MicroAdam: Accurate Adaptive Optimization with Low Space Overhead and Provable Convergence

Ionut-Vlad Modoranu[1], Mher Safaryan[1], Grigory Malinovsky[2], Eldar Kurtic[1], Thomas Robert[1], Peter Richtarik[2], Dan Alistarh[1]

[1] Institute of Science and Technology Austria (ISTA)
[2] King Abdullah University of Science and Technology (KAUST)

November 12th, 2024

Memory usage of Adam

- model size **d**
- two momentum buffers (**m**, **v**): **2d** additional memory
  - float32 (4 Bytes): **8d** (B)
  - bfloat16 (2 Bytes): **4d** (B)
- Existing work
  - AdamW-8bit: **2d** (B)
  - Can we do better?

MicroAdam

- Designed for finetuning
  - not all gradient entries are necessary for optimization

- Store a window of **m** sparse gradients
  - Top-K
  - 99% sparse
  - store largest 1% components (indices and values)

- Error Feedback
  - stores the compression error and corrects the gradient
  - quantized to 4 bits

- Constructs the Adam update dynamically at each step using CUDA kernels
  - We do not store **m** and **v**, so no additional memory used

- Memory footprint (bytes)
  - **0.9d** compared to 2d for AdamW-8bit

MicroAdam algorithm

Input: $\beta_1, \beta_2, \epsilon, \mathcal{G}, T, d, k$

$m_0, v_0 \leftarrow 0_d, 0_d$

$\delta_1, \Delta_1 \leftarrow 0, 0$

$e_1 \leftarrow 0_d^{4b}$

for $t = \{1, 2, ..., T\}$ **do**

$\quad g_t \leftarrow \widetilde{\nabla}_\theta f(\theta_t)$

$\quad a_t \leftarrow g_t + Q^{-1}(e_t, \delta_t, \Delta_t)$

$\quad \mathcal{I}_t, \mathcal{V}_t \leftarrow T_k(|a_t|)$

$\quad a_t[\mathcal{I}_t] \leftarrow 0$

$\quad \delta_{t+1}, \Delta_{t+1} \leftarrow min(a_t), max(a_t)$

$\quad e_{t+1} \leftarrow Q(a_t, \delta_{t+1}, \Delta_{t+1})$

$\quad \mathcal{G}_{i,:} \leftarrow (\mathcal{I}_t, \mathcal{V}_t)$

$\quad \hat{m}_t \leftarrow \textsc{AdamStats}(\beta_1, \mathcal{G})$

$\quad \hat{v}_t \leftarrow \textsc{AdamStats}(\beta_2, \mathcal{G}^2)$

$\quad \theta_{t+1} \leftarrow \theta_t - \eta_t \frac{\hat{m}_t}{\epsilon + \sqrt{\hat{v}_t}}$

$\quad i \leftarrow (i+1)\% m$

end for

MicroAdam algorithm

Input: $\beta_1, \beta_2, \epsilon, \mathcal{G}, T, d, k$
$m_0, v_0 \leftarrow 0_d, 0_d$
$\delta_1, \Delta_1 \leftarrow 0, 0$
$e_1 \leftarrow 0_d^{4b}$
for $t = \{1, 2, ..., T\}$ **do**
    $g_t \leftarrow \widetilde{\nabla}_\theta f(\theta_t)$
    $a_t \leftarrow g_t + Q^{-1}(e_t, \delta_t, \Delta_t)$
    $\mathcal{I}_t, \mathcal{V}_t \leftarrow T_k(|a_t|)$
    $a_t[\mathcal{I}_t] \leftarrow 0$
    $\delta_{t+1}, \Delta_{t+1} \leftarrow min(a_t), max(a_t)$
    $e_{t+1} \leftarrow Q(a_t, \delta_{t+1}, \Delta_{t+1})$
    $\mathcal{G}_{i,:} \leftarrow (\mathcal{I}_t, \mathcal{V}_t)$
    $\hat{m}_t \leftarrow \textsc{AdamStats}(\beta_1, \mathcal{G})$
    $\hat{v}_t \leftarrow \textsc{AdamStats}(\beta_2, \mathcal{G}^2)$
    $\theta_{t+1} \leftarrow \theta_t - \eta_t \frac{\hat{m}_t}{\epsilon + \sqrt{\hat{v}_t}}$
    $i \leftarrow (i+1)\%m$
end for

MicroAdam algorithm

Input: $\beta_1, \beta_2, \epsilon, \mathcal{G}, T, d, k$

$m_0, v_0 \leftarrow 0_d, 0_d$

$\delta_1, \Delta_1 \leftarrow 0, 0$

$e_1 \leftarrow 0_d^{4b}$

for $t = \{1, 2, ..., T\}$ **do**

$\quad g_t \leftarrow \widetilde{\nabla}_\theta f(\theta_t)$

$\quad a_t \leftarrow g_t + Q^{-1}(e_t, \delta_t, \Delta_t)$

$\quad \mathcal{I}_t, \mathcal{V}_t \leftarrow T_k(|a_t|)$

$\quad a_t[\mathcal{I}_t] \leftarrow 0$

$\quad \delta_{t+1}, \Delta_{t+1} \leftarrow min(a_t), max(a_t)$

$\quad e_{t+1} \leftarrow Q(a_t, \delta_{t+1}, \Delta_{t+1})$

$\quad \mathcal{G}_{i,:} \leftarrow (\mathcal{I}_t, \mathcal{V}_t)$

$\quad \hat{m}_t \leftarrow \text{AdamStats}(\beta_1, \mathcal{G})$

$\quad \hat{v}_t \leftarrow \text{AdamStats}(\beta_2, \mathcal{G}^2)$

$\quad \theta_{t+1} \leftarrow \theta_t - \eta_t \frac{\hat{m}_t}{\epsilon + \sqrt{\hat{v}_t}}$

$\quad i \leftarrow (i+1)\% m$

end for

MicroAdam algorithm

Input: $\beta_1, \beta_2, \epsilon, \mathcal{G}, T, d, k$

$m_0, v_0 \leftarrow 0_d, 0_d$

$\delta_1, \Delta_1 \leftarrow 0, 0$

$e_1 \leftarrow 0_d^{4b}$

for $t = \{1, 2, ..., T\}$ **do**

- $g_t \leftarrow \widetilde{\nabla}_\theta f(\theta_t)$
- $a_t \leftarrow g_t + Q^{-1}(e_t, \delta_t, \Delta_t)$
- $\mathcal{I}_t, \mathcal{V}_t \leftarrow T_k(|a_t|)$
- $a_t[\mathcal{I}_t] \leftarrow 0$
- $\delta_{t+1}, \Delta_{t+1} \leftarrow min(a_t), max(a_t)$
- $e_{t+1} \leftarrow Q(a_t, \delta_{t+1}, \Delta_{t+1})$
- $\mathcal{G}_{i,:} \leftarrow (\mathcal{I}_t, \mathcal{V}_t)$
- $\hat{m}_t \leftarrow \text{ADAMSTATS}(\beta_1, \mathcal{G})$
- $\hat{v}_t \leftarrow \text{ADAMSTATS}(\beta_2, \mathcal{G}^2)$
- $\theta_{t+1} \leftarrow \theta_t - \eta_t \frac{\hat{m}_t}{\epsilon + \sqrt{\hat{v}_t}}$
- $i \leftarrow (i+1)\%m$

end for

MicroAdam algorithm

Input: $\beta_1, \beta_2, \epsilon, \mathcal{G}, T, d, k$

$m_0, v_0 \leftarrow 0_d, 0_d$

$\delta_1, \Delta_1 \leftarrow 0, 0$

$e_1 \leftarrow 0_d^{4b}$

for $t = \{1, 2, ..., T\}$ **do**

$\quad g_t \leftarrow \widetilde{\nabla}_\theta f(\theta_t)$

$\quad a_t \leftarrow g_t + Q^{-1}(e_t, \delta_t, \Delta_t)$

$\quad \mathcal{I}_t, \mathcal{V}_t \leftarrow T_k(|a_t|)$

$\quad a_t[\mathcal{I}_t] \leftarrow 0$

$\quad \delta_{t+1}, \Delta_{t+1} \leftarrow min(a_t), max(a_t)$

$\quad e_{t+1} \leftarrow Q(a_t, \delta_{t+1}, \Delta_{t+1})$

$\quad \mathcal{G}_{i,:} \leftarrow (\mathcal{I}_t, \mathcal{V}_t)$

$\quad \hat{m}_t \leftarrow \text{ADAMSTATS}(\beta_1, \mathcal{G})$

$\quad \hat{v}_t \leftarrow \text{ADAMSTATS}(\beta_2, \mathcal{G}^2)$

$\quad \theta_{t+1} \leftarrow \theta_t - \eta_t \frac{\hat{m}_t}{\epsilon + \sqrt{\hat{v}_t}}$

$\quad i \leftarrow (i+1)\%m$

end for

MicroAdam algorithm

Input: $\beta_1, \beta_2, \epsilon, \mathcal{G}, T, d, k$

$m_0, v_0 \leftarrow 0_d, 0_d$

$\delta_1, \Delta_1 \leftarrow 0, 0$

$e_1 \leftarrow 0_d^{4b}$

for $t = \{1, 2, ..., T\}$ **do**

$\quad g_t \leftarrow \widetilde{\nabla}_\theta f(\theta_t)$

$\quad a_t \leftarrow g_t + Q^{-1}(e_t, \delta_t, \Delta_t)$

$\quad \mathcal{I}_t, \mathcal{V}_t \leftarrow T_k(|a_t|)$

$\quad a_t[\mathcal{I}_t] \leftarrow 0$

$\quad \delta_{t+1}, \Delta_{t+1} \leftarrow min(a_t), max(a_t)$

$\quad e_{t+1} \leftarrow Q(a_t, \delta_{t+1}, \Delta_{t+1})$

$\quad \mathcal{G}_{i,:} \leftarrow (\mathcal{I}_t, \mathcal{V}_t)$

$\quad \hat{m}_t \leftarrow \text{AdamStats}(\beta_1, \mathcal{G})$

$\quad \hat{v}_t \leftarrow \text{AdamStats}(\beta_2, \mathcal{G}^2)$

$\quad \theta_{t+1} \leftarrow \theta_t - \eta_t \frac{\hat{m}_t}{\epsilon + \sqrt{\hat{v}_t}}$

$\quad i \leftarrow (i + 1)\% m$

end for

MicroAdam algorithm

Input: $\beta_1, \beta_2, \epsilon, \mathcal{G}, T, d, k$

$m_0, v_0 \leftarrow 0_d, 0_d$

$\delta_1, \Delta_1 \leftarrow 0, 0$

$e_1 \leftarrow 0_d^{4b}$

for $t = \{1, 2, ..., T\}$ **do**

$\quad g_t \leftarrow \widetilde{\nabla}_\theta f(\theta_t)$

$\quad a_t \leftarrow g_t + Q^{-1}(e_t, \delta_t, \Delta_t)$

$\quad \mathcal{I}_t, \mathcal{V}_t \leftarrow T_k(|a_t|)$

$\quad a_t[\mathcal{I}_t] \leftarrow 0$

$\quad \delta_{t+1}, \Delta_{t+1} \leftarrow min(a_t), max(a_t)$

$\quad e_{t+1} \leftarrow Q(a_t, \delta_{t+1}, \Delta_{t+1})$

$\quad \mathcal{G}_{i,:} \leftarrow (\mathcal{I}_t, \mathcal{V}_t)$

$\quad \hat{m}_t \leftarrow \textsc{AdamStats}(\beta_1, \mathcal{G})$

$\quad \hat{v}_t \leftarrow \textsc{AdamStats}(\beta_2, \mathcal{G}^2)$

$\quad \theta_{t+1} \leftarrow \theta_t - \eta_t \frac{\hat{m}_t}{\epsilon + \sqrt{\hat{v}_t}}$

$\quad i \leftarrow (i+1)\% m$

end for

MicroAdam - LLM Finetuning Results

Table 2: FFT results for Llama-2 7B/13B on GSM-8k.

LLaMA-2 size	**Optimizer**	**Accuracy**	**State**	**Total**	**Runtime**
7B	**Adam**	34.50%	25.1 GB	55.2 GB	1h 17m
	Adam-8b	34.34%	12.55 GB	42.5 GB	1h 18m
	MicroAdam ($m = 10$)	34.72%	5.65 GB	37.1 GB	1h 8m
	MicroAdam ($m = 20$)	35.10%	8.25 GB	39.7 GB	1h 37m

Theory: Gradient and Error Compression

Assumption 1. *The gradient compressor $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ is q-contractive with $0 \le q < 1$, i.e.,*

$$\|\mathcal{C}(x) - x\| \le q \|x\|, \quad \textit{for any } x \in \mathbb{R}^d.$$

Assumption 2. *The error compressor $\mathcal{Q} : \mathbb{R}^d \to \mathbb{R}^d$ is unbiased and ω-bounded with $\omega \ge 0$, namely,*

$$\mathbb{E}[\mathcal{Q}(x)] = x, \quad \|\mathcal{Q}(x) - x\| \le \omega \|x\|, \quad \textit{for any } x \in \mathbb{R}^d.$$

Assumption 3 (Lower bound and smoothness). *The loss function $f : \mathbb{R}^d \to \mathbb{R}$ is lower bounded by some $f^* \in \mathbb{R}$ and L-smooth, i.e., $\|\nabla f(\theta) - \nabla f(\theta')\| \le L \|\theta - \theta'\|$, for any $\theta, \theta' \in \mathbb{R}^d$.*

Assumption 4 (Unbiased and bounded stochastic gradient). *For all iterates $t \ge 1$, the stochastic gradient g_t is unbiased and uniformly bounded by a constant $G \ge 0$, i.e., $\mathbb{E}[g_t] = \nabla f(\theta_t)$, $\|g_t\| \le G$.*

Assumption 5 (Bounded variance). *For all iterates $t \ge 1$, the variance of the stochastic gradient g_t is uniformly bounded by some constant $\sigma^2 \ge 0$, i.e., $\mathbb{E}[\|g_t - \nabla f(\theta_t)\|^2] \le \sigma^2$.*

Assumption 6 (PL-condition). *For some $\mu > 0$ the loss f satisfies Polyak-Lojasiewicz (PL) inequality*

$$\|\nabla f(\theta)\|^2 \ge 2\mu(f(\theta) - f^*), \quad \textit{for any } \theta \in \mathbb{R}^d.$$

Theory: Non-convex and PL Convergence Rates

Theorem 1. (Non-convex convergence rate) *Let Assumptions 1, 2, 3, 4, 5 hold and* $\boxed{q_\omega := (1+\omega)q < 1}$. *Then, choosing* $\eta = \frac{1}{\sqrt{T}} \le \frac{\epsilon}{4LC_0}$, MICROADAM *(Algorithm 4) satisfies*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\theta_t)\|^2] \le 2C_0\left(\frac{f(\theta_1)-f^*}{\sqrt{T}} + \frac{L(\sigma^2 + C_2^2G^2)}{\epsilon\sqrt{T}}\right) + \mathcal{O}\left(\frac{G^3(G+d)}{T}\right)$$

with constants $C_0 := \sqrt{\frac{4(1+q_\omega^2)^3}{(1-q_\omega^2)^2}G^2 + \epsilon}$ *and* $C_2 := \omega q(1 + \frac{2q_\omega}{1-q_\omega^2})$.

Theorem 2. (PL convergence rate) *Let Assumptions 1, 2, 3, 4, 5 and 6 hold, and* $\boxed{q_\omega < 1}$. *Then, choosing* $\eta = \frac{C_0 \log T}{\mu T} \le \frac{\epsilon}{4LC_0}$, MICROADAM *(Algorithm 4) satisfies*

$$\mathbb{E}[f(\theta_{T+1})] - f^* \le \frac{f(\theta_1)-f^*}{T} + \frac{\log T}{T}\left(\frac{LC_0^2}{\mu}\frac{\sigma^2 + (C_1 + C_2^2)G^2}{\mu\epsilon} + \frac{C_0(1+C_1)(1+d)G^2}{\mu\sqrt{\epsilon}}\right) + \widetilde{\mathcal{O}}\left(\frac{G^4(G+d)}{T^2}\right)$$

with constant $C_1 := \frac{\beta_1}{1-\beta_1}(1 + C_2) + \frac{2q_\omega}{1-q_\omega^2}$.

Thank you!