}}$ 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 $\mathbb{P}[|\hat{v}_t^{\text{corr}} - \hat{v}_t^c|_i \geq \xi] \leq \alpha$ 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 $\nu = (\frac{4\sigma^2 C^2}{B^2})\sqrt{\frac{1-\beta_2^{2t}}{1-\beta_2^2}}, b = \frac{4\sigma^2 C^2}{B^2}$.

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 $\hat{v}_t^{\text{corr}}$ error are concentrated around their mean $\hat{v}_t^c$, with $\hat{v}_t^{\text{corr}} - \hat{v}_t^c$ 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 $\mathbb{E}[\hat{v}_t^{\text{corr}}]$ 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 Défossez et al. (2022), adding the bounded gradient assumption from Li et al. (2023a) 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 technically 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 (Krizhevsky 2009), SNLI (Bowman et al. 2015), QNLI (Wang et al. 2019) and ogbn-arxiv (Hu et al. 2021) 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 (Devlin et al. 2018); for node classification, we train a DP-GCN model (Daigavane et al. 2021) 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 δ = 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.

		$\epsilon \approx 1$	$\epsilon \approx 3$	$\epsilon \approx 7$
SNLI	DP-SGD	**48.0 (1.3)**	45.1 (1.8)	51.0 (0.5)
	DP-Adam	44.7 (1.3)	47.5 (1.8)	52.6 (1.9)
	DP-AdamBC	45.2 (1.0)	**50.1 (1.6)**	**56.1 (1.0)**
QNLI	DP-SGD	57.1 (1.6)	58.9 (1.2)	58.3 (0.9)
	DP-Adam	58.0 (2.1)	60.7 (1.1)	61.2 (1.3)
	DP-AdamBC	**58.3 (1.9)**	**61.4 (1.0)**	**62.8 (1.6)**
CIFAR10	DP-SGD	**52.4 (0.5)**	**57.3 (0.8)**	**65.3 (0.3)**
	DP-Adam	51.9 (0.7)	54.1 (0.4)	62.2 (0.1)
	DP-AdamBC	49.8 (0.6)	54.3 (0.2)	63.4 (0.4)
		$\epsilon \approx 3$	$\epsilon \approx 6$	$\epsilon \approx 12$
obgn-arxiv	DP-SGD	45.4 (1.4)	49.1 (1.9)	**54.2 (0.6)**
	DP-Adam	46.6 (0.5)	52.0 (0.5)	54.0 (0.2)
	DP-AdamBC	**50.5 (0.6)**	**53.4 (0.3)**	53.8 (0.3)

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.

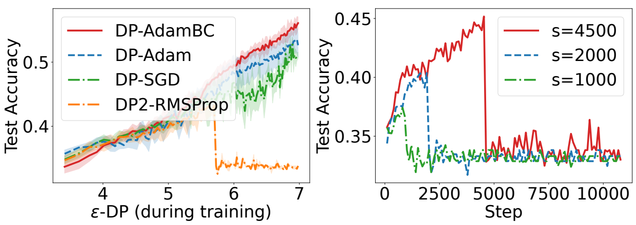

Figure 3: (Left:) Comparison between DP2RMSProp, DP-AdamBC, DP-Adam and DP-SGD and (Right:) the performance of DP2RMSProp with different phase switching frequency s on SNLI with Bert-base.

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 (Kunstner et al. 2023), 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 Kunstner et al. (2023)); 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 (Wilson et al. (2018), Table 1 in Daigavane et al. (2021), 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 developped for DP, named DP^2 (Li et al. 2023b). 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 $\Delta_t^c = \hat{m}_t^p/\sqrt{\hat{v}_t^c}, \Delta_t^p = \hat{m}_t^p/\sqrt{\hat{v}_t^p}$ and $\Delta_t^{\text{corr}} = \hat{m}_t^p/\sqrt{\hat{v}_t^{\text{corr}}}$. Table 2 shows the summary statistics of these variables near end of training, computed with the SNLI dataset with $B = 256, C = 0.1, \sigma = 0.4, \Phi \approx 2.441$e-8 in the limit of t. We observe 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, suggesting that the DP bias Φ dominates over the un-noised estimates of second moment $\hat{v}_t^c$. We also observe the scale of $\hat{v}_t^p$ is generally close to Φ, suggesting the private estimate of the second moments are largely affected by the DP noise. The scale of the corrected second moment estimates, $\hat{v}_t^{\text{corr}} = \max(\hat{v}_t - \Phi, \gamma')$ is closer to the scale of $\hat{v}_t^c$, with the numerical stability constant ($\gamma' = 3$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).

	Min	Q1	Median	Q3	Max	Mean
m_t^c	-8e-05	-7.e-08	-2e-18	7e-08	8e-05	4e-10
$\hat{m}_t^p$	-2e-04	-2e-05	1e-08	2e-05	2e-04	6e-09
$\hat{v}_t^c$	4e-24	8e-14	4e-13	1e-12	3e-08	4e-12
$\hat{v}_t^p$	2e-08	2e-08	3e-08	3e-08	6e-08	3e-08
$\hat{v}_t^{\text{corr}}$	3e-10	3e-10	3e-10	7e-10	3e-08	6e-10
Δ_t^c	-3e+07	-4e+01	2e-02	4e+01	2e+07	-3e+01
Δ_t^p	-1.2	-2e-01	8e-05	2e-01	1.1	4e-05
Δ_t^{corr}	-1e+01	-1.1	6e-04	1.1	1e+01	4e-04

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

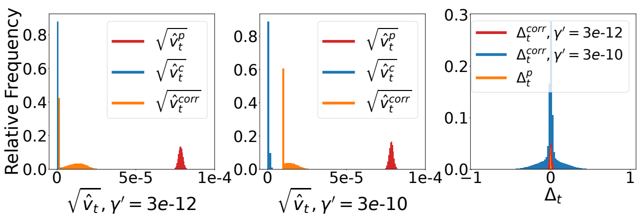

Figure 4: Histogram of private ($\hat{v}_t^p$), un-noised ($\hat{v}_t^c$) and corrected ($\hat{v}_t^{\text{corr}}$) second moment estimates with (Left:) $\gamma' = 3e$-12, (Middle:) $\gamma' = 3e$-10. (Right:) private (Δ_t^p) and corrected (Δ_t^{corr}) Adam updates with respect to m_t^p.

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 ($\sqrt{\hat{v}_t^{\text{corr}}}$) second moment estimates, when $\gamma' = 3e$-12 and $3e$-10 respectively. We see 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 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 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 $\mathbb{E}[\bar{g}_t^2]$.

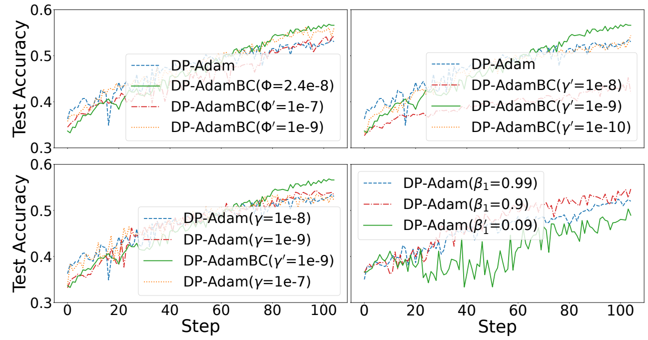

Figure 5: Performance when (Upper Left:) subtracting different (fake) values of Φ, (Upper Right:) tuning γ' in DP-AdamBC, (Lower Left:) tuning γ in DP-Adam, (Lower Right:) tuning βs in DP-Adam. Tuning hyperparameters in DP-Adam cannot replace DP-AdamBC's bias correction.

Effect of the Numerical Stability Constant. The numerical stability constant γ can affect the performance of Adam in the non-private setting, and γ is often tuned as a hyperparameter (Reddi, Kale, and Kumar 2019). 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 impact 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.

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 Kingma and Ba (2014), 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.

Acknowledgments

We are grateful for the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) [reference number RGPIN-2022-04469], as well as a Google Research Scholar award. This research was enabled by computational support provided by the Digital Research Alliance of Canada (alliancecan.ca), and by the University of British Columbia's Advanced Research Computing (UBC ARC).

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*.

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

Balles, L.; and Hennig, P. 2020. Dissecting Adam: The Sign, Magnitude and Variance of Stochastic Gradients. arXiv:1705.07774.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632–642. Lisbon, Portugal: Association for Computational Linguistics.

Carlini, N.; Chien, S.; Nasr, M.; Song, S.; Terzis, A.; and Tramer, F. 2022. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*.

Carlini, N.; Liu, C.; Erlingsson, Ú.; Kos, J.; and Song, D. 2019. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. In *USENIX Security Symposium*.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T. B.; Song, D.; Erlingsson, U.; et al. 2021. Extracting Training Data from Large Language Models. In *USENIX Security Symposium*.

Daigavane, A.; Madan, G.; Sinha, A.; Thakurta, A. G.; Aggarwal, G.; and Jain, P. 2021. Node-level differentially private graph neural networks. *arXiv preprint arXiv:2111.15521*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*.

Défossez, A.; Bottou, L.; Bach, F.; and Usunier, N. 2022. A Simple Convergence Proof of Adam and Adagrad. arXiv:2003.02395.

Hu, W.; Fey, M.; Zitnik, M.; Dong, Y.; Ren, H.; Liu, B.; Catasta, M.; and Leskovec, J. 2021. Open Graph Benchmark: Datasets for Machine Learning on Graphs. arXiv:2005.00687.

Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images. Technical report.

Kunstner, F.; Chen, J.; Lavington, J. W.; and Schmidt, M. 2023. Heavy-tailed Noise Does Not Explain the Gap Between SGD and Adam, but Sign Descent Might. In *International Conference on Learning Representations*.

Li, T.; Zaheer, M.; Liu, K. Z.; Reddi, S. J.; McMahan, H. B.; and Smith, V. 2023a. Differentially Private Adaptive Optimization with Delayed Preconditioners. arXiv:2212.00309.

Li, T.; Zaheer, M.; Liu, K. Z.; Reddi, S. J.; McMahan, H. B.; and Smith, V. 2023b. Differentially Private Adaptive Optimization with Delayed Preconditioners. arXiv:2212.00309.

Li, X.; Tramèr, F.; Liang, P.; and Hashimoto, T. 2021. Large Language Models Can Be Strong Differentially Private Learners.

Mohapatra, S.; Sasy, S.; He, X.; Kamath, G.; and Thakkar, O. 2021. The Role of Adaptive Optimizers for Honest Private Hyperparameter Selection. arXiv:2111.04906.

Reddi, S. J.; Kale, S.; and Kumar, S. 2019. On the Convergence of Adam and Beyond.

Tang, Q.; and Lécuyer, M. 2023. DP-Adam: Correcting DP Bias in Adam's Second Moment Estimation. arXiv:2304.11208.

Tang, Q.; Shpilevskiy, F.; and Lécuyer, M. 2023. DP-AdamBC: Your DP-Adam Is Actually DP-SGD (Unless You Apply Bias Correction).

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. arXiv:1804.07461.

Wasserman, L.; and Zhou, S. 2010. A statistical framework for differential privacy. *Journal of the American Statistical Association*.

Wilson, A. C.; Roelofs, R.; Stern, M.; Srebro, N.; and Recht, B. 2018. The Marginal Value of Adaptive Gradient Methods in Machine Learning. arXiv:1705.08292.