DP-Adam: Correcting DP Bias in Adam's Second Moment Estimation

Qiaoyue Tang Mathias Lécuyer

University of British Columbia

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.

The Adam Update under Differential Privacy

Interpretation of Adam as Sign Descent

- Adam maintains exponential moving average for estimating $\mathbb{E}[g_t]$ and $\mathbb{E}[g_t^2]$: the vector of first and second moment of updates to each parameter during training.
- Previous evidence supports the hypothesis: Adam may derive its empirical performance from being a smoothed out version of sign descent.

Using Adam with Differential Privacy

- Existing DP approaches using Adam substitutes mini-batch gradient g_t with privatized $\tilde{g}_t$ to preserve privacy.
- $g_n = \nabla f(\theta_t, x_n)$ is the gradient for sample n; B, C, σ are batch size, maximum $L2$-norm clipping value and noise multiplier,

$$\bar{g}_t = (1/B)\sum_{n\in B} g_n/\max\left(1, \|g_n\|_2/C\right),$$
$$\tilde{g}_t = \bar{g}_t + (1/B)z_t,\ \ z_t \sim \mathcal{N}(0, \sigma^2 C^2 \mathbb{I}^d).$$

DP Noise Shifts Second Moment Estimates

The independent DP noise has no impact on the first moment in expectation,

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

v_t^p is a biased estimate of the second moment of the mini-batch clipped gradient $\bar{g}_t$,

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

When $|\mathbb{E}[\bar{g}_t]|_i \approx \sqrt{\mathbb{E}[\bar{g}_t^2]}_i$, the Adam update becomes $\pm\frac{|\mathbb{E}[\bar{g}_t]|_i}{\sqrt{\mathbb{E}[\bar{g}_t^2]_i + \Phi}}$ instead of ± 1.

Correcting for DP noise in DP-Adam

We correct for this bias by changing the Adam update Δ_t as:

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

It enables a sign descent interpretation for DP-Adam closely follows Adam, **except that the variance is compared to Φ instead of 0**:

For each parameter i,

- If $|\mathbb{E}[\bar{g}_t]|_i \approx 0$:
  - If $\mathrm{Var}\,[\bar{g}_t]_i \gg \Phi$, then $\Delta_t \approx 0$.
  - If $\mathrm{Var}\,[\bar{g}_t]_i \lesssim \Phi$, the γ' parameter ensures $\Delta_t \approx 0$.
- If $|\mathbb{E}[\bar{g}_t]|_i \gg 0$, then our correction restores the smoothed sign descent behavior of DP-Adam:
  - If $\mathrm{Var}\,[\bar{g}_t]_i \gg \Phi$, $|\mathbb{E}[\bar{g}_t]|_i \approx \sqrt{\mathbb{E}[\bar{g}_t^2]}_i$, and $\Delta_t \approx \pm 1$.
  - If $\mathrm{Var}\,[\bar{g}_t]_i \lesssim \Phi$, $\Delta_t \in [0, 1]$, performing a smooth (variance scaled) version of sign descent.

Privacy Analysis

The privacy guarantee holds following the post-processing property of DP, and composition over training iterations.

Algorithm: 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; initialize moment estimates $m_0 = 0, v_0 = 0$;
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

The Empirical Effect of Correcting for DP Noise

Performance of Uncorrected, Corrected DP-Adam and DP-SGD

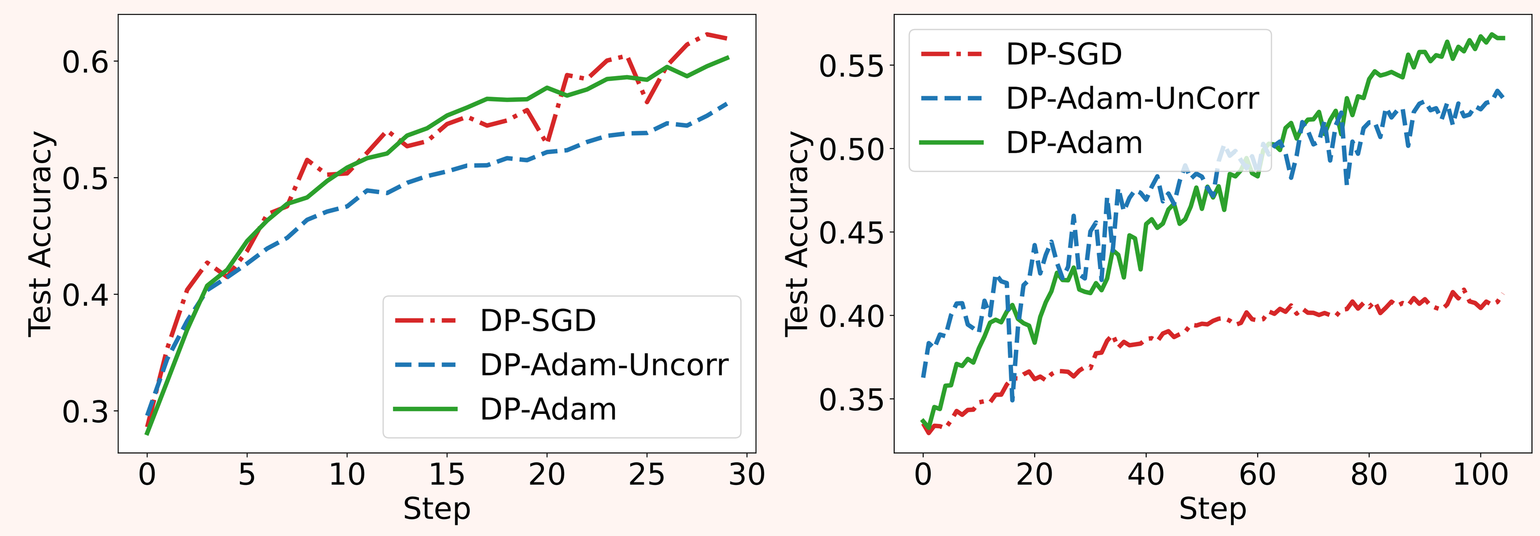

Figure 1. Comparing the performance of DP-Adam, DP-Adam-Uncorr and DP-SGD on **Left:** CIFAR10 and **Right:** SNLI. Tested under training-from-scratch setting. η, γ (or γ') are tuned at a coarse granularity for $\epsilon \approx 7$.

First and Second Moment Estimates of Clipped and Private Gradients

		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.

- Scale and spread of v_t^p is different from that of v_t^c, suggesting v_t^p largely affected by the DP noise,
- Φ dominates the size of $\mathrm{Var}\,[\bar{g}_t]$ (and hence v_t^p) thus making Δ_t smaller.

Correcting Second Moment with Different Values

- Correcting for a different value (of $\Phi' > \Phi$ or $\Phi' < \Phi$) does not provide a good estimate for $\mathrm{Var}\,[\bar{g}_t]$.

Effect of the Numerical Stability Constant

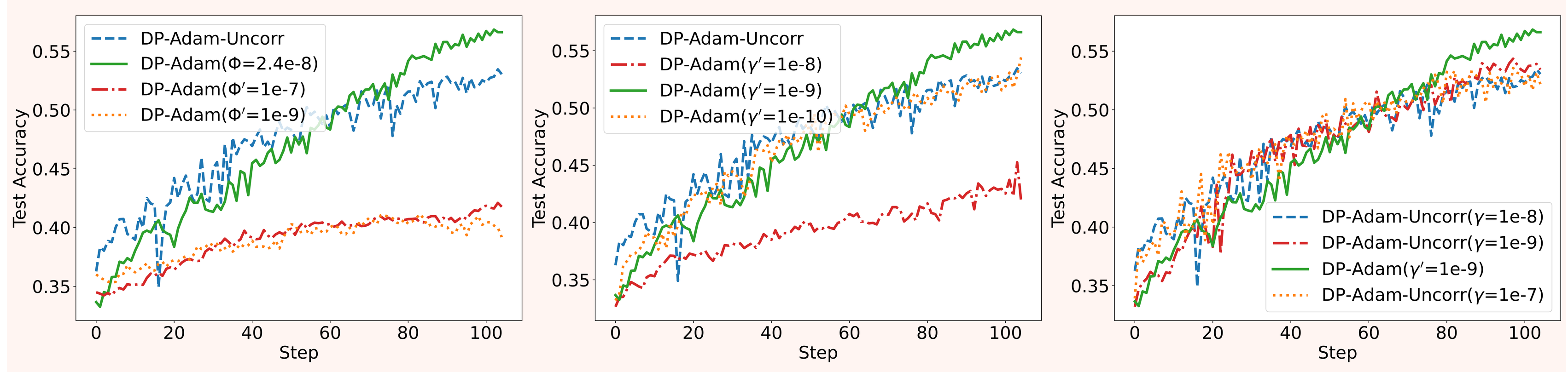

Figure 2. Compare the performance when **Left:** subtracting different (fake) values of Φ, **Middle:** tuning γ' in DP-Adam, **Right:** tuning γ in DP-Adam-Uncorr.

- γ' impacts the performance of DP-Adam: v_t^p are small, changing γ' avoids magnifying parameters with tiny estimates of v_t^c,
- Tuning γ with DP-Adam-Uncorr does not lead to the same effect as correcting Φ in DP-Adam, and DP-Adam achieves higher accuracy.