DP-ADAM: CORRECTING DP BIAS IN ADAM'S SECOND MOMENT ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We observe that the traditional use of DP with the Adam optimizer introduces a bias in the second moment estimation, due to the addition of independent noise in the gradient computation. This bias leads to a different scaling for low variance parameter updates, that is inconsistent with the behavior of non-private Adam, and Adam's sign descent interpretation. Empirically, correcting the bias introduced by DP noise significantly improves the optimization performance of DP-Adam.

1 INTRODUCTION

The Adam optimization algorithm (Kingma & Ba, 2014) is the default optimizer for several deep learning architectures and tasks. However we observe that when combined with Differential Privacy (DP), Adam does not perform as well: it suffers a large degradation of performance compared to SGD on vision tasks, and NLP tasks perform poorly when training from scratch. Recent empirical investigations suggest the performance of Adam may be linked to its update rule performing a smooth version of the sign descent update (Kunstner et al., 2023). The key components of Adam's update are 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. This additive change to the second moment estimate moves the Adam update away from that of sign descent, by scaling the gradient dimensions with different magnitudes differently. We show that we can correct for this DP noise induced bias. Empirically, correcting Adam's second moment estimate for DP noise significantly increases test performance for Adam with Differential Privacy, on both vision and NLP tasks.

2 THE ADAM UPDATE UNDER DIFFERENTIAL PRIVACY

The Adam update (Kingma & Ba, 2014) is defined as follows, let $g_t = (1/B)\nabla f(\theta_{t-1})$ be the mean gradient over a mini-batch of size B with respect to loss function f at step t; let β_1 and β_2 be the decay coefficients: at each step, Adam updates $m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1)\, g_t,\ \ \widehat{m}_t \leftarrow m_t/\left(1-\beta_1^t\right)$ and $v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2)\, g_t^2,\ \ \widehat{v}_t \leftarrow v_t/\left(1-\beta_2^t\right)$. The model's parameters are updated as $\theta_t \leftarrow \theta_{t-1} - \eta\Delta_t$, with learning rate η, $\Delta_t = \hat{m}_t/(\sqrt{\hat{v}_t}+\gamma)$ the update direction, and $\gamma > 0$ a small constant added for numerical stability. Intuitively, Adam's m_t and 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.

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 (e.g., NLP) where sign descent also performs well, at least when running on 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:

(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* $|\mathbb{E}[g_t]|_i$.

(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 $[0, 1]$, 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 with a constant size ± 1 update, and rescales the update of parameters with high variance gradients towards 0. As we describe next, adding Differential Privacy to gradient computations breaks this interpretation of Adam as sign descent.

Using Adam with Differential Privacy. Most optimization approaches for deep learning models with Differential Privacy (DP) follow a common recipe: compute each update (averaged gradients 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, let C, σ be the maximum $L2$-norm clipping value and the noise multiplier respectively. For a mini-batch B, $\overline{g}_t = (1/B)\sum_n g_n / \max\left(1, \|g_n\|_2/C\right)$ is the mean of clipped gradient, which is a biased estimate of g_t. Then, the private gradient update is $\tilde{g}_t = \overline{g}_t + (1/B)z_t,\ z_t \sim \mathcal{N}(0, \sigma^2 C^2 \mathbb{I}^d)$. With this recipe, any optimizer that only takes mini-batch updates as input, such as Adam, can be applied to the DP update. 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 $m_t^p \leftarrow \beta_1 m_{t-1}^p + (1-\beta_1)\,\tilde{g}_t,\ \ \widehat{m}_t^p \leftarrow m_t^p/\left(1-\beta_1^t\right),\ \ v_t^p \leftarrow \beta_2 v_{t-1}^p + (1-\beta_2)\,\tilde{g}_t^2,\ \ \widehat{v}_t^p \leftarrow v_t/\left(1-\beta_2^t\right), \theta_t \leftarrow \theta_{t-1} - \eta \hat{m}_t/(\sqrt{\hat{v}_t} + \gamma)$.

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, 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:

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

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:

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

where $\mathbb{E}[m_t^c]$ and $\mathbb{E}[v_t^c]$ are the quantities estimated under regular Adam, computed with respect to $\bar{g}_t$ (due to DP clipping). Under exact expectations, the Adam update becomes $\Delta_t = \mathbb{E}[\tilde{g}_t]/\sqrt{\mathbb{E}[\tilde{g}_t^2]} = \mathbb{E}[\bar{g}_t]/\sqrt{\mathbb{V}[\bar{g}_t] + \mathbb{E}(\bar{g}_t)^2 + \Phi}$.

To understand the implication of DP noise bias Φ, let us follow Kingma & Ba (2014) and interpret the update under the assumption that $\mathbb{E}[m_t^c] \approx (1-\beta_1^t)\mathbb{E}[\bar{g}_t]$ and $\mathbb{E}[v_t^c] \approx (1-\beta_2^t)\mathbb{E}[\bar{g}_t^2]$—e.g., the first and second moment $\mathbb{E}[\bar{g}_\tau], \mathbb{E}[\bar{g}_\tau^2], \tau = 0, 1, 2, \ldots, t$ are stationary; or decay coefficients β_1, β_2 are such that past g_τ are assigned small weights, t is large enough, and β_1, β_2 are such that $1 - \beta_1^t/\sqrt{1-\beta_2^t} = 1$. Let Φ be the bias accumulated from DP noise variance up to step t in v_t^p (Eq. (2)). This bias in the second moment rescales the Adam update, which becomes incompatible with the sign descent interpretation. 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.

Correcting for DP noise in DP-Adam. 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:

$$\Delta_t = \hat{m}_t / \sqrt{\max\left(\hat{v}_t - (\sigma C/B)^2, \gamma'\right)}. \tag{3}$$

Equation 3 enables a sign descent interpretation for DP-Adam which closely tracks that of Adam. Ignoring the stochasticity introduced by measurements with DP noise for now, we have that:

(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), $\Delta_t \in [0, 1]$, 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 $\mathbb{V}[\bar{g}_t]_i$).

Of course, the exponential moving averages over DP noise introduce measurement errors: it is 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, and we expect that $\sqrt{\gamma'} \gg \gamma$. While it can still happen that $|\Delta_{i,t}| \geq 1$, we show in §3 that debiasing the second moment to follow the sign descent interpretation yields an important improvement in model accuracy. Algorithm 1 shows the modified DP-Adam with a corrected estimate for the second moments.

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

Privacy Analysis. Our bias corrected version of DP-Adam follows the same DP analysis as that of Adam without correction, 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, ensures privacy. The correction is based only on public parameters of DP-Adam: β_2, step t, batch size B, and the DP noise variance $(\sigma C)^2$. In experiments (§3) we use Rényi DP for composition, though other techniques would also apply.

3 THE EMPIRICAL EFFECT OF CORRECTING FOR DP NOISE

Performance of DP-Adam, DP-Adam-Biased and DP-SGD. We compare the performance of DP-Adam, DP-Adam-Biased (with no correction in v_t^p) and DP-SGD on image and text classification tasks with CIFAR10 (Krizhevsky et al.) and SNLI (Bowman et al., 2015) datasets. We test under 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 block are initialized randomly and the other layers inherit weights from pre-trained BERT-base model (Devlin et al., 2018). For each optimizer, we tune the learning rate and γ (or γ') at a coarse granularity, for $\epsilon \approx 7$. The final hyperparamters are: for CIFAR10 $C = 1, \sigma = 1, B = 2048$, η, γ (or γ') for DP-SGD/DP-Adam-Biased/DP-Adam are respectively 4/0.001/0.001, (not applicable)/1e-8/1e-8; for SNLI $C = 0.1, \sigma = 0.4, B = 256$, η, γ (or γ') for DP-SGD/DP-Adam-Biased/DP-Adam are respectively 4/0.01/0.001, (not applicable)/1e-8/1e-9. Figure 1 shows that for CIFAR10, on which Adam often performs worse than SGD in non-private settings, DP-Adam-Biased performs much worse than DP-SGD, whereas DP-Adam brings the performance close to that of DP-SGD. For SNLI, DP-Adam performs better than DP-Adam-Biased and both are better than DP-SGD.

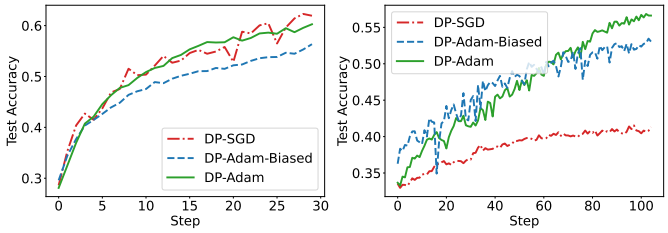

Figure 1: Comparing the performance of DP-Adam, DP-Adam-Biased and DP-SGD on **Left:** CIFAR10 and **Right:** SNLI. At the end of training ϵ-DP ≈ 7 in both experiments.

		Min	Q1	Median	Q3	Max	Mean
t = 5000	v_t^c	4.757e-21	1.802e-13	6.333e-13	1.481e-12	2.647e-8	4.050e-12
	v_t^p	2.025e-8	2.363e-8	2.426e-8	2.463e-8	5.324e-8	2.436e-8
t = 20000	v_t^c	6.584e-22	5.867e-14	2.925e-13	8.372e-13	1.060e-8	3.673e-12
	v_t^p	2.065e-8	2.408e-8	2.460e-8	2.513e-8	3.657e-8	2.461e-8

Table 1: Summary statistics of v_t^p with the SNLI dataset.

First and second moment estimates of clipped and private gradients. We compare the scale of v_t^c and v_t^p by measuring their summary statistics at the early ($t = 5000$) and late ($t = 20000$) training stages. In this experiment, $B = 256, C = 0.1, \sigma = 0.4, \beta_2 = 0.999$, $\Phi \approx 2.441$e-8. The results on the SNLI dataset are summarized in Table 1. We observe that both the scale and spread of v_t^p is quite different from that of v_t^c, which suggests that the values of v_t^p are largely affected by the DP noise. If no correction is imposed, Φ dominates the size of $\mathbb{V}[\overline{g}_t]$, and the tuned learning rate is larger to compensate. Since Φ dominates, the update Δ_t is proportional to the first moment $\mathbb{E}[\overline{g}_t]$ (§2). This is not compatible with the behavior of sign descent.

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 2(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 hyperparameters except changing value of Φ to Φ'. 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 $\mathbb{E}[\overline{g}_t^2]$.

Effect of the numerical stability constant. The numerical stability constant γ in known to affect the performance of Adam in the non-private setting, and γ is often tuned as a hyperparameter (Reddi et al., 2019). Following the same logic, we test the effect of γ' and γ on the performance of DP-Adam and DP-Adam-Biased. Figure 2 (Middle) 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 2 (Right) shows the effect of tuning γ in DP-Adam-Biased. We observe that it has a smaller effect than with DP-Adam, since the large scale of Φ makes the estimates of v_t^p relatively large and similar among parameters. We also observe that tuning γ with DP-Adam-Biased does not lead to the same effect as correcting Φ in DP-Adam, and DP-Adam achieves higher accuracy.

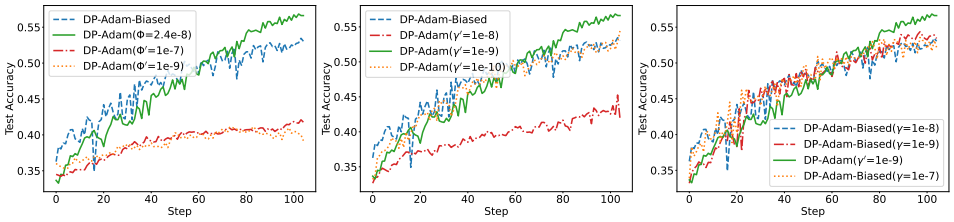

Figure 2: Compare the performance when **Left:** subtracting different (fake) values of Φ, **Middle:** tuning γ in DP-Adam, **Right:** tuning γ in DP-Adam-Biased.

REFERENCES

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL https://arxiv.org/abs/1810.04805.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Heavy-tailed noise does not explain the gap between SGD and adam, but sign descent might. In *International Conference on Learning Representations*, 2023.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners, 2021. URL https://arxiv.org/abs/2110.05679.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond, 2019. URL https://arxiv.org/abs/1904.09237.