DP-AdamBC:

DP-Adam is actually DP-SGD (unless you apply Bias Correction)

Qiaoyue Tang
University of British Columbia

DP-AdamBC:

DP-Adam is actually DP-SGD (unless you apply Bias Correction)

Qiaoyue Tang
University of British Columbia

Frederick Shpilevskiy

Mathias Lécuyer

RUNNING EXAMPLE

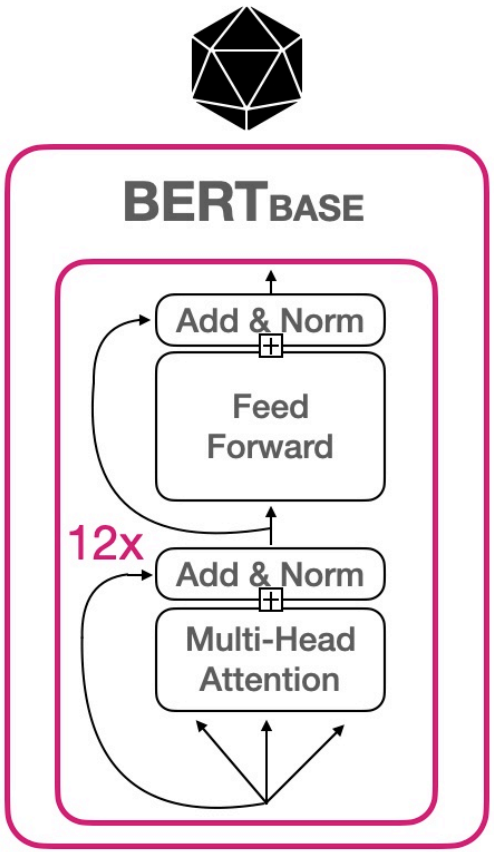
Imagine that we want to solve the following task (SNLI dataset).

Premise	Hypothesis	Label
Children smiling and waving at camera	They are smiling at their parents	neutral
Children smiling and waving at camera	There are children present	entailment
Children smiling and waving at camera	The kids are frowning	contradiction

RUNNING EXAMPLE

Imagine that we want to solve the following task (SNLI dataset).

Premise	Hypothesis	Label
Children smiling and waving at camera	They are smiling at their parents	neutral
Children smiling and waving at camera	There are children present	entailment
Children smiling and waving at camera	The kids are frowning	contradiction



110M Parameters

https://huggingface.co

Now, we need to train the model (Optimization background)

Stochastic gradient descent (SGD)

Initialize θ_0

for $t = 1 \ldots T$ **do**

$$g_t = \frac{1}{|B|} \sum_{x \in B} \nabla_\theta \ell(\theta_{t-1}, x)$$

$$\Delta_t = g_t$$

$$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$$

Return θ_T

Stochastic gradient descent (SGD)

Initialize θ_0

for $t = 1 \ldots T$ **do**

$g_t = \frac{1}{|B|} \sum_{x \in B} \nabla_\theta \ell(\theta_{t-1}, x)$ ← Compute gradient on minibatch B

$\Delta_t = g_t$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

Return θ_T

Stochastic gradient descent (SGD)

Initialize θ_0

for $t = 1 \ldots T$ **do**

$g_t = \frac{1}{|B|} \sum_{x \in B} \nabla_\theta \ell(\theta_{t-1}, x)$

$\Delta_t = g_t$ ← The parameter update is the gradient

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

Return θ_T

Stochastic Gradient Descent (SGD)

Initialize θ_0

for $t = 1 \ldots T$ **do**

$g_t = \frac{1}{|B|} \sum_{x \in B} \nabla_\theta \ell(\theta_{t-1}, x)$

$\Delta_t = g_t$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$ ← Move opposite the update by LR η

Return θ_T

Stochastic gradient descent (SGD)

Initialize θ_0

for $t = 1 \ldots T$ **do**

$g_t = \frac{1}{|B|} \sum_{x \in B} \nabla_\theta \ell(\theta_{t-1}, x)$

$\Delta_t = g_t$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

Return θ_T

If you have trained language models before, you may want to use a variation of SGD.

Stochastic gradient descent with momentum (SGDM)

There are two variations of interest today: SGDM

Initialize θ_0, $m_0 = 0$

for $t = 1...T$ **do**

$g_t = \frac{1}{|B|} \sum_{x \in B} \nabla_\theta \ell(\theta_{t-1}, x)$

$m_t \leftarrow \beta m_{t-1} + (1 - \beta) g_t$

$\Delta_t = m_t$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

Return θ_T

STOCHASTIC GRADIENT DESCENT WITH MOMENTUM (SGDM)

There are two variations of interest today: SGDM

Initialize θ_0, $m_0 = 0$

for $t = 1 \ldots T$ **do**

$g_t = \frac{1}{|B|} \sum_{x \in B} \nabla_\theta \ell(\theta_{t-1}, x)$

$m_t \leftarrow \beta m_{t-1} + (1 - \beta) g_t$ $\approx \mathbb{E}[g_t]$

$\Delta_t = m_t$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

Return θ_T

Stochastic gradient descent with momentum (SGDM)

There are two variations of interest today: SGDM

Initialize θ_0, $m_0 = 0$

for $t = 1 \ldots T$ **do**

$g_t = \frac{1}{|B|} \sum_{x \in B} \nabla_\theta \ell(\theta_{t-1}, x)$

$m_t \leftarrow \beta m_{t-1} + (1 - \beta) g_t$ $\approx \mathbb{E}[g_t]$

$\Delta_t = m_t$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

Return θ_T

Probably still not the optimizer you want for language models.

ADAM (SIMPLIFIED FOR THIS DISCUSSION)

There are two variations of interest today: SGDM, ADAM

Initialize θ_0, $m_0 = 0$, $v_0 = 0$

for $t = 1 \ldots T$ **do**

$g_t = \frac{1}{|B|} \sum_{x \in B} \nabla_\theta \ell(\theta_{t-1}, x)$

$m_t \leftarrow \beta m_{t-1} + (1-\beta) g_t$ $\approx \mathbb{E}[g_t]$

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

$\Delta_t = \frac{m_t}{\sqrt{v_t} + \gamma}$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

Return θ_T

ADAM (SIMPLIFIED FOR THIS DISCUSSION)

There are two variations of interest today: SGDM, ADAM

Initialize θ_0, $m_0 = 0$, $v_0 = 0$

for $t = 1 \ldots T$ **do**

$g_t = \frac{1}{|B|} \sum_{x \in B} \nabla_\theta \ell(\theta_{t-1}, x)$

$m_t \leftarrow \beta m_{t-1} + (1-\beta) g_t$

$v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2) g_t^2$ $\approx \mathbb{E}[g_t^2]$

$\Delta_t = \frac{m_t}{\sqrt{v_t} + \gamma}$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

Return θ_T

ADAM (SIMPLIFIED FOR THIS DISCUSSION)

There are two variations of interest today: SGDM, ADAM

Initialize θ_0, $m_0 = 0$, $v_0 = 0$

for $t = 1 \ldots T$ **do**

$g_t = \frac{1}{|B|} \sum_{x \in B} \nabla_\theta \ell(\theta_{t-1}, x)$

$m_t \leftarrow \beta m_{t-1} + (1 - \beta) g_t$

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

$\Delta_t = \frac{m_t}{\sqrt{v_t} + \gamma}$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

Return θ_T

ADAM (SIMPLIFIED FOR THIS DISCUSSION)

There are two variations of interest today: SGDM, ADAM

Initialize θ_0, $m_0 = 0$, $v_0 = 0$

for $t = 1 \ldots T$ **do**

$g_t = \frac{1}{|B|} \sum_{x \in B} \nabla_\theta \ell(\theta_{t-1}, x)$

$m_t \leftarrow \beta m_{t-1} + (1 - \beta) g_t$

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

$\Delta_t = \frac{m_t}{\sqrt{v_t} + \gamma}$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

Return θ_T

Why would this be a good idea?

ADAM (SIMPLIFIED FOR THIS DISCUSSION)

There are two variations of interest today: SGDM, ADAM

Initialize θ_0, $m_0 = 0$, $v_0 = 0$

for $t = 1 \ldots T$ **do**

$g_t = \frac{1}{|B|} \sum_{x \in B} \nabla_\theta \ell(\theta_{t-1}, x)$

$m_t \leftarrow \beta m_{t-1} + (1 - \beta) g_t$

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

$\Delta_t = \frac{m_t}{\sqrt{v_t} + \gamma}$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

Return θ_T

Why would this be a good idea? Adam works when sign descent works.

ADAM (SIMPLIFIED FOR THIS DISCUSSION)

There are two variations of interest today: SGDM, ADAM

Initialize θ_0, $m_0 = 0$, $v_0 = 0$

for $t = 1 \ldots T$ **do**

$g_t = \frac{1}{|B|} \sum_{x \in B} \nabla_\theta \ell(\theta_{t-1}, x)$

$m_t \leftarrow \beta m_{t-1} + (1-\beta) g_t$

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

$\Delta_t = \frac{m_t}{\sqrt{v_t} + \gamma}$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

$\Delta_t \approx \frac{\mathbb{E}[g_t]}{\sqrt{\mathbb{E}[g_t^2]}}$

Return θ_T

Why would this be a good idea? Adam works when sign descent works.

ADAM (SIMPLIFIED FOR THIS DISCUSSION)

There are two variations of interest today: SGDM, ADAM

Initialize θ_0, $m_0 = 0$, $v_0 = 0$

for $t = 1 \ldots T$ **do**

$g_t = \frac{1}{|B|} \sum_{x \in B} \nabla_\theta \ell(\theta_{t-1}, x)$

$m_t \leftarrow \beta m_{t-1} + (1-\beta) g_t$

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

$\Delta_t = \frac{m_t}{\sqrt{v_t} + \gamma}$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

$$\Delta_t \approx \frac{\mathbb{E}[g_t]}{\sqrt{\mathbb{E}[g_t^2]}} \approx \frac{\mathbb{E}[g_t]}{\sqrt{\mathbb{E}[g_t]^2 + \mathbb{V}[g_t]}}$$

Return θ_T

Why would this be a good idea? Adam works when sign descent works.

ADAM (SIMPLIFIED FOR THIS DISCUSSION)

There are two variations of interest today: SGDM, ADAM

Initialize θ_0, $m_0 = 0$, $v_0 = 0$

for $t = 1 \ldots T$ **do**

$$g_t = \frac{1}{|B|} \sum_{x \in B} \nabla_\theta \ell(\theta_{t-1}, x)$$

$$m_t \leftarrow \beta m_{t-1} + (1-\beta) g_t$$

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

$$\Delta_t = \frac{m_t}{\sqrt{v_t} + \gamma}$$

$$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$$

Return θ_T

$$\Delta_t \approx \frac{\mathbb{E}[g_t]}{\sqrt{\mathbb{E}[g_t^2]}} \approx \frac{\mathbb{E}[g_t]}{\sqrt{\mathbb{E}[g_t]^2 + \mathbb{V}[g_t]}}$$

$$\approx \begin{cases} \pm 1, & \text{if } \mathbb{V}[g_t] \gg \mathbb{E}[g_t]^2 \end{cases}$$

Why would this be a good idea? Adam works when sign descent works.

ADAM (SIMPLIFIED FOR THIS DISCUSSION)

There are two variations of interest today: SGDM, ADAM

Initialize θ_0, $m_0 = 0$, $v_0 = 0$

for $t = 1 \ldots T$ **do**

$g_t = \frac{1}{|B|} \sum_{x \in B} \nabla_\theta \ell(\theta_{t-1}, x)$

$m_t \leftarrow \beta m_{t-1} + (1-\beta) g_t$

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

$\Delta_t = \frac{m_t}{\sqrt{v_t}+\gamma}$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

Return θ_T

$$\Delta_t \approx \frac{\mathbb{E}[g_t]}{\sqrt{\mathbb{E}[g_t^2]}} \approx \frac{\mathbb{E}[g_t]}{\sqrt{\mathbb{E}[g_t]^2 + \mathbb{V}[g_t]}}$$

$$\approx \begin{cases} \pm 1, & \text{if } \mathbb{V}[g_t] \gg \mathbb{E}[g_t]^2 \\ \to 0, & \text{when } \mathbb{V}[g_t] \ll \mathbb{E}[g_t]^2 \end{cases}$$

Why would this be a good idea? Adam works when sign descent works.

ADAM (SIMPLIFIED FOR THIS DISCUSSION)

There are two variations of interest today: SGDM, ADAM

Initialize θ_0, $m_0 = 0$, $v_0 = 0$

for $t = 1 \ldots T$ **do**

$g_t = \frac{1}{|B|} \sum_{x \in B} \nabla_\theta \ell(\theta_{t-1}, x)$

$m_t \leftarrow \beta m_{t-1} + (1-\beta) g_t$

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

$\Delta_t = \frac{m_t}{\sqrt{v_t} + \gamma}$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

Return θ_T

$$\Delta_t \approx \frac{\mathbb{E}[g_t]}{\sqrt{\mathbb{E}[g_t^2]} + \gamma} \approx \frac{\mathbb{E}[g_t]}{\sqrt{\mathbb{E}[g_t]^2 + \mathbb{V}[g_t]} + \gamma}$$

$$\approx \begin{cases} \pm 1, & \text{if } \mathbb{V}[g_t] \gg \mathbb{E}[g_t]^2 \\ \to 0, & \text{when } \mathbb{V}[g_t] \ll \mathbb{E}[g_t]^2 \end{cases}$$

Why would this be a good idea? Adam works when sign descent works.

ADAM (SIMPLIFIED FOR THIS DISCUSSION)

There are two variations of interest today: SGDM, ADAM

Initialize θ_0, $m_0 = 0$, $v_0 = 0$

for $t = 1 \ldots T$ **do**

$g_t = \frac{1}{|B|} \sum_{x \in B} \nabla_\theta \ell(\theta_{t-1}, x)$

$m_t \leftarrow \beta m_{t-1} + (1-\beta) g_t$

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

$\Delta_t = \frac{m_t}{\sqrt{v_t}+\gamma}$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

Return θ_T

$$\Delta_t \approx \frac{\mathbb{E}[g_t]}{\sqrt{\mathbb{E}[g_t^2]}+\gamma} \approx \frac{\mathbb{E}[g_t]}{\sqrt{\mathbb{E}[g_t]^2+\mathbb{V}[g_t]}+\gamma}$$

$$\approx \begin{cases} \pm 1, & \text{if } \mathbb{V}[g_t] \gg \mathbb{E}[g_t]^2 \\ \to 0, & \text{when } \mathbb{V}[g_t] \ll \mathbb{E}[g_t]^2 \\ \to 0, & \text{when } |\mathbb{E}[g_t]| \ll \gamma \end{cases}$$

Why would this be a good idea? Adam works when sign descent works.

ADAM (SIMPLIFIED FOR THIS DISCUSSION)

We now have a good optimizer, and can train our model!

Is our data safe this way? (Optim + Differential Privacy)

LEARNING OUR MODEL WILL EXPOSE THE TRAINING DATA

Using the trained model, an attacker can:

LEARNING OUR MODEL WILL EXPOSE THE TRAINING DATA

Using the trained model, an attacker can:

- learn membership in the training set,

LEARNING OUR MODEL WILL EXPOSE THE TRAINING DATA

Using the trained model, an attacker can:

- learn membership in the training set,
- reconstruct training examples.

LEARNING OUR MODEL WILL EXPOSE THE TRAINING DATA

Using the trained model, an attacker can:

- learn membership in the training set,
- reconstruct training examples.

Differential Privacy (DP) prevents this leakage.

LEARNING OUR MODEL WILL EXPOSE THE TRAINING DATA

Using the trained model, an attacker can:

- learn membership in the training set,
- reconstruct training examples.

Differential Privacy (DP) prevents this leakage. And training deep learning models with DP is easy! (conceptually)

DP stochastic gradient descent (DP-SGD)

Initialize θ_0

for $t = 1 \ldots T$ **do**

$g_t = \frac{1}{|B|} \sum_{x \in B} \nabla_\theta \ell(\theta_{t-1}, x)$

$\Delta_t = g_t$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

Return θ_T

DP stochastic gradient descent (DP-SGD)

Initialize θ_0

for $t = 1 \ldots T$ **do**

$\bar{g}_t = \frac{1}{|B|} \sum_{x \in B} \text{clip}\left(\nabla_\theta \ell(\theta_{t-1}, x), C\right)$

$\Delta_t = g_t$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

Return θ_T

- Clip gradients to control sensitivity.

DP stochastic gradient descent (DP-SGD)

Initialize θ_0

for $t = 1 \ldots T$ **do**

$\bar{g}_t = \frac{1}{|B|} \sum_{x \in B} \text{clip}\left(\nabla_\theta \ell(\theta_{t-1}, x), C\right)$

$\tilde{g}_t = \bar{g}_t + z, \quad z \sim \mathcal{N}(0, \sigma^2 \mathbb{I}^d)$

$\Delta_t = g_t$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

Return θ_T

- Clip gradients to control sensitivity.
- Add DP noise.

DP stochastic gradient descent (DP-SGD)

Initialize θ_0

for $t = 1 \ldots T$ **do**

$\bar{g}_t = \frac{1}{|B|} \sum_{x \in B} \text{clip}\left(\nabla_\theta \ell(\theta_{t-1}, x), C\right)$

$\tilde{g}_t = \bar{g}_t + z, \quad z \sim \mathcal{N}(0, \sigma^2 \mathbb{I}^d)$

$\Delta_t = g_t$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

Return θ_T

- Clip gradients to control sensitivity.
- Add DP noise.

We can do this with all the optimizers we saw!

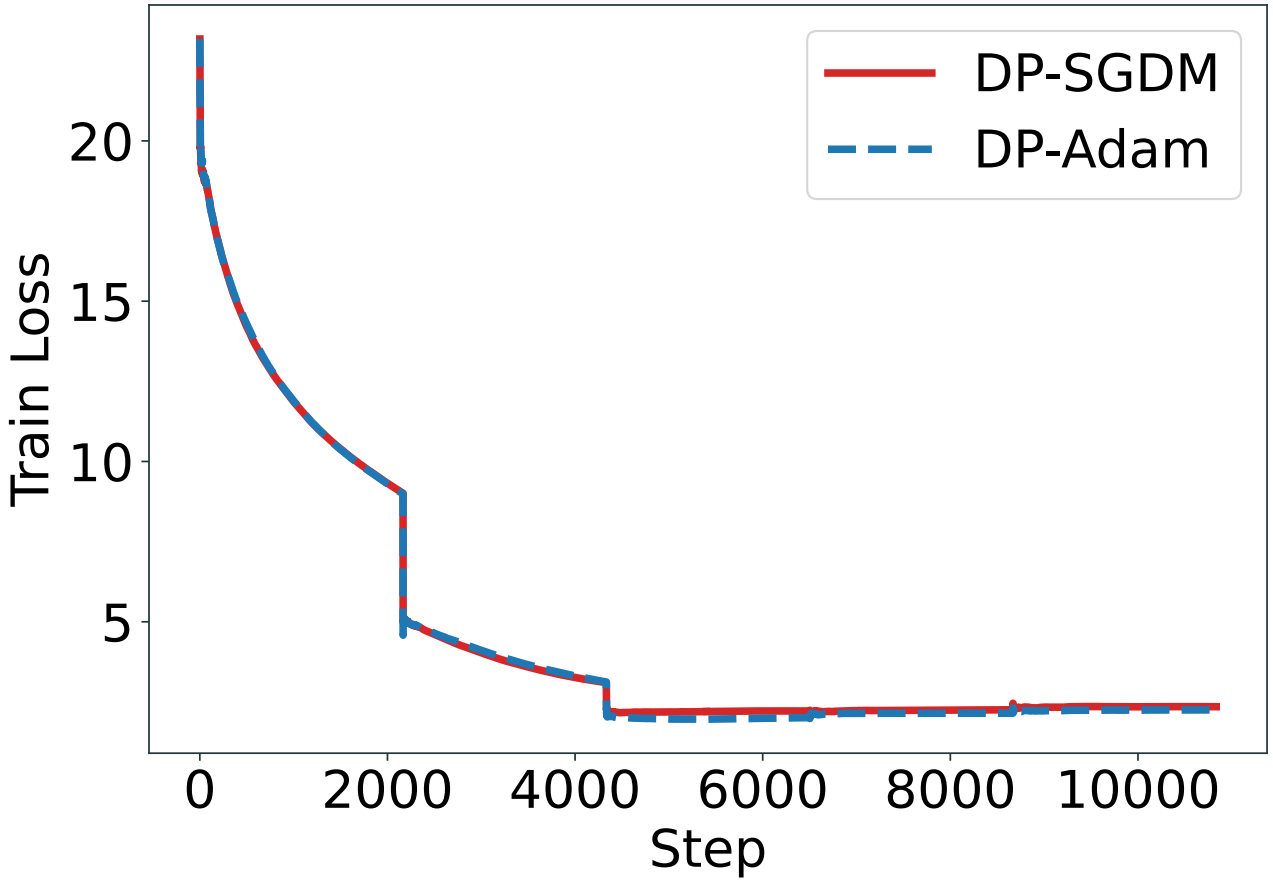
(DP-SGD, DP-SGDM, DP-Adam)

TESTING THE DP-* OPTIMIZERS

We implement the various DP optimizers, and tune their respective learning rates.

TESTING THE DP-* OPTIMIZERS

We implement the various DP optimizers, and tune their respective learning rates.



DP-Adam is Actually DP-SGDM

DP NOISE MAKES DP-ADAM BEHAVE LIKE DP-SGDM

Remember Adam's update: $\Delta_t \approx \frac{\mathbb{E}[g_t]}{\sqrt{\mathbb{E}[g_t^2]}}$.

DP NOISE MAKES DP-ADAM BEHAVE LIKE DP-SGDM

Remember Adam's update: $\Delta_t \approx \frac{\mathbb{E}[g_t]}{\sqrt{\mathbb{E}[g_t^2]}}$.

- Under DP, we would expect DP-Adam to update with $\Delta_t \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]}}$.

DP NOISE MAKES DP-ADAM BEHAVE LIKE DP-SGDM

Remember Adam's update: $\Delta_t \approx \frac{\mathbb{E}[g_t]}{\sqrt{\mathbb{E}[g_t^2]}}$.

- Under DP, we would expect DP-Adam to update with $\Delta_t \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]}}$.
- But we get $\Delta_t \approx \frac{\mathbb{E}[\tilde{g}_t]}{\sqrt{\mathbb{E}[\tilde{g}_t^2]}}$.

DP NOISE MAKES DP-ADAM BEHAVE LIKE DP-SGDM

Remember Adam's update: $\Delta_t \approx \frac{\mathbb{E}[g_t]}{\sqrt{\mathbb{E}[g_t^2]}}$.

- Under DP, we would expect DP-Adam to update with $\Delta_t \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]}}$.
- But we get $\Delta_t \approx \frac{\mathbb{E}[\tilde{g}_t]}{\sqrt{\mathbb{E}[\tilde{g}_t^2]}} = \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]+\sigma^2}}$.

DP NOISE MAKES DP-ADAM BEHAVE LIKE DP-SGDM

Remember Adam's update: $\Delta_t \approx \frac{\mathbb{E}[g_t]}{\sqrt{\mathbb{E}[g_t^2]}}$.

- Under DP, we would expect DP-Adam to update with $\Delta_t \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]}}$.
- But we get $\Delta_t \approx \frac{\mathbb{E}[\tilde{g}_t]}{\sqrt{\mathbb{E}[\tilde{g}_t^2]}} = \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]+\sigma^2}}$.

DP noise biases the second moment estimate.

DP NOISE MAKES DP-ADAM BEHAVE LIKE DP-SGDM

We expect DP-Adam's update to be: $\Delta_t \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]}}$, but have $\Delta_t \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]+\sigma^2}}$ instead.

DP NOISE MAKES DP-ADAM BEHAVE LIKE DP-SGDM

We expect DP-Adam's update to be: $\Delta_t \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]}}$, but have $\Delta_t \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]+\sigma^2}}$ instead.

- Under typical DP parameters, $\sigma^2 \gg \mathbb{E}[\bar{g}_t^2]_i$.

DP NOISE MAKES DP-ADAM BEHAVE LIKE DP-SGDM

We expect DP-Adam's update to be: $\Delta_t \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]}}$, but have $\Delta_t \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]+\sigma^2}}$ instead.

- Under typical DP parameters, $\sigma^2 \gg \mathbb{E}[\bar{g}_t^2]_i$.
- So $\Delta_t \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]+\sigma^2}}$.

DP NOISE MAKES DP-ADAM BEHAVE LIKE DP-SGDM

We expect DP-Adam's update to be: $\Delta_t \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]}}$, but have $\Delta_t \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]+\sigma^2}}$ instead.

- Under typical DP parameters, $\sigma^2 \gg \mathbb{E}[\bar{g}_t^2]_i$.
- So $\Delta_t \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]+\sigma^2}} \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\sigma^2}}$.

DP NOISE MAKES DP-ADAM BEHAVE LIKE DP-SGDM

We expect DP-Adam's update to be: $\Delta_t \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]}}$, but have $\Delta_t \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]+\sigma^2}}$ instead.

- Under typical DP parameters, $\sigma^2 \gg \mathbb{E}[\bar{g}_t^2]_i$.
- So $\Delta_t \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]+\sigma^2}} \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\sigma^2}} \approx \frac{\eta}{\sigma}\mathbb{E}[\bar{g}_t]$.

DP NOISE MAKES DP-ADAM BEHAVE LIKE DP-SGDM

We expect DP-Adam's update to be: $\Delta_t \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]}}$, but have $\Delta_t \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]+\sigma^2}}$ instead.

- Under typical DP parameters, $\sigma^2 \gg \mathbb{E}[\bar{g}_t^2]_i$.
- So $\Delta_t \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\mathbb{E}[\bar{g}_t^2]+\sigma^2}} \approx \frac{\mathbb{E}[\bar{g}_t]}{\sqrt{\sigma^2}} \approx \frac{\eta}{\sigma}\mathbb{E}[\bar{g}_t]$.

This is the update we expect from DP-SGDM! (with a rescaled learning rate $\eta' = \frac{\eta}{\sigma}$)

Can we get Adam back under DP?

DP-ADAMBC (simplified for this discussion)

We can correct for DP noise bias when estimating the gradient's second moment:

Initialize θ_0, $m_0 = 0$, $v_0 = 0$

for $t = 1 \ldots T$ **do**

$$\tilde{g}_t = \frac{1}{|B|} \sum_{x \in B} \text{clip}\left(\nabla_\theta l(\theta_{t-1}, x)\right) + z$$

$$m_t \leftarrow \beta m_{t-1} + (1 - \beta) g_t$$

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

$$\Delta_t = \frac{m_t}{\sqrt{\max(v_t - \sigma^2, \gamma')}}$$

$$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$$

Return θ_T

DP-ADAMBC (simplified for this discussion)

We can correct for DP noise bias when estimating the gradient's second moment:

Initialize θ_0, $m_0 = 0$, $v_0 = 0$

for $t = 1 \ldots T$ **do**

$\tilde{g}_t = \frac{1}{|B|} \sum_{x \in B} \text{clip}\left(\nabla_\theta l(\theta_{t-1}, x)\right) + z$

$m_t \leftarrow \beta m_{t-1} + (1-\beta) g_t$

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

$\Delta_t = \frac{m_t}{\sqrt{\max(v_t - \sigma^2, \gamma^2)}}$

Correct DP bias

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

Return θ_T

DP-ADAMBC (simplified for this discussion)

We can correct for DP noise bias when estimating the gradient's second moment:

Initialize θ_0, $m_0 = 0$, $v_0 = 0$

for $t = 1 \ldots T$ **do**

$\tilde{g}_t = \frac{1}{|B|} \sum_{x \in B} \text{clip}\left(\nabla_\theta l(\theta_{t-1}, x)\right) + z$

$m_t \leftarrow \beta m_{t-1} + (1-\beta) g_t$

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

$\Delta_t = \frac{m_t}{\sqrt{\max(v_t - \sigma^2, \gamma^2)}}$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

Return θ_T

Correct DP bias

$v_t - \sigma^2$ can be small or negative

DP-ADAMBC (simplified for this discussion)

We can correct for DP noise bias when estimating the gradient's second moment:

Initialize θ_0, $m_0 = 0$, $v_0 = 0$

for $t = 1 \ldots T$ **do**

$\tilde{g}_t = \frac{1}{|B|} \sum_{x \in B} \text{clip}\left(\nabla_\theta l(\theta_{t-1}, x)\right) + z$

$m_t \leftarrow \beta m_{t-1} + (1-\beta) g_t$

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

$\Delta_t = -\frac{m_t}{\sqrt{\max(v_t - \sigma^2, \gamma')}}$

$\theta_t \leftarrow \theta_{t-1} - \eta \Delta_t$

Return θ_T

Correct DP bias

$v_t - \sigma^2$ can be small or negative

Adapt "stability" constant

DP-AdamBC results

We show DP-AdamBC has some good properties:

- preserves privacy,

DP-AdamBC results

We show DP-AdamBC has some good properties:

- preserves privacy,
- is consistent,

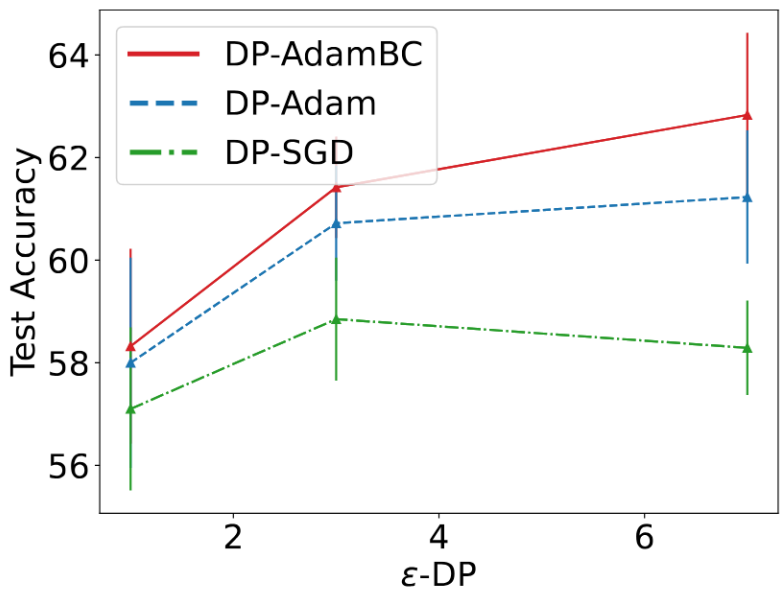
DP-AdamBC results

We show DP-AdamBC has some good properties:

- preserves privacy,
- is consistent,
- restores sign descent behaviour.

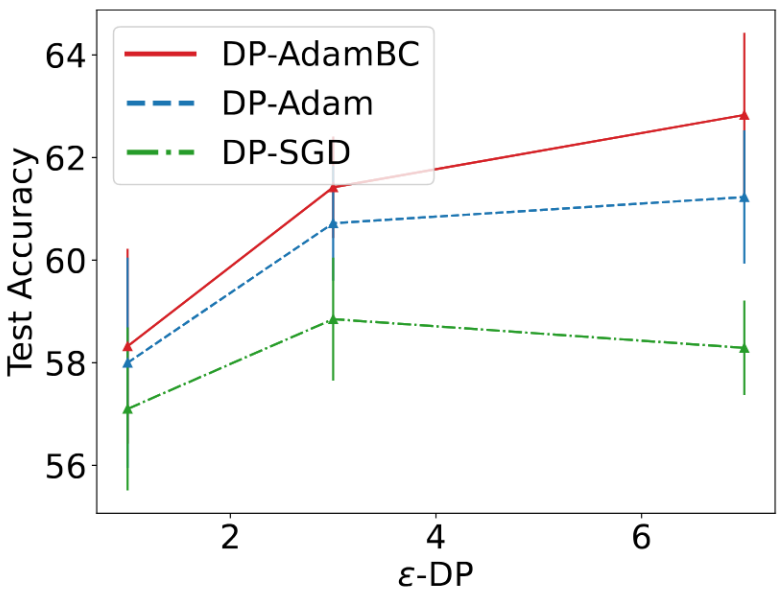
DP-AdamBC results

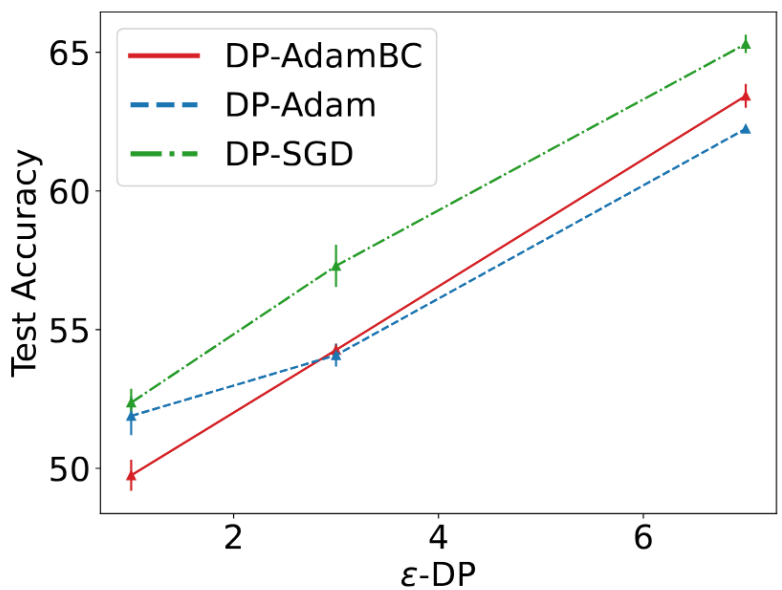
Test accuracy for SNLI data, BERT base model (left), CIFAR10 data, CNN model (right).



DP-AdamBC results

Test accuracy for SNLI data, BERT base model (left), CIFAR10 data, CNN model (right).





SUMMARY

- DP noise introduces bias in Adam's second moment estimates.

SUMMARY

- DP noise introduces bias in Adam's second moment estimates.
- DP-Adam breaks sign descent behaviour of Adam and behaves like DP-SGDM.

SUMMARY

- DP noise introduces bias in Adam's second moment estimates.
- DP-Adam breaks sign descent behaviour of Adam and behaves like DP-SGDM.
- DP-AdamBC: a practical correction to remove DP bias in estimating second moment of gradient.

SUMMARY

- DP noise introduces bias in Adam's second moment estimates.
- DP-Adam breaks sign descent behaviour of Adam and behaves like DP-SGDM.
- DP-AdamBC: a practical correction to remove DP bias in estimating second moment of gradient.
- Empirically DP-AdamBC has good performance especially when SGD underperforms.